

Stefan Lorkowski, Paul Cullen (Editors)
Analysing Gene Expression

A Handbook of Methods
Possibilities and Pitfalls

Related Titles from WILEY-VCH

Douglas T. Gjerde, Christopher P. Hanna, and David Hornby

DNA Chromatography

2002

ISBN 3-527-60074-4

Alfred Pingoud, Claus Urbanke, Jim Hoggett, and Albert Jeltsch

Biochemical Methods

A Concise Guide for Students and Researchers includes CD-ROM

2002

ISBN 3-527-30299-9

Reiner Westermeier and Tom Naven

Proteomics in Practice

A Laboratory Manual of Proteome Analysis

2002

ISBN 3-527-30354-5

Reiner Westermeier

Electrophoresis in Practice

3rd Edition

2001

ISBN 3-527-30300-6

Analysing Gene Expression

A Handbook of Methods: Possibilities and Pitfalls

Edited by

Stefan Lorkowski and Paul Cullen

Editors

Dr. Stefan Lorkowski

Institute of Arteriosclerosis Research
University of Münster
Domagkstr. 3
48149 Münster
Germany

Dr. Paul Cullen

Ogham GmbH
Mendelstr. 11
48149 Münster
Germany

Cover Illustration: The molecular structure of a hydrogel-coated surface plasmon resonance protein biochip. A covalently-attached polysaccharide monolayer serves as an immobilisation matrix for receptor molecules and reduces the possibility of non-specific interactions with the surface of the biochip.

The binding of ligands is detected in real-time via modulations in the intensity of a reflected laser beam (Reprinted with kind permission from XanTec bioanalytics GmbH, Münster, Germany. Copyright XanTec bioanalytics GmbH).

■ This book was carefully produced. Nevertheless, editors, authors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate. Each section in this book is the exclusive responsibility of the authors of that section.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by

Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at
<<http://dnb.ddb.de>>.

© 2003 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation in other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Printed in the Federal Republic of Germany.
Printed on acid-free paper.

Composition Hagedorn Kommunikation, Viernheim

Printing Druckhaus Darmstadt GmbH, Darmstadt

Bookbinding J. Schäffer GmbH & Co. KG, Grünstadt

ISBN 3-527-30488-6

Foreword

In the past century, humankind has made truly remarkable progress in understanding the mystery of life, from the discovery of DNA as a basic form of genetic material, the revelation of the chemical and structural nature of genes, and the establishment of a genetic central dogma, to the complete sequencing of the human genome. Although the central genetic dogma applied to a majority of life forms is defined as genetic information flow from DNA, to messenger RNA, then to protein, the complete sequencing of the three billion-letter human genome has shed little light on how precisely such unidirectional information flow of tens of thousands of genes is programmed. If the achievement of complete sequencing of the one-dimensional linear genetic code of the human genome can be compared with man landing on the moon, then the complexity and difficulty in interpreting the genomic instruction into the form of gene expression in a four-dimensional biological context, such as during development and disease, will prove to be a much more challenging and daunting task than that of getting man back from the moon to the earth.

Although classical genetics has been powerful in dissecting molecular diseases affected by the loss of function of a protein encoded by a single gene, such an approach has proved to be less fruitful for the understanding of phenotypes or diseases that are controlled by multiple genes, such as cancer, type 2 diabetes and heart disease. In fact, many of these genes themselves are signalling molecules, each of which controls the expression of a subset of downstream genes. Thus, analysing differential gene expression, or RNA genetics, a concept put forth by the late Ruth Sager, has become one of the most widely practiced strategies for studying the more complicated biological systems. Perhaps one of the earliest success stories of this approach was the discovery of the p53 tumour-suppressor protein in late 1970s as a protein over-expressed when normal cells were infected by DNA tumour viruses. Two-dimensional protein gel electrophoresis later was developed to provide a more complete picture of cellular protein expression. Methodologies that focused on mRNA expression, such as differential screening and subtractive hybridisation, were invented in the early 1980s, which proved to be more comprehensive, sensitive and informative in gene identification than two-dimensional protein gels. The discovery of T-cell receptors by Mark Davis and colleagues, when they compared the differences in mRNA expression between T and B-cells using such strategies,

provided the most beautiful example of gene discovery through the analysis of gene expression. The success of T-cell receptor discovery fuelled a great flood of biomedical research using gene expression analysis as a basic strategy in understanding a wide spectrum of biological systems. Several modern and more sophisticated molecular biological tools for the global analysis of gene expression at mRNA levels were invented in the 1990s. These methods, including differential display (DD), serial analysis of gene expression (SAGE) and DNA microarrays, have led to an explosion in the amount of research in gene expression analysis. The combined Medline hits of DD, SAGE and DNA microarrays have exceeded 6,000! Thereafter, numerous modifications of these technologies, as well as new approaches focusing on the analysis of gene expression at various levels both *in vivo* and *in vitro*, recently have been described.

This timely book entitled “Analysing Gene Expression” edited by the German scientist Dr. Stefan Lorkowski and the Irish physician Dr. Paul Cullen, and published by Wiley-VCH, took years to complete and represents a collective effort of over 200 researchers in the world, many of whom are leaders in the field of gene expression analysis. Unlike many previous books that focus on a single technology, “Analysing Gene Expression” provides a comprehensive description of nearly every technique and methodology ever invented for the analysis of gene expression, making it truly an encyclopaedia on this emerging subject in modern biology. The book includes not only basic background knowledge of gene expression at the levels of transcription, post-transcription, translation and post-translation, but also step-by-step protocols with a balanced and unbiased treatment of each methodology. Thus, whether you are a novice or a veteran in the field, this book will guide you through the jungle of old and new methodologies for gene expression analysis, and allow you to make informed choices as to which method(s) may best fit the type of biological problem under investigation.

For those who have never been there before, it cannot be emphasised enough that no methodology in gene expression analysis is foolproof, and finding a truly differentially expressed gene is only the first important step of a long journey. Undoubtedly, this book has everything you need to know to make that right first step. Ultimately, it will be the functional characterisations of each gene by genetic, cell biological and biochemical methods, that likely will provide the real proof (or dis-proof) of the relevance of a gene to the biological system under investigation. In the preface to a methodology book on protein purification, Dr. Arthur Kornberg once quoted an admonition of Efraim Racker, who said, “Don’t waste clean thinking on dirty enzymes” to illustrate the importance of the good biochemical practices that are at the core of enzymology. A similar doctrine, “Don’t waste clear thinking on dirty data”, certainly will continue to produce a better quality of science in the field of gene expression analysis in the new millennium.

Peng Liang
Nashville, September 2002

Preface

In February 2001, the draft sequence of the human genome was published. While at the time of writing many of the details still remain to be worked out, broad consensus now exists on the architecture of our genetic makeup. Understandably, therefore, the focus of cutting-edge research is now on functional genomics, *i.e.* the study of gene expression, and of its regulation, at the mRNA and the protein level.

Techniques for studying gene expression have burgeoned. However, as far as we are aware there does not yet exist a single comprehensive work devoted to this topic. This is what inspired us to compile and co-author this book. Our aim was two-fold. First, to provide a compendium of current methods of analysing gene expression with sufficient detail to allow the novice to decide what technique is most suitable for a particular application. Second, to put these different methods into perspective with relation to each other and to highlight the relative advantages and disadvantages associated with each.

We have divided this book into seven chapters, not because of the biblical charm of the number seven, but because we found this to be the most logical way to organise the content. Chapter 1 describes the fundamental biology of gene expression and chapter 2 outlines the tools needed to prepare samples and carry out gene expression analysis. Chapter 3 describes methods of mRNA expression analysis that can be implemented in the normal research setting, while chapter 4 is devoted to high-throughput methods more suitable to the industrial environment. Chapter 5 describes methods for analysing protein expression, chapter 6 is devoted to methods for analysing gene expression *in situ* and in the living organism at the mRNA and protein levels. Finally, chapter 7 rounds off the book by describing currently available bioinformatics approaches and internet databases.

To our knowledge, this work represents the most complete text currently available devoted solely to the topic of analysing gene expression. We have done our best to make it as up-to-date as possible. Most contributions represent the state of the art in March/April 2002. Although the two volumes of this book run to nearly 1,000 pages, it is still not possible for them to provide all the experimental detail needed for the researcher to implement the protocols in his or her laboratory. For that reason, we have made a special effort to comprehensively cite relevant literature, and to ensure that the papers and books cited are as recent as possible.

This is reflected in the large number of citations from 2001 and 2002. Despite our best efforts, it is possible that this book has overlooked important techniques or approaches. We apologize for any egregious gaps, responsibility for which is entirely our own. Wherever possible, we have tried to eliminate redundancy in the text. However, in a compendium of this nature, a degree of repetition is unavoidable, and we ask readers to overlook any particularly irritating examples that still remain.

No scholarly endeavour can today ignore the impact of the internet. In deference to this fact, we have included as much information as possible on currently available internet resources. One of the drawbacks of the world wide web is its ephemerality, and we hope that readers will excuse us if some of the links cited no longer function, or if their content is not as described.

This book would not have been possible without the encouragement and unflagging support of our publishers at Wiley-VCH Verlag. We are particularly grateful to Dr. Hans-Joachim Kraus for hours of helpful discussion and advice with regard to content and scope, and to Hans-Jochen Schmitt for expert advice with regard to the layout of the text. We thank Professor Gerd Assmann of the Institute of Arteriosclerosis Research for providing us with the facilities to carry out this work. Furthermore, we are grateful to our partners, Bernadette Biermann and Susanne Cullen for their support at all stages of the project. We also thank Bernadette Biermann for countless hours of correction and proof-reading. Parts of the work were made possible by a grant from the European Union to Dr. Paul Cullen (grant no. QLG1, 1999-01007). Limitations of space prevent us from naming the funding sources of all contributors, for which we would like to sincerely express our gratitude and that of the respective authors.

Throughout the text, we have referred to researchers in the masculine form only. This is purely a convention to save space and is in no way intended as a slight on our female colleagues.

Finally, we would like to thank all the contributors who made this work possible. We hope that it proves useful to our fellow-scientists in their attempt to navigate the choppy seas of the post-genomic era.

Stefan Lorkowski and Paul Cullen
Münster, September 2002

Contents

Volume 1

Foreword V

Preface VII

List of Contributors XXII

1	Basic concepts of gene expression	1
1.1	Introduction	1
1.2	Basics of transcription and translation in the cell	2
1.2.1	Introduction	2
1.2.2	Transcription	3
1.2.3	Translation	5
1.2.4	Summary	9
1.3	Regulation of transcription	9
1.3.1	Introduction	9
1.3.2	mRNA expression profiles – the transcriptome	9
1.3.3	Protein expression profiles – the proteome	13
1.3.4	Interaction between genes and proteins – the interactome	14
1.3.5	The transcription machinery and core promoters	16
1.3.6	Regulatory promoters	20
1.3.7	Enhancers	21
1.3.8	Locus control regions	22
1.3.9	Matrix attachment regions	23
1.3.10	Insulators	23
1.3.11	RIDGEs – Regions of increased gene expression	24
1.3.12	Enhanceosomes	25
1.3.13	Chromatin	25
1.3.14	Silencer elements	26
1.3.15	Transcription factors, repressors and co-repressors	26
1.3.16	Epigenetics	29
1.3.17	Summary and conclusions	30

1.4	Post-transcriptional regulation	30
1.4.1	Introduction	30
1.4.2	Regulation of RNA stability and degradation	31
1.4.3	Regulation of transcription elongation	33
1.4.3.1	Introduction	33
1.4.3.2	Regulation of transcription elongation in prokaryotes	34
1.4.3.3	Regulation of transcription elongation in eukaryotes	35
1.4.3.4	Conclusions	36
1.4.4	Differential/alternative pre-mRNA splicing	38
1.4.5	<i>Trans</i> -RNA splicing	39
1.4.6	Regulation of mRNA transport	40
1.4.7	Directed intracellular mRNA localisation	40
1.4.8	Regulation of mRNA polyadenylation	45
1.4.9	Antisense RNA	45
1.4.10	RNA editing	47
1.4.11	Summary and conclusions	49
1.5	Post-translational modification of proteins	51
1.5.1	Introduction	51
1.5.2	Proteolytic cleavage of proteins	52
1.5.3	Acylation	53
1.5.4	Prenylation	53
1.5.5	Methylation	54
1.5.6	Sulphation	54
1.5.7	Phosphorylation	54
1.5.8	Ubiquitination	55
1.5.9	Glycosylation	56
1.5.10	Conclusions	57
1.6	Correlation of mRNA and protein expression	58
1.6.1	Introduction	58
1.6.2	Levels of mRNA and protein expression: correlations and discrepancies	59
1.6.3	Conclusions	62
1.7	Housekeeping genes, internal and external standards	63
1.7.1	What are housekeeping genes?	63
1.7.2	Survey of the most important housekeeping genes	65
1.7.2.1	Glyceraldehyde-3-phosphate dehydrogenase	65
1.7.2.2	β -Actin	66
1.7.3	Other commonly used housekeeping genes	66
1.7.3.1	Ribosomal RNA (rRNA)	67
1.7.4	New identified 'maintenance genes'	67
1.7.5	Methods of quantification	69
1.7.5.1	Normalisation (<i>via</i> internal standards)	69
1.7.5.2	Standard curves (<i>via</i> external standards)	69
1.7.5.3	Alternative methods (<i>via</i> internal standards)	70
1.7.6	Summary	71

1.8	Classification of differential gene expression technologies	72
1.8.1	Introduction	72
1.8.2	From single genes to transcriptomes	72
1.8.3	Methods of classification	74
1.8.4	Conclusions	76
1.9	Summary	77
1.10	References	78
2	Sample preparation and supplementary tools	97
2.1	Introduction	97
2.2	Preparation of cells and tissues	99
2.2.1	Immunopurification of cells	99
2.2.2	Differential-velocity centrifugation/countercurrent elutriation	103
2.2.3	Surface affinity chromatography	105
2.2.4	Density-gradient centrifugation	106
2.2.5	Flow cytometry	107
2.2.5.1	Introduction	107
2.2.5.2	Analysers and cell sorters	108
2.2.5.3	Fluidics	109
2.2.5.4	Parameters and optics	112
2.2.5.5	Compensation	114
2.2.5.6	Cell sorting	116
2.2.5.7	Practical considerations	117
2.2.5.8	Discussion	119
2.2.6	Tissue microdissection techniques	120
2.2.6.1	Introduction	120
2.2.6.2	The force of focused light	120
2.2.6.3	LPC – Laser pressure catapulting	121
2.2.6.4	μCUT – Laser microdissection ultraviolet cut	125
2.2.6.5	Laser scissors	125
2.2.6.6	LMD – Laser microdissection	125
2.2.6.7	LCM – Laser capture microdissection	126
2.2.6.8	Ultrasonically oscillating needle	126
2.2.6.9	Manual tissue dissection	127
2.2.6.10	Discussion	127
2.2.7	Miscellaneous cell isolation and culturing techniques	127
2.3	Preparation of nucleic acids and proteins	130
2.3.1	Introduction	133
2.3.2	Isolation of total RNA and mRNA	133
2.3.2.1	Introduction	133
2.3.2.2	General considerations for isolation of RNA	134
2.3.2.3	Disruption and homogenisation of starting materials for isolation of RNA	135
2.3.2.4	RNA isolation technologies	136
2.3.3	Stabilisation of RNA prior to isolation	140

2.3.3.1	Introduction	140
2.3.3.2	RNA stabilisation in cells and tissues	140
2.3.3.3	RNA stabilisation in bacteria	140
2.3.3.4	RNA stabilisation in blood	141
2.3.3.5	Methods for stabilisation of the RNA expression profile	141
2.3.3.6	Commercially available RNA stabilisation reagents	143
2.3.4	Preparation of protein samples from cell cultures and tissues	144
2.3.4.1	Introduction	144
2.3.4.2	Separation and disruption of cells	144
2.3.4.3	Fractionation of cell extracts	145
2.3.4.4	Sample preparation for downstream applications	145
2.3.4.5	Immunoprecipitation	147
2.3.4.6	Conclusions	149
2.4	Summary	150
2.5	References	150

3 Methods for analysing mRNA expression 163

3.1	Introduction	163
3.2	Hybridisation-based methods	164
3.2.1	Branched DNA assay	164
3.2.2	Northern blot and related techniques	167
3.2.3	Nuclear run-on assay	173
3.2.4	Subtractive hybridisation	178
3.2.4.1	DISH – Differentiation induction subtraction hybridisation	178
3.2.4.2	DSC – Differential subtraction chain	185
3.2.4.3	EDS – Enzymatic degrading subtraction	192
3.2.4.4	GES – Gene expression screen	196
3.2.4.5	LCS – Linker capture subtraction	201
3.2.4.6	RaSH – Rapid subtraction hybridisation	206
3.2.4.7	cDNA-RDA – Representational difference analysis of cDNA	214
3.2.4.8	RSDD – Reciprocal subtraction differential RNA display	219
3.2.4.9	SABRE – Selective amplification <i>via</i> biotin and restriction-mediated enrichment	227
3.2.4.10	Subtractive hybridisation of transcriptionally active genomic DNA fragments	233
3.2.4.11	SuRF – Subtracted RNA fingerprinting	237
3.2.4.12	SSH – Suppressive subtractive hybridisation	244
3.2.5	Multiplexed DNA and RNA analysis on fluorescent microbeads	249
3.2.6	RPA – Ribonuclease protection assay	255
3.2.7	Virtual Northern blot	260
3.3	PCR-based methods	267
3.3.1	ADDER – Amplification of double-stranded cDNA end restriction fragments	267
3.3.2	ATAC-PCR – Adapter-tagged competitive PCR	278

3.3.3	cDNA-AFLP – cDNA-based amplified-fragment length polymorphism fingerprinting	285
3.3.4	Competitive RT-PCR	296
3.3.5	GEF – Gene expression fingerprinting	302
3.3.6	IBDD – Indexing-based differential display	308
3.3.7	Molecular indexing of 3'-cDNA ends	312
3.3.8	MSP-PCR – Multiplex PCR using colour-tagged module-shuffling primers	321
3.3.9	ODD – Ordered differential display	326
3.3.10	PACS – Preferential amplification of coding sequences	332
3.3.11	RAP-PCR – RNA arbitrarily primed PCR fingerprinting	340
3.3.12	Real-time reverse transcription polymerase chain reaction	345
3.3.13	RD-PCR – Restriction display polymerase chain reaction	358
3.3.14	SIP RT-PCR	363
3.3.15	Targeted display	370
3.4	Further methods for analysing gene expression based on other principles	373
3.4.1	RLCS – Restriction landmark cDNA scanning	373
3.4.2	RNA patterns method	379
3.5	Summary and conclusions	384
3.6	References	385

Volume 2

4	High-throughput and industrial methods for mRNA expression analysis	409
4.1	Introduction	409
4.2	Hybridisation-based techniques	410
4.2.1	DNA microarrays	410
4.2.1.1	Introduction	410
4.2.1.2	How DNA microarrays work	411
4.2.1.3	Production of DNA microarrays	416
4.2.1.4	Outlook	432
4.2.1.5	Transcriptome-based oligonucleotide design	432
4.2.1.6	State-of-the-art of nucleic acids labelling and detection	434
4.2.1.7	Conclusions	462
4.2.2	OFP – Oligonucleotide fingerprinting	463
4.2.3	Quantikine® mRNA assay	471
4.3	PCR-based techniques	478
4.3.1	ADGE – Amplified differential gene expression	478
4.3.2	DEPD® – Digital expression pattern display	484
4.3.3	FDD – Automated fluorescent mRNA differential display	494
4.3.4	GeneCalling	501
4.3.5	iAFLP – Introduced amplified fragment length polymorphism	506
4.3.6	RAGE – Rapid analysis of gene expression	511

4.3.7	READS™ – Restriction enzyme analysis of differential sequences	518
4.3.8	TOGA – Total gene expression analysis	529
4.4	Sequencing-based techniques	538
4.4.1	MPSS™ – Massively parallel signature sequencing	538
4.4.2	SAGE – Serial analysis of gene expression	551
4.4.3	MicroSAGE	557
4.4.4	MiniSAGE	568
4.4.5	ORESTES – Open reading frame expressed sequence tags	570
4.4.6	PCR-SAGE and SAGE-Lite	577
4.4.7	SADE – SAGE adaption for downsized extracts	581
4.4.8	TALEST – Tandem arrayed ligation of expressed sequence tags	585
4.5	Summary	592
4.6	References	594
5	Protein expression analysis	623
5.1	Introduction	623
5.2	Sample separation in proteomics	624
5.2.1	Introduction	624
5.2.2	Conventional slab gel electrophoresis	625
5.2.2.1	Isoelectric focusing (IEF)	625
5.2.2.2	Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and Western blotting	626
5.2.2.3	Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)	629
5.2.2.4	Two-dimensional difference gel electrophoresis (2D-DIGE)	630
5.2.3	High-performance liquid chromatography (HPLC)	630
5.2.3.1	Ion exchange chromatography (IEC)	633
5.2.3.2	Size exclusion chromatography (SEC)	633
5.2.3.3	Affinity chromatography	634
5.2.3.4	Reverse phase chromatography (RPC)	634
5.2.3.5	Normal phase chromatography	634
5.2.4	Capillary electrophoresis (CE) and capillary electrochromatography (CEC)	635
5.2.5	Multi-dimensional separations	635
5.2.6	New approaches	636
5.2.6.1	Isotope-coded affinity tag method (ICAT™)	636
5.2.6.2	Multi-epitope ligand-Kartographie (MELK)	637
5.2.7	Summary	638
5.3	Detection and identification of proteins by mass spectrometry	638
5.3.1	Introduction	638
5.3.2	How does mass spectrometry work?	639
5.3.2.1	Ionisation of molecules	640
5.3.2.2	Detection of ions	640
5.3.2.3	Appearance of spectra	642
5.3.2.4	Modes of operation	644
5.3.3	Proteomic mass spectrometry	646

5.3.3.1	Data analysis tools for mass spectroscopic experiments	651
5.3.3.2	Proteomics servers and databases	651
5.3.4	An example: Identification of an isolated and unknown protein	652
5.3.5	Is quantification possible with mass spectrometry?	655
5.3.6	Analysis of protein modifications	655
5.3.6.1	Phosphorylation	655
5.3.6.2	Glycosylation	656
5.3.7	Access to protein tertiary structure	657
5.3.8	Analysis of intact proteins and non-covalent protein complexes	659
5.3.9	Affinity mass spectrometry	659
5.3.10	Conclusion and outlook	660
5.4	Protein microarrays	660
5.4.1	Introduction	660
5.4.2	Principles and basics, typical results and applications	661
5.4.3	Discussion	670
5.5	Protein expression analysis with immunoassays	671
5.5.1	Introduction	671
5.5.2	Immunoassay reagents (binding agents)	671
5.5.3	Design options of immunoassays	674
5.5.3.1	Competitive immunoassays	674
5.5.3.2	Immunometric immunoassays	674
5.5.3.3	Homogenous and heterogeneous immunoassays	676
5.5.4	Types of label	678
5.5.4.1	Enzyme-linked immunosorbent assay (ELISA)	679
5.5.4.2	Direct label of antibody	682
5.5.5	Interferences	684
5.5.6	Conclusions	686
5.6	Flow cytometry and protein expression analysis	687
5.6.1	Introduction	687
5.6.2	Principles and basics	687
5.6.3	Results	689
5.6.4	Discussion	694
5.7	Summary	694
5.8	References	695
6	Methods for mRNA and protein expression analysis <i>in situ</i> and <i>in vivo</i>	703
6.1	Introduction	703
6.2	<i>In situ</i> hybridisation, immunocytochemistry and immunohistochemistry	704
6.2.1	Introduction	704
6.2.2	<i>In situ</i> hybridisation	705
6.2.2.1	Probe types	705
6.2.2.2	Probe labels and detection methods	707
6.2.2.3	Probe design strategy	708
6.2.2.4	Sample acquisition and fixation	709

6.2.2.5	Preparation and hybridisation of specimens	709
6.2.2.6	Optimisation and control of <i>in situ</i> hybridisation analysis	710
6.2.2.7	Increasing sensitivity of detection	711
6.2.3	Immunocytochemistry and immunohistochemistry	711
6.2.4	Advanced techniques	713
6.2.5	Application in the study of drug efficacy and toxicity	714
6.3	Visualising RNA in living cells	715
6.3.1	Introduction	715
6.3.2	Fluorescence <i>in vivo</i> hybridisation	716
6.3.2.1	Probe selection	716
6.3.2.2	Cell transfection	719
6.3.2.3	Detection of hybridised RNA in living cells	720
6.3.3	Alternative methods for visualising RNA <i>in vivo</i>	722
6.3.3.1	Direct and indirect labelling	722
6.3.3.2	Two plasmid expression	722
6.4	MRI – Magnetic resonance imaging	724
6.4.1	Introduction	724
6.4.2	Imaging mRNA with magnetic switches	725
6.4.3	Imaging proteins with targeted probes	727
6.4.4	High-throughput imaging	727
6.4.5	Magnetic resonance-sensitive marker genes	728
6.4.5.1	Tyrosinase	728
6.4.5.2	Activatable enzyme-sensitive magnetic resonance probes	729
6.4.5.3	Internalising receptors	731
6.5	MRS – Magnetic resonance spectroscopy	732
6.5.1	Introduction	732
6.5.2	Principles and basics	733
6.5.3	<i>In vivo</i> spectroscopy of mRNA and protein expression	735
6.5.4	Discussion	737
6.6	PET – Positron emission tomography	739
6.6.1	Introduction	739
6.6.2	From clinical PET to quantitative small laboratory animal tomography	742
6.6.3	Examples of molecular imaging and imaging of gene expression with PET	742
6.6.4	Quantitative imaging in oncology	743
6.6.5	Quantitative imaging of transgene and endogenous gene expression	743
6.6.6	Discussion and conclusions	746
6.7	SPECT – Single photon emission computed tomography	747
6.7.1	Introduction	747
6.7.2	Principles and basics	747
6.7.3	Direct imaging of gene expression	749
6.7.4	Indirect imaging <i>via</i> expressed proteins	750
6.7.5	Conclusions	752
6.8	Optical <i>in vivo</i> imaging	752
6.8.1	Introduction	752

6.8.2	Fluorescent proteins	753
6.8.3	Bioluminescence	755
6.8.4	NIRF – Near-infrared fluorescent imaging	755
6.8.4.1	Optical imaging of mRNA with molecular beacons	755
6.8.5	Enzyme-sensing, activatable NIRF probes	756
6.8.5.1	Receptor-targeted NIRF-probes	758
6.9	Summary	758
6.10	References	759
7	Computational methods and bioinformatic tools	769
7.1	Introduction	769
7.2	Comparative expressed sequence tag analysis	770
7.2.1	Introduction	770
7.2.2	Processing expressed sequence tags prior to content analysis	771
7.2.3	Gene content and annotation of expressed sequence tags	772
7.2.4	Expressed sequence tags in comparative genomics	774
7.2.5	<i>In silico</i> subtraction using clustered sets of expressed sequence tags	776
7.2.6	Expressed sequence tag data repositories and cDNA clone distribution centres	777
7.3	Data management and data mining	778
7.3.1	Introduction	778
7.3.1.1	Current situation	778
7.3.1.2	Future development	779
7.3.1.3	Taking part in bioinformatics	779
7.3.1.4	Hardware and software demands	780
7.3.1.5	Data types, structures and processing	780
7.3.1.6	Communication structures	781
7.3.2	Building a test scenario	781
7.3.2.1	Microarray experiments	781
7.3.2.2	Analysing the workflow – getting things done	783
7.3.2.3	Designing the question and choosing the right tools for the answer	784
7.3.2.4	Scaling up	784
7.3.3	Strategies of data mining	784
7.3.3.1	Data evaluation and representation	784
7.3.3.2	Principles of query languages	785
7.3.3.3	Data mining	785
7.3.3.4	Custom solutions	786
7.3.4	Summary	787
7.4	Integration of heterogeneous high-throughput gene expression data	787
7.4.1	Introduction	787
7.4.2	Steps towards data integration	788
7.4.3	Initial steps in realising data integration	788
7.4.4	Conclusions	789
7.5	Cluster analysis of gene expression profiles	790
7.5.1	Introduction	790

7.5.2	Information content of gene expression clusters	790
7.5.3	Similarity matrices and gene expression matrices	792
7.5.4	Clustering algorithms	792
7.5.4.1	Hierarchical clustering	793
7.5.4.2	Self-organising maps	794
7.5.4.3	K-means	795
7.5.4.4	Gene shaving	796
7.5.5	Evaluation of gene expression clusters	797
7.5.6	Conclusion	798
7.6	Promoter finding in eukaryotic genomes	798
7.6.1	Introduction	798
7.6.2	Transcription regulation in eukaryotes	799
7.6.2.1	Promoter structure	799
7.6.2.2	Transcription factors	799
7.6.2.3	Combinatorial nature of transcription regulation	800
7.6.3	Databases on transcriptional regulation	801
7.6.4	<i>In silico</i> study of gene transcription regulation	802
7.6.4.1	Recognition of <i>cis</i> -regulatory elements	802
7.6.4.2	Recognition of composite regulatory elements	805
7.6.4.3	Recognition of promoters	808
7.6.5	Conclusions	811
7.7	GeneEMAC – Three-dimensional visualisation of gene expression	811
7.7.1	Introduction	811
7.7.2	Principles and basics of the GeneEMAC concept	813
7.7.3	Specimen preparation	814
7.7.3.1	Whole-mount <i>in situ</i> hybridisation	814
7.7.3.2	Embedding	814
7.7.3.3	Introduction of external markers	814
7.7.3.4	Capturing of a reference image	815
7.7.3.5	Histological sectioning	815
7.7.4	Microscopy and digital image processing	815
7.7.4.1	Image capturing	815
7.7.4.2	Image congruencing	816
7.7.4.3	Image segmentation	816
7.7.4.4	Generation of a three-dimensional model	816
7.7.5	Visualisations of models	817
7.7.6	Examples	817
7.7.6.1	Whole-organism screens	817
7.7.6.2	High-resolution analysis	818
7.7.7	Discussion	819
7.8	RNA-based gene expression databases and analyses tools	819
7.8.1	Introduction	819
7.8.1.1	Choice of a database management systems	820
7.8.1.2	cDNA database based on relational database management systems	820
7.8.1.3	Spatial and temporal gene expression data	820

7.8.1.4	Data format	821
7.8.1.5	Types of data	822
7.8.1.6	Use of other databases	823
7.8.2	ASDB – The Alternative Splicing Database	824
7.8.3	AsMamDB – The Alternative Splice Database of Mammals	825
7.8.4	BodyMap – An anatomical gene expression database of human and mouse	825
7.8.5	The CYTOMER® Gene Expression Database on human organs and cell types	826
7.8.6	Database of three-dimensional visualisation of gene expression	828
7.8.7	dbEST – The Database of Expressed Sequence Tags	829
7.8.8	DDD – Digital Differential Display	830
7.8.9	The Genexpress IMAGE Knowledge Base of the Human Genome and Transcriptomes	832
7.8.10	The Gencarta™ Database	833
7.8.11	GEO – Gene Expression Omnibus Database	834
7.8.12	GXD – The Mouse Gene Expression Database	835
7.8.13	ISG Database – Interferon-Stimulated Gene Database	836
7.8.14	The Kidney Development Database	837
7.8.15	MAGEST – The Maboya Gene Expression Patterns and Sequence Tags Database	838
7.8.16	MethDB – The DNA Methylation Database	839
7.8.17	EMAGE – The Edinburgh Mouse Atlas Gene Expression Database	840
7.8.18	RAD – The RNA Abundance Database	842
7.8.19	The Rochester Muscle Database	843
7.8.20	SAGEmap – The serial analysis of gene expression tag to gene mapping database	844
7.8.21	SGD – The <i>Saccharomyces</i> Genome Database and its Expression Connection	845
7.8.22	SMD – The Stanford Microarray Database	845
7.8.23	TRIPLES – The Database of Transposon-Insertion Phenotypes, Localisation, and Expression in <i>Saccharomyces cerevisiae</i>	847
7.8.24	UTRdb and UTRsite – The Specialised Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs	849
7.8.25	yMGV – The Yeast Microarray Global Viewer	850
7.9	Protein-based gene expression databases and analyses tools	851
7.9.1	Introduction to protein-based gene expression databases	851
7.9.2	The Proteome Analysis Database	853
7.9.3	SWISS-2DPAGE – A two-dimensional polyacrylamide gel electrophoresis database	854
7.10	Further gene-expression databases in the internet	855
7.11	Summary	860
7.12	References	871

List of Contributors

The complete postal addresses of all authors including their telephone and fax numbers as well as the email addresses are available at the homepage of the book publication at <http://www.wiley-vch.de/home/3-527-30488-6>.

Ruedi Aebersold,
The Institute for Systems Biology,
Seattle (USA)
Author of section 1.6

Glenn Albrecht,
Lynx Therapeutics, Inc.,
Hayward, (USA)
Author of section 4.4.1

Heiko Alfke,
Philipps University,
Marburg (Germany)
*First and corresponding author of section 6.5;
senior author of section 6.6; author of section 6.7*

Alon Amit,
Compugen Ltd.,
Jamesburg (USA)
*Corresponding author of sections 4.2.1.3
and 7.8.10*

Rolf Apweiler,
The European Bioinformatics Institute,
Hinxton (UK)
Senior and corresponding author of section 7.9.2

Charles Auffray,
Genexpress,
Centre National de la Recherche Scientifique,
Villejuif Cedex (France)
Senior and corresponding author of section 7.8.9

Christian W.B. Bachem,
Wageningen University and Research Centre
(The Netherlands)
First and corresponding author of section 3.3.3

Sabine Bahn,
Babraham Institute,
Cambridge (UK)
Author of section 3.3.6

Richard Baldock,
Western General Hospital,
Edinburgh (UK)
Author of section 7.8.17

Catherine A. Ball,
School of Medicine,
Stanford University, Stanford (USA)
*First and corresponding author of section 7.8.21;
author of section 7.8.22*

Oliver Bauer,
Max-Planck-Institut für Molekulare Genetik,
Berlin (Germany)
First author of section 4.2.2

Martin P. Béhé,
Philipps University,
Marburg (Germany)
Author of section 6.6; first author of section 6.7

Thomas M. Behr,
Philipps University,
Marburg (Germany)
*First and corresponding author of section 6.6;
senior and corresponding author of section 6.7*

Nicholas J. Beauchamp,
University of Amsterdam,
Amsterdam (The Netherlands)
Author of section 4.4.2

Alexander V. Belyavsky,
Engelhardt Institute of Molecular Biology,
Moscow (Russia)
Corresponding author of section 3.3.5

Paul Bertone,
Yale University,
New Haven (USA)
Author of section 7.8.23

Ralf Bickel,
Clondia Chip Technologies GmbH,
Jena (Germany)
*First and corresponding author
of sections 4.2.1.4 and 4.2.1.4.4*

Gail Binkley,
School of Medicine,
Stanford University,
Stanford (USA)
Author of sections 7.8.21 and 7.8.22

Margaret Biswas,
ViaLactia Biosciences Ltd.,
Auckland (New Zealand)
First author of section 7.9.2

Mark E. Bolander,
Mayo Clinic and Foundation,
Rochester (USA)
Author of sections 3.2.4.11 and 3.3.10

Sylvie Bortoli,
Commissariat à l'Énergie Atomique (CEA),
Evry Cedex (France)
Author of section 7.8.9

David Botstein,
School of Medicine, Stanford University,
Stanford (USA)
Author of section 7.8.21 and section 7.8.22

Annabelle Bouchardon,
INSERM U371 "Cerveau et Vision",
Bron (France)
Author of section 3.2.4.9

Lucas D. Bowler,
Trafford Centre for Medical Research,
Brighton (UK)
Exclusive author of section 3.2.4.7

Christoph Bremer,
University of Münster,
Münster (Germany)
First author of sections 6.4 and 6.8

Sydney Brenner,
Lynx Therapeutics, Inc.,
Hayward (USA)
Author of section 4.4.1

Donald D. Brown,
Carnegie Institution of Washington,
Baltimore (USA)
Senior author of section 3.2.4.4

Patrick O. Brown,
Stanford University,
Stanford (USA)
Author of section 7.8.22

Iris Bruchhaus,
Bernhard-Nocht-Institut für Tropenmedizin,
Hamburg (Germany)
Senior author of section 3.2.7

Tim Burcham,
Lynx Therapeutics, Inc.,
Hayward (USA)
Author of section 4.4.1

Renate Burgemeister,
P. A. L. M. Microlaser Technologies AG,
Bernried (Germany)
First and corresponding author of section 2.2.6

Stephen A. Bustin,
Queen Mary's School of Medicine
and Dentistry, Royal London Hospital,
London (UK)
Exclusive author of section 3.3.12

Claude Chelala,
Genexpress,
Centre National de la Recherche Scientifique,
Villejuif Cedex (France)
Author of section 7.8.9

Lianggui Chen,
Digital Gene Technologies,
La Jolla (USA)
Author of section 4.3.8

Xin Chen,
The National Laboratory of Protein
Engineering and Plant Genetic Engineering,
Peking University,
Beijing (People's Republic of China)
Author of section 7.8.5

Zhijian J. Chen,
Fox Chase Cancer Center,
Philadelphia (USA)
First author of section 4.3.1

J. Michael Cherry,
School of Medicine,
Stanford University,
Stanford (USA)
Author of section 7.8.21

Kei-Hoi Cheung,
School of Medicine, Yale University,
New Haven (USA)
Author of section 7.8.23

Kevin Corcoran,
Lynx Therapeutics, Inc.,
Hayward (USA)
Author of section 4.4.1

Jonathan Crabtree,
Center for Bioinformatics,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

Paul Cullen,
Ogham GmbH,
Münster (Germany)
*Senior author of sections 1.3.1, 1.4.1 to 1.4.3,
1.4.5 to 1.4.8, 1.4.10, 1.4.11, 1.5, 1.8, 2.1, 2.2.1
to 2.2.4, 2.2.7, 2.3.1, 2.4, 3.1, 3.2.4, 4.2.1.4.1 to
4.2.1.4.3, 4.2.1.4.5 to 4.2.1.4.10, 4.2.1.5, 4.3.7,
4.4.4, 4.4.5, 7.9.1, 7.10; first and corresponding
author of sections 1.1, 1.3.7, 1.9, 3.5, 4.1, 4.2.1.1,
4.2.1.2, 4.5, 5.7, 6.1, 6.9, 7.1, 7.11; author of
sections 1.7, 3.2.4.3, 3.3.15, 3.4.2, 4.4.7, 6.3*

Nicole A. Datson,
Leiden/Amsterdam Center for Drug Research,
Leiden University Medical Centre,
Leiden (The Netherlands)
First and corresponding author of section 4.4.3

Duncan R. Davidson,
Western General Hospital,
Edinburgh (UK)
Corresponding author of section 7.8.17

Jamie A. Davies,
University of Edinburgh, Medical School,
Edinburgh (UK)
Exclusive author of section 7.8.14

Warren Davis, Jr.,
Fox Chase Cancer Center,
Philadelphia (USA)
Author of section 4.3.1

Michael J. de Veer,
The Walter and Eliza Hall Institute of Medical
Research, The Royal Melbourne Hospital,
Victoria (Australia)
First author of section 7.8.13

Vivian de Waard,
Academic Medical Center,
University of Amsterdam (The Netherlands)
First author of section 4.4.2

Charles Decraene,
Commissariat à l'Énergie Atomique (CEA),
Evry Cedex (France)
Author of section 7.8.9

Colette Dehay,
INSERM U371 "Cerveau et Vision",
Bron (France)
Author of section 3.2.4.9

Marie-Dominique Devignes,
LORIA – Langue et dialogue,
Vandoeuvre les Nancy (France)
Author of section 7.8.9

Kara Dolinski,
School of Medicine,
Stanford University,
Stanford (USA)
Author of section 7.8.21

Peter A. Doris,
Houston Health Sciences Center,
Houston (USA)
Senior and corresponding author of section 3.3.4

Inna Dubchak,
Lawrence Berkeley National Laboratory,
Berkeley (USA)
Exclusive author of section 7.8.2

Selina S. Dwight,
School of Medicine, Stanford University,
Stanford (USA)
Author of section 7.8.21

Ralf Ehricht,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Jean-Marc Elalouf,
Commissariat à l'Énergie Atomique (CEA)
Saclay,
Gif-sur-Yvette (France)
Senior and corresponding author of section 4.4.7

Thomas Ellinger,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Michael Erren,
University of Münster,
Münster (Germany)
Senior author of section 5.5

Eugen Ermantraut,
Clondia Chip Technologies GmbH,
Jena (Germany)
Senior author of sections 4.2.1.4 and 4.2.1.4.4

Eric Eveno,
Genexpress,
Centre National de la Recherche Scientifique,
Villejuif Cedex (France)
First author of section 7.8.9

Asun Fernandes del Carmen,
Wageningen University and Research Centre,
Wageningen (The Netherlands)
Author of section 3.3.3

Paul B. Fisher,
Columbia University,
New York (USA)
Senior and corresponding author of sections 3.2.4.1, 3.2.4.6 and 3.2.4.8

Jeffrey S. Fisher,
GenHunter Corporation,
Nashville (USA)
First author of section 4.3.3

Manfred Fobker,
University of Münster
Münster (Germany)
First and corresponding author of section 5.5

Ellen Fricke,
BIOBASE GmbH,
Wolfenbüttel (Germany)
First author of section 7.8.5

Laurent Gate,
Fox Chase Cancer Center,
Philadelphia (USA)
Author of section 4.3.1

Erk Gedig,
XanTec bioanalytics GmbH,
Münster (Germany)
Exclusive author of section 5.4

Carsten Gerdemann,
Abbott GmbH,
Wiesbaden (Germany)
Exclusive author of sections 3.2.4.10 and 3.3.13

Neil I. Goldstein,
DGI Biotechnologies, Inc.,
Edison (USA)
Author of section 3.2.4.1

Jeremy Gollub,
Stanford University,
Stanford (USA)
Author of section 7.8.22

David Goodlett,
The Institute for Systems Biology,
Seattle (USA)
Corresponding author of section 1.6

Rahul V. Gopalkrishnan,
Columbia University,
New York (USA)
First author of section 3.2.4.1

Gregory R. Grant,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

David P. L. Green,
University of Otago,
Dunedin (New Zealand)
Senior and corresponding author of section 4.4.6

Jens Grote,
University of Münster,
Münster (Germany)
Author of section 5.3

Christoph Grunau,
Institute de Génétique Humain,
Montpellier (France)
Exclusive author of section 7.8.16

Anna Guerasimova,
Max-Planck-Institut für Molekulare Genetik,
Berlin (Germany)
Author of section 4.2.2

Zhen Guo,
GenHunter Corporation,
Nashville (USA)
Author of section 4.3.3

Jörg Haberland,
Ogham GmbH,
Münster (Germany)
Exclusive author of sections 2.3.4 and 5.2

Jennifer Harrington,
R&D Systems, Inc.,
Minneapolis, MS (USA)
Senior author of section 4.2.3

Martin Haubrock,
BIOBASE GmbH,
Wolfenbüttel (Germany)
Author of section 7.8.5

Joan Hebert,
Center for Clinical Sciences Research,
Stanford (USA)
Author of section 7.8.22

Steffen Hennig,
Max Planck Institut für Molekulare Genetik,
Berlin (Germany)
First and corresponding author section of 7.2

Tina Hernandez-Boussard,
Center for Clinical Sciences Research,
Stanford (USA)
Author of section 7.8.22

Ralf Herwig,
Max-Planck Institute for Molecular Genetics,
Berlin (Germany)
*Author of section 4.2.2, first and corresponding
author of section 7.5*

Beate Hess,
Würzburg (Germany)
Exclusive author of section 3.2.2 and 3.2.4.12

Johannes T. Heverhagen,
Philipps University (Germany)
Senior author of section 6.5

Brian S. Hilbush,
Digital Gene Technologies, Inc.,
La Jolla (USA)
First author of section 4.3.8

Cruz A. Hinojos,
University of Texas-Houston,
Health Sciences Center,
Houston (USA)
Author of section 3.3.4

Christine Hoogland,
Swiss Institute of Bioinformatics,
Geneva (Switzerland)
Exclusive author of section 7.9.3

Sebastian Horstmann,
University of Cologne,
Cologne (Germany)
Exclusive author of section 3.2.1 and 3.3.11

Sandrine Imbeaud,
Genexpress,
Centre National de la Recherche Scientifique,
Villejuif Cedex (France)
Author of section 7.8.9

Laurie Issel-Tarver,
School of Medicine, Stanford University,
Stanford (USA)
Author of section 7.8.21

Takashi Ito,
Cancer Research Institute,
Kanazawa University,
Ishikawa (Japan)
Author of section 3.3.8

Michal Janitz,
Max-Planck-Institut für Molekulare Genetik,
Berlin (Germany)
Author of section 4.2.2

Hongkai Ji,
Tsinghua University,
Beijing (People's Republic of China)
Exclusive author of section 7.8.3

Hongping Jiang,
Medco Health Solutions, Inc.,
New Jersey (USA)
Author of section 3.2.4.6

Heng Jin,
Stanford University,
Stanford (USA)
Author of section 7.8.22

Miroslava Kaloper,
School of Medicine, Stanford University,
Stanford (USA)
Author of section 7.8.22

Hideki Kambara,
Hitachi Ltd.,
Tokyo (Japan)
Author of section 3.3.8

Dong-chul Kang,
Columbia University,
New York (USA)
*Author of section 3.2.4.1; first author of sections
3.2.4.6 and 3.2.4.8*

Dagmar Karas,
BIOBASE GmbH,
Wolfenbüttel (Germany)
Author of section 7.8.5

Kikuya Kato,
Nara Institute of Science and Technology,
Nara (Japan)
Exclusive author of sections 3.3.2 and 3.3.7

Shoko Kawamoto,
Medical Institute for Bioregulation,
Fukuoka (Japan)
Senior author of sections 4.3.5 and 7.8.4

Takeshi Kawashima,
Kyoto University,
Kyoto (Japan)
First author of section 7.8.1

Alexander Kel,
BIOBASE GmbH,
Wolfenbüttel (Germany)
First and corresponding author of section 7.6

Paul Kersey,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton (UK)
Author of section 7.9.2

Clive N. Kind,
AstraZeneca R & D Charnwood,
Loughborough (UK)
Exclusive author of section 6.2

Hanns-Georg Klein,
IMGM Laboratories GmbH,
Martinsried (Germany)
*Senior and corresponding author
of sections 1.3.6 to 1.3.16, 1.4.4, 1.4.9*

Simone König,
University of Münster,
Münster (Germany)
*Exclusive author of section 5.1;
first and corresponding author of section 5.3*

Benoît Kornmann,
University of Geneva,
Geneva (Switzerland)
First author of section 3.3.1

Eberhard Korsching,
University of Münster,
Münster (Germany)
Exclusive author of section 7.3

Mario Kratz,
University of Münster,
Münster (Germany)
Exclusive author of section 1.2

Anuj Kumar,
Cellular and Developmental Biology,
Yale University,
New Haven (USA)
First author of section 7.8.23

Karl J. Lackner,
University Hospital Mainz,
Mainz (Germany)
Exclusive author of section 3.2.5

Sigrid Land,
BIOBASE GmbH,
Wolfenbüttel (Germany)
Author of section 7.8.5

Claus Langer,
University of Münster,
Münster (Germany)
First and corresponding author of section 3.2.3

Daniel J. Lavery,
Purdue Pharma L. P.,
Cranbury (USA)
Corresponding author of section 3.2.4.9

Phuc V. Le,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

Irina V. Lebedeva,
Columbia University,
New York (USA)
Author of section 3.2.4.1

Hans Lehrach,
Max Planck Institut für Molekulare Genetik,
Berlin (Germany)
Author of sections 4.2.2, 7.2 and 7.5

Magdalena Leszczyniecka,
Novartis Pharmaceuticals,
Summit (USA)
Author of section 3.2.4.1

Ines Leube,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Peng Liang,
Vanderbilt University,
Nashville (USA)
Senior and corresponding author of section 4.3.3

Junmin Liu,
University of Pennsylvania,
Philadelphia (USA)
Senior author of section 7.8.18

Yang Liu,
School of Medicine, Yale University,
New Haven (USA)
Author of section 7.8.23

David Lo,
Digital Gene Technologies,
La Jolla (USA)
Corresponding author of section 4.3.8

Gerhard Lorkowski,
Pharma Consulting,
Münster (Germany)
Author of sections 4.2.1.1 and 4.2.1.2

Stefan Lorkowski,
University of Münster,
Münster (Germany)
Senior author of sections 1.1, 1.3.17, 3.5, 4.1, 4.2.1.1, 4.2.1.2, 4.5, 5.7, 6.1, 6.9, 7.1, 7.11; senior and corresponding author of sections 1.7, 3.2.4.3, 3.4.2, 3.3.15, 6.3; first and corresponding author of sections 1.3.1, 1.4.1 to 1.4.3, 1.4.5 to 1.4.8, 1.4.10, 1.4.11, 1.5, 1.8, 1.9, 2.1, 2.2.1 to 2.2.4, 2.2.7, 2.3.1, 2.4, 3.1, 3.2.4, 4.2.1.4.1 to 4.2.1.4.3, 4.2.1.4.5 to 4.2.1.4.10, 4.2.1.5, 4.3.7, 4.4.4, 4.4.5, 7.9.1, 7.10; first author of section 4.4.7

Hermann Lübbert,
Biofrontera Pharmaceuticals AG,
Leverkusen (Germany)
Exclusive author of section 4.3.2

Jian-Hua Luo,
School of Medicine, University of Pittsburgh,
Pittsburgh (USA)
First and corresponding author of section 3.2.4.2

Michael C. MacLeod,
University of Texas,
Smithville (USA)
Senior and corresponding author of section 4.3.6

Harin Mahadeva,
University of Leicester,
Leicester (UK)
Author of section 3.3.6

Kazuhiro W. Makabe,
Kyoto University,
Kyoto (Japan)
Senior and corresponding author of section 7.8.1; exclusive author of section 7.8.15

Elisabetta Manduchi,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

Philippe Marc,
Ecole Normale Supérieure,
Paris (France)
Exclusive author of section 7.8.25

Luis Marengo,
Yale University School of Medicine,
New Haven (USA)
Author of section 7.8.23

Régine Mariage-Samson,
Genexpress,
Centre National de la Recherche Scientifique,
Villejuif Cedex (France)
Author of section 7.8.9

Christoph Marschall,
IMGM Laboratories GmbH,
Martinsried (Germany)
First author of sections 1.3.6 to 1.3.16, 1.4.4, 1.4.9

Peter Masiar,
Yale University School of Medicine,
New Haven (USA)
Author of section 7.8.23

John C. Matese,
Stanford University,
Stanford (USA)
First and corresponding author of section 7.8.22

Mikhail V. Matz,
University of Florida,
St. Augustine (USA)
Exclusive author of section 3.3.9

Douglas J. McGarvey,
QIAGEN GmbH,
Hilden (Germany)
First and corresponding author of sections 2.3.2 and 2.3.3

Shannon K. McWeeney,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

Jonathan Meade,
GenHunter Corporation,
Nashville (USA)
Author of section 4.3.3

Perry Miller,
School of Medicine, Yale University,
New Haven (USA)
Author of section 7.8.23

Fumihito Miura,
Kanazawa University,
Ishikawa (Japan)
Author of section 3.3.8

Gerd B. Müller,
University of Vienna,
Vienna (Austria)
Senior author of sections 7.7 and 7.8.6

Junko Nishida,
Hitachi Ltd.,
Tokyo (Japan)
Author of section 3.3.8

Kazunori Okano,
Hitachi Ltd.,
Tokyo (Japan)
Author of section 3.3.8

Kousaku Okubo,
Kyushu University,
Fukuoka (Japan)
*First and corresponding author of section 4.3.5
and 7.8.4*

Hans Pannekoek,
University of Amsterdam,
Amsterdam (The Netherlands)
Corresponding author of section 4.4.2

Georgia Panopoulou,
Max-Planck Institut für Molekulare Genetik,
Berlin (Germany)
Author section of 7.2

Wanda Panter,
R&D Systems, Inc.,
Minneapolis (USA)
First and corresponding author of section 4.2.3

John E. Pearl,
The Trudeau Institute,
New York (USA)
Senior author of section 5.6

Graziano Pesole,
University of Milan,
Milan (Italy)
Exclusive author of 7.8.24

Oxana Pickeral,
Human Genome Sciences, Inc.,
Rockville (USA)
*Exclusive author of sections 7.4, 7.8.7, 7.8.11
and 7.8.20*

Geneviève Piétu,
Commissariat à l'Énergie Atomique,
Evry Cedex (France)
Author of section 7.8.9

Angel Pizarro,
University of Pennsylvania,
Philadelphia (USA)
Author of section 7.8.18

Joan Pontius,
National Center for Biotechnology
Information, National Institute of Health,
Bethesda (USA)
Exclusive author of section 7.8.8

Albert Poustka,
Max-Planck Institut für Molekulare Genetik,
Berlin (Germany)
Senior author of section 7.2

Nicolas Preitner,
University of Geneva,
Geneva (Switzerland)
Author of section 3.3.1

Stephen W. Pursley,
DakoCytomation, Inc.,
Fort Collins (USA)
First and corresponding author of section 5.6

Achim Quandt,
QIAGEN GmbH,
Hilden (Germany)
Senior author of sections 2.3.2 and 2.3.3

Uwe Radelof,
Deutsches Ressourcenzentrum
für Genomforschung GmbH,
Berlin (Germany)
Author of section 4.2.2

Thomas Rasmussen,
University of Copenhagen,
Copenhagen (Denmark)
Senior author of section 3.3.14

Jill M. Ray,
Capital Genomix, Inc.,
Gaithersburg (USA)
First author of section 4.3.6

Daniele Rifat,
University of Geneva,
Geneva (Switzerland)
Author of section 3.3.1

Martin Ringwald,
The Jackson Laboratory,
Bar Harbor (USA)
Exclusive author of section 7.8.12

Thomas Roeder,
University of Würzburg,
Würzburg (Germany)
First and corresponding author of section 3.2.7

Stella Rotert,
BIOBASE GmbH,
Wolfenbüttel (Germany)
Author of section 7.8.5

Norrie Russell,
Lynx Therapeutics, Inc.,
Hayward (USA)
Senior author of section 4.4.1

Jana Sachtschal,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Yoshiyuki Sakaki,
University of Tokyo,
Tokyo (Japan)
Senior author of section 3.3.8

Giovanni Salerno,
DakoCytomation Flow Center GmbH,
Freiburg i. Br. (Germany)
Exclusive author of section 2.2.5

Devanand Sarkar,
Columbia University,
New York (USA)
Author of section 3.2.4.1

Gobinda Sarkar,
Mayo Clinic and Foundation,
Rochester (USA)
*Corresponding and senior author
of sections 3.2.4.11 and 3.3.10*

Moirá Sauane,
Columbia University,
New York (USA)
Author of section 3.2.4.1

Ueli Schibler,
University of Geneva,
Geneva (Switzerland)
Senior and corresponding author of section 3.3.1

Michael Schnoor,
University of Münster,
Münster (Germany)
First author of sections 3.3.15 and 6.3

Mark Schroeder,
School of Medicine, Stanford University,
Stanford (USA)
Author of section 7.8.21

Karin Schütze,
P. A. L. M. Microlaser Technologies AG,
Bernried (Germany)
Senior author of section 2.2.6

Udo Seedorf,
University of Münster,
Münster (Germany)
*Senior author of section 3.2.3; exclusive
author of section 3.2.6*

Gavin J. Sherlock,
Stanford University,
Stanford (USA)
*Author of section 7.8.21; senior author
of section 7.8.22*

Anne Sherwood,
Inbios International, Inc.,
Seattle (USA)
First author of sections 3.2.4.11 and 3.3.10

Richard A. Shimkets,
CuraGen Corporation,
Branford (USA)
Exclusive author of section 4.3.4

Sergey V. Shmelkov,
Weill Medical College of Cornell University,
New York (USA)
First author of section 3.3.5

Frank Sivo,
Columbia University,
New York (USA)
Author of section 3.2.4.1

Michael Snyder,
Yale University,
New Haven (USA)
Corresponding author of section 7.8.23

Jo-Ann L. Stanton,
University of Otago,
Dunedin (New Zealand)
First author of section 4.4.6

Christian J. Stoeckert, Jr.,
University of Pennsylvania,
Philadelphia (USA)
First and corresponding author of section 7.8.18

Dominic G. Spinella,
Chugai Biopharmaceuticals, Inc.,
San Diego (USA)
Exclusive author of section 4.4.8

Michael P. Starkey,
Human Genome Mapping Project Resource
Centre, Wellcome Trust Genome Campus,
Hinxton (UK)
First author of section 3.3.6

Katrin Stolle,
University of Münster,
Münster (Germany)
First author of sections 3.2.4.3 and 3.4.2

Johannes Streicher,
University of Vienna,
Vienna (Austria)
*First and corresponding author of sections 7.7
and 7.8.6*

Zao-zhong Su,
Columbia University,
New York (USA)
Author of sections 3.2.4.1, 3.2.4.6 and 3.2.4.8

J. Gregor Sutcliffe,
The Scripps Research Institute,
La Jolla (USA)
Senior author of section 4.3.8

Arthur J. Sytkowski,
Harvard Medical School,
Boston (USA)
Senior and corresponding author of section 3.2.4.5

Dmitrij Tchekmenev,
BIOBASE GmbH,
Wolfenbüttel (Germany)
Author of section 7.6

Kenneth D. Tew,
Fox Chase Cancer Center,
Philadelphia (USA)
Senior and corresponding author of section 4.3.1

Kim Theilgaard-Mönch,
Rigshospitalet,
Copenhagen (Denmark)
First and corresponding author of section 3.3.14

Nick Tosches,
School of Medicine, Yale University,
New Haven (USA)
Author of section 7.8.23

Chihiro Uematsu,
Central Research Laboratory, Hitachi Ltd.,
Tokyo (Japan)
First and corresponding author of section 3.3.8

Anton-Jan van Zonneveld,
Academic Medical Center,
University of Amsterdam,
Amsterdam (The Netherlands)
Author of section 4.4.2

Richard G. F. Visser,
Wageningen University and Research Centre,
Wageningen (The Netherlands)
Author of section 3.3.3

David J. Volsky,
St. Luke's-Roosevelt Hospital Center,
Columbia University,
New York (USA)
Author of section 3.2.4.6

Erno Vreugdenhil,
Leiden University Medical Centre,
Leiden (The Netherlands)
Senior author of section 4.4.3

Annette Wagenhaus,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Gerd Wagner,
Clondia Chip Technologies GmbH,
Jena (Germany)
Author of sections 4.2.1.4 and 4.2.1.4.4

Jamie Walden,
GenHunter Corporation,
Nashville (USA)
Author of section 4.3.3

Eldon M. Walker,
Lerner Research Institute,
Cleveland Clinic Foundation,
Cleveland (USA)
Author of section 7.8.13

Yuxun Wang,
Beth Israel Deaconess Medical Center,
Harvard Medical School,
Boston (USA)
First author of section 3.2.4.5

Zhou Wang,
Feinberg School of Medicine,
Northwestern University,
Chicago (USA)
First and corresponding author of section 3.2.4.4

Sachihiko Watanabe,
Genomic Sciences Center,
Yokohama Institute,
Yokohama (Japan)
Exclusive author of section 3.4.1

Ralph Weissleder,
Massachusetts General Hospital,
Charlestown (USA)
*Senior and corresponding author
of sections 6.4 and 6.8*

Stephen Welle,
University of Rochester,
Rochester (USA)
Exclusive author of section 7.8.19

Shuai Wenig,
Stanford University,
Stanford (USA)
Author of section 7.8.21

Claudia Wenner,
University of Münster,
Münster (Germany)
First author of section 1.7

Michel Werner,
Service de Biochimie
et Génétique Moléculaire,
Gif-sur-Yvette (France)
Exclusive author of section 1.3.2 to 1.3.5

Bryan R. G. Williams,
Lerner Research Institute,
Cleveland Clinic Foundation,
Cleveland (USA)
Senior and corresponding author of section 7.8.13

Edgar Wingender,
BIOBASE GmbH,
Wolfenbüttel (Germany)
*Senior author of section 7.6; senior and
corresponding author of section and 7.8.5*

Rick Woychik,
Lynx Therapeutics, Inc.,
Hayward (USA)
First and corresponding author of section 4.4.1

Yan Ping Yu,
School of Medicine, University of Pittsburgh,
Pittsburgh (USA)
First and corresponding author of section 3.2.4.2

Martin Zeller,
University of Münster,
Münster (Germany)
Senior author of section 5.3

Kunbo Zhang,
Mayo Clinic and Foundation,
Rochester (USA)
Author of sections 3.2.4.11 and 3.3.10

1

Basic concepts of gene expression

1.1

Introduction

Following the publication of the draft sequence of the human genome in February 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001), the focus of many groups in academia and industry has shifted to the complex and demanding study of gene expression. Before discussing the methods for this endeavour, it is useful to review basic concepts. Experienced workers may wish to skip this section and move directly to the technique that interests them. However, the novice, or those from a non-biomedical background, may find this chapter useful.

The following pages include a description of how the information contained within the genes is expressed, *i. e.* an overview of the machinery and the regulation of gene transcription and protein translation in the eukaryotic cell. As everywhere within the book, there is an undisguised bias towards the situation in the human. We make no apology for this as we feel that many of our readers are likely to approach the book from a biomedical perspective.

This description of the mechanism of gene expression is rounded off by a brief classification of the technologies for differential gene expression that make up the remainder of the book.

At present, there is a lively debate between the proponents of mRNA expression analysis and the champions of protein analysis as to which method provides the truest picture of functional gene expression. The mRNA camp points to the very high degree of sensitivity, the speed and the completeness of their methodology. Aficionados of protein expression, by contrast, point to the obvious discrepancies between the gene complement of an organism and its ultimate phenotype. Thus, the butterfly and the caterpillar have identical genomes but radically different proteomes. We believe that this debate has produced much heat but little light. The reality is that the study of mRNA expression and protein expression are complementary techniques, each of which has an irreplaceable contribution to make to the understanding of gene expression. Indeed, this logic also applies to the invidious comparisons which are sometimes drawn between individual methods. Here too, the truth is that each method has its own particular usefulness and that the choice of method depends on the question being asked and the financial and temporal resources available to the investigator.

1.2

Basics of transcription and translation in the cell

1.2.1

Introduction

Proteins play important roles in nearly all biological processes. They catalyse chemical reactions, are crucial for the structure and stability of cells, regulate cellular uptake and storage of other essential molecules, and are responsible for coordinated growth and development. Thus, the synthesis of proteins is of prime importance for all cells. The structure of proteins is defined in the sequence of bases in the DNA. However, the DNA is not translated into proteins directly (figure 1.1). Instead, in a process called *transcription* a molecule of messenger RNA (mRNA) is copied from the genomic DNA. A single strand of mRNA then leaves the nucleus and serves as a blueprint for the assembly of a protein. This process, called *translation*,

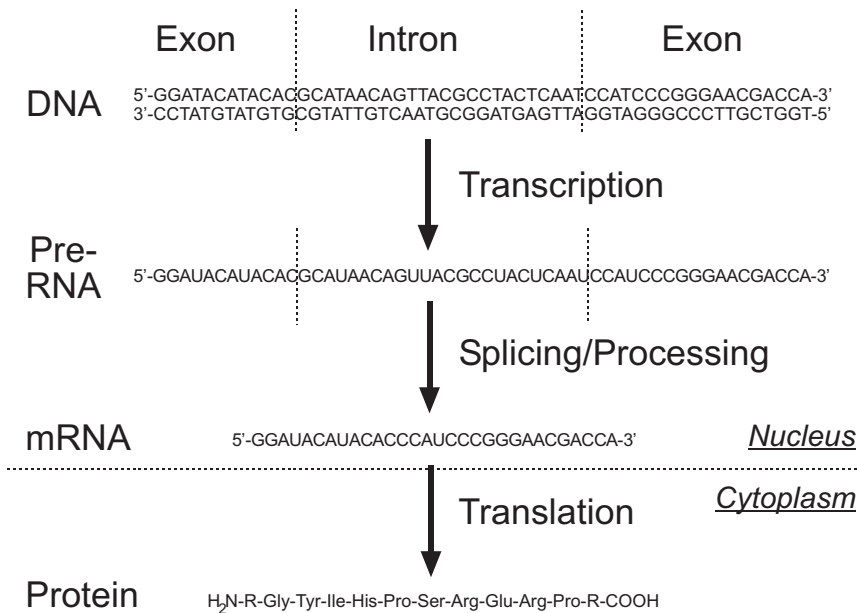


Figure 1.1 Schematic overview of the flow of the genetic information. The genetic code, which is stored within the sequence of bases within the DNA, is the blueprint for the assembly of proteins. However, DNA is not immediately translated into proteins. First, a copy of one DNA strand is made in a process called transcription, which leads to a primary RNA transcript (preliminary RNA, preRNA). This RNA is identical to the coding strand of the DNA, except that thymine is exchanged for

uracil. Then, the non-coding parts of the RNA (introns) are resected in a process called splicing, which results in the formation of messenger RNA (mRNA), which is modified by the addition of a cap at the 5'-end and a poly(A) tail at the 3'-end. The mRNA is translocated into the cytoplasm, where it is translated into protein by special organelles called ribosomes. 'R' indicates the remaining amino acids of the protein.

is mediated by ribosomes, small organelles consisting of several subunits of ribosomal RNA (rRNA), which are preloaded with auxiliary protein factors. By the action of ribosomes, the sequence of nucleotides in the mRNA molecule is translated into a corresponding sequence of amino acids to produce a protein chain. Each of the possible pool of 20 different amino acids from which the protein is to be built is brought to the ribosome by its specific transfer RNA (tRNA) molecule. This chapter outlines the basic concepts of transcription and translation, with a focus on the mechanisms in eukaryotes. The similarities and differences between eukaryotes and prokaryotes will be summarised at the end of the subchapters.

1.2.2

Transcription

Transcription generates both the mRNAs, which are the blueprint for the assembly of proteins, and the different types of RNA that are needed for RNA processing and protein synthesis (small nuclear RNA (snRNA), tRNA, rRNA). These RNA molecules are synthesised by an enzyme called RNA polymerase, which copies the sequence of one strand of DNA, the so-called coding strand. The enzyme binds tightly to specific DNA sequences called promoters, which indicate the starting point for RNA synthesis. Promoters are characterised by consensus sequences, *i. e.* typical sequences of bases within the DNA, which are in special positions relative to the coding region (figure 1.2). By means of these promoters, RNA polymerase recognises the starting point for transcription. After binding to the promoter, the enzyme starts transcription by first unwinding a region of the double helix and subsequent coupling of two ribonucleotides, which must be complementary to the template DNA strand. The RNA polymerase molecule then assembles the mRNA strand by moving along the DNA, unwinding the helix, and adding one complementary nucleotide at a time (figure 1.3). This results in the transient formation of a RNA/DNA hybrid helix. This helix, however, is relatively short – about twelve bases – because the assembled RNA chain is immediately displaced, allowing the DNA to rewind at the rear of the RNA polymerase.

The elongation process continues until the enzyme encounters a termination signal in the DNA, where the polymerase stops and releases both the DNA and the newly made RNA chain, which is typically between 70 and 10,000 nucleotides long. Little is known about termination signals in eukaryotes. Eukaryotic preliminary RNA is cleaved by a specific endonuclease, which recognises a cleavage signal containing the consensus sequence AAUAAA. The endonuclease cleaves behind the uracil residue, and a polymerase adds about 250 adenylate residues, resulting in a polyadenylate tail at the 3'-end. This poly(A) tail is thought to protect the mRNA from digestion by nucleases and also appears to play a role in promoting translation. The RNA resulting from transcription is an identical copy of the coding strand of the DNA, except that thymine is replaced by uracil.

Besides smaller differences, for example in the exact position and consensus sequences of the promoters, there are also some more fundamental differences in transcription between eukaryotes and prokaryotes. In prokaryotes, all mRNA

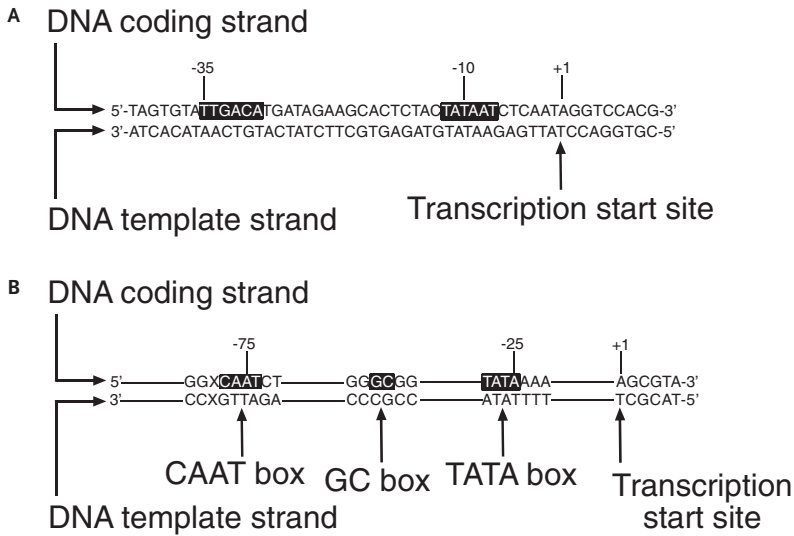


Figure 1.2 Prokaryotic A) and eukaryotic B) promoters. In bacteria, two sequences on the 5'-side of the first nucleotide to be transcribed into mRNA, which is defined as +1, are important. One of these is positioned around the -10 position and has the consensus sequence TATAAT. The other is found at position -35 with the consensus sequence TTGACA. Eukaryotic genes have promoter sites with a TATAAA con-

sensus sequence centred at about position -25, and often also a CAAT box with a GGXCAATCT consensus sequence positioned at about position -75. Constitutive genes often also show GC-rich sequences in their promoters. The CAAT and GC boxes can affect promoter activity when positioned on either the coding or the template strand.

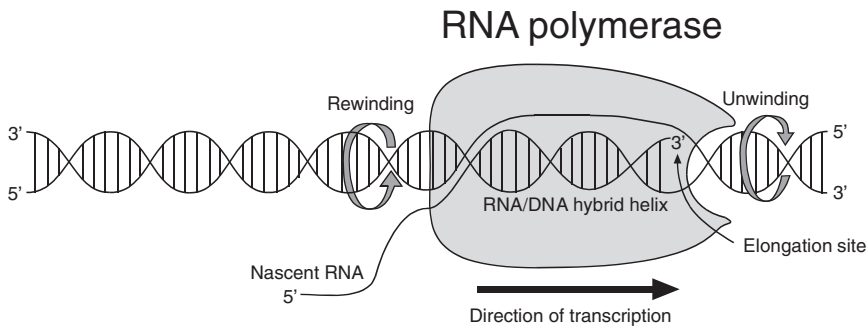


Figure 1.3 Basic mechanism of transcription. After binding to the promoter, RNA polymerase starts transcription by first unwinding a small piece of DNA. Then, the RNA strand is assembled by adding one complementary nucleotide at a time as the RNA polymerase moves along the DNA. This results in the formation of an RNA/DNA hybrid helix, which is

relatively short due to the fact that the RNA chain is immediately displaced, allowing the DNA to rewind again at the rear of the enzyme. This process continues until the enzyme encounters a termination signal in the DNA, where the polymerase stops and releases both the DNA and the newly synthesised RNA. The figure is adapted from Stryer (1995).

is produced by a single RNA polymerase, while in eukaryotes three distinct RNA polymerase enzymes exist. Furthermore, transcription and translation take place simultaneously in prokaryotes, *i. e.* translation begins before the whole mRNA molecule is transcribed. By contrast, transcription and translation are spatially and temporally separated in eukaryotes, which enables more subtle regulation of gene expression. In contrast to prokaryotes, where mRNA molecules are produced directly by DNA transcription, most primary RNA transcripts in eukaryotic cells are altered extensively. These alterations include the addition of a 7-methylguanosine cap at the 5'-end, the attachment of the poly(A) tail at the 3'-end, and most importantly a process called splicing, which is the resection of non-coding parts of the RNA (introns) and the joining of the coding parts (exons) to form a continuous sequence that codes for a specific polypeptide chain. Splicing is also of interest with regard to the regulation of gene expression, because many primary RNA transcripts in eukaryotes can be processed in more than one way to produce different mRNAs and thus different polypeptides. The primary transcript contains molecular signals for alternative processing pathways, and the pathway favoured in a given cell is determined by special RNA-binding proteins. The exact mechanisms of this so-called alternative splicing and its consequences for gene expression will be dealt with in a later chapter of this book.

1.2.3

Translation

Cells translate the sequence of the bases into an amino acid sequence by means of codons, which are triplets of nucleotides on the mRNA that specify a single amino acid (table 1.1). Four different nucleotides (adenine, cytosine, guanine and uracil) exist, thus, any series of the three nucleotides of a codon can result in $4^3 (= 64)$ possible combinations. Since there are only 20 different amino acids, it is clear that some or all amino acids must be specified by several codons. In fact, most amino acids are coded by between two and six different codons. Only tryptophan and methionine are coded by a single codon. Three of the 64 codons do not code for amino acids but act as termination signals (stop codons). How do cells translate the sequence of these codons into proteins? This is done by the use of a set of tRNAs, which are small RNA molecules characterised by two major binding domains: one, the anticodon, is specific for a codon in the mRNA, the other for the amino acid specified by that codon. Each tRNA is designed to carry only one of the 20 amino acids used for protein synthesis. Before an amino acid is incorporated into a polypeptide chain, it is attached by its carboxyl end to the 3'-end of an appropriate tRNA molecule (figure 1.4). This attachment covalently links the amino acid to a tRNA containing the correct anticodon. Furthermore, this amino acid attachment generates a high-energy linkage at the carboxyl end, thereby facilitating its reaction with the amino group of the next amino acid in the protein sequence to form a peptide bond. As the sequence of proteins within the synthesised protein is solely determined by the tRNA molecule, and not by its attached amino acid, the accuracy of protein synthesis is crucially dependent on the accuracy of the me-

Table 1.1 The genetic code.

1 st position	2 nd position				3 rd position
	U	C	A	G	
U	Phenylalanine	Serine	Tyrosine	Cysteine	U
	Phenylalanine	Serine	Tyrosine	Cysteine	C
	Leucine	Serine	STOP	STOP	A
	Leucine	Serine	STOP	Tryptophan	G
C	Leucine	Proline	Histidine	Arginine	U
	Leucine	Proline	Histidine	Arginine	C
	Leucine	Proline	Glutamine	Arginine	A
	Leucine	Proline	Glutamine	Arginine	G
A	Isoleucine	Threonine	Asparagine	Serine	U
	Isoleucine	Threonine	Asparagine	Serine	C
	Isoleucine	Threonine	Lysine	Arginine	A
	Methionine	Threonine	Lysine	Arginine	G
G	Valine	Alanine	Aspartic acid	Glycine	U
	Valine	Alanine	Aspartic acid	Glycine	C
	Valine	Alanine	Glutamic acid	Glycine	A
	Valine	Alanine	Glutamic acid	Glycine	G

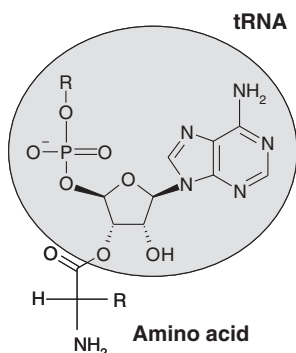


Figure 1.4 Structure of an aminoacyl tRNA. Before being incorporated into the polypeptide chain, an amino acid is linked with its carboxyl end to the 3'-end of its specific tRNA. The letter R indicates any possible residue of the specific amino acid.

chanism that links the amino acid specifically to its corresponding tRNA. This accuracy is guaranteed by 20 different enzymes called aminoacyl tRNA synthetases.

Protein synthesis requires a complex catalytic machinery. Above all, it must be ensured that each successive codon in the mRNA engages precisely with the anticodon of the tRNA molecule and does not slip by even one nucleotide. This precise movement is catalysed by the ribosomes, which are large complexes composed of one large and one small rRNA subunit as well as of several proteins. A ribosome contains three binding sites for RNA molecules: one for mRNA and two for tRNAs. One site, called the peptidyl tRNA-binding site, or P-site, holds the tRNA molecule

that is linked to the growing end of the polypeptide chain. Another site, called the aminoacyl tRNA-binding site, or A-site, holds the incoming tRNA molecule charged with an amino acid (figure 1.5). The A and P-sites are adjacent to each other so that the anticodons of two tRNAs bound here form base pairs with neighbouring codons in the mRNA molecule.

During the initiation phase of protein synthesis, the large and small subunits of the ribosome are brought together at the exact spot on the mRNA where the polypeptide chain is to begin. For this purpose, an initiator tRNA molecule bound to methionine, which is the amino acid encoded by the start codon (AUG), is required. This tRNA binds to the P-site of the small ribosomal subunit, and then binds to the first AUG start codon on the mRNA that is in a particular context – the highly conserved Kozak sequence with the consensus CCACCaugG – (Peri & Pandey, 2001; Kozak 1987, 1991 and 2000), thereby bringing the ribosome into exactly the correct position. As soon as the mRNA and the first tRNA are bound, the assembly of the ribosome is completed by the binding of the large ribosomal subunit.

The elongation of the polypeptide chain is a reiterating three-step cycle. After translation has been started by the fixation of the initiator tRNA in the P-site of the ribosome, the aminoacyl tRNA molecule with the fitting anticodon becomes bound to the ribosomal A-site by forming base pairs with the codon exposed at the A-site (step 1). A tRNA molecule is held tightly at either site only if its anticodon forms base pairs with a complementary codon on the mRNA molecule that is bound to the ribosome. Then, a peptide bond is formed by a reaction between the amino group of the amino acid linked to the tRNA in the A-site and the carboxyl group of the polypeptide chain that is bound to the tRNA molecule in the P-site (step 2). This reaction is catalysed by an enzyme called peptidyl transferase, which is an enzymatic function of the large ribosomal subunit. The free tRNA bound to the P-site is then released, and the new peptidyl tRNA in the A-site is translocated to the P-site as the ribosome moves exactly three nucleotides further along the mRNA molecule in the 3'-direction (step 3). After completion of this last step, the A-site is free to accept a new tRNA molecule, which starts the cycle again.

Protein synthesis is terminated when so-called release factors bind to a stop codon in the A-site. This binding alters the activity of the peptidyl transferase, causing it to catalyse the addition of a water molecule instead of an amino acid to the peptidyl tRNA. This reaction releases the carboxyl end of the growing polypeptide chain from its tRNA molecule, thus discarding the complete protein into the cytoplasm. The ribosome then also releases the mRNA and dissociates into its two separate subunits.

The major difference between prokaryotic and eukaryotic translation is that prokaryotic mRNA can have multiple start sites and can serve as the template for the synthesis of several proteins, while each eukaryotic mRNA is the template for only one protein. In addition, slight differences exist between eukaryotes and prokaryotes in the size of the ribosomes, the nature of the initiator tRNA, and the number and structure of initiation, elongation and termination factors.

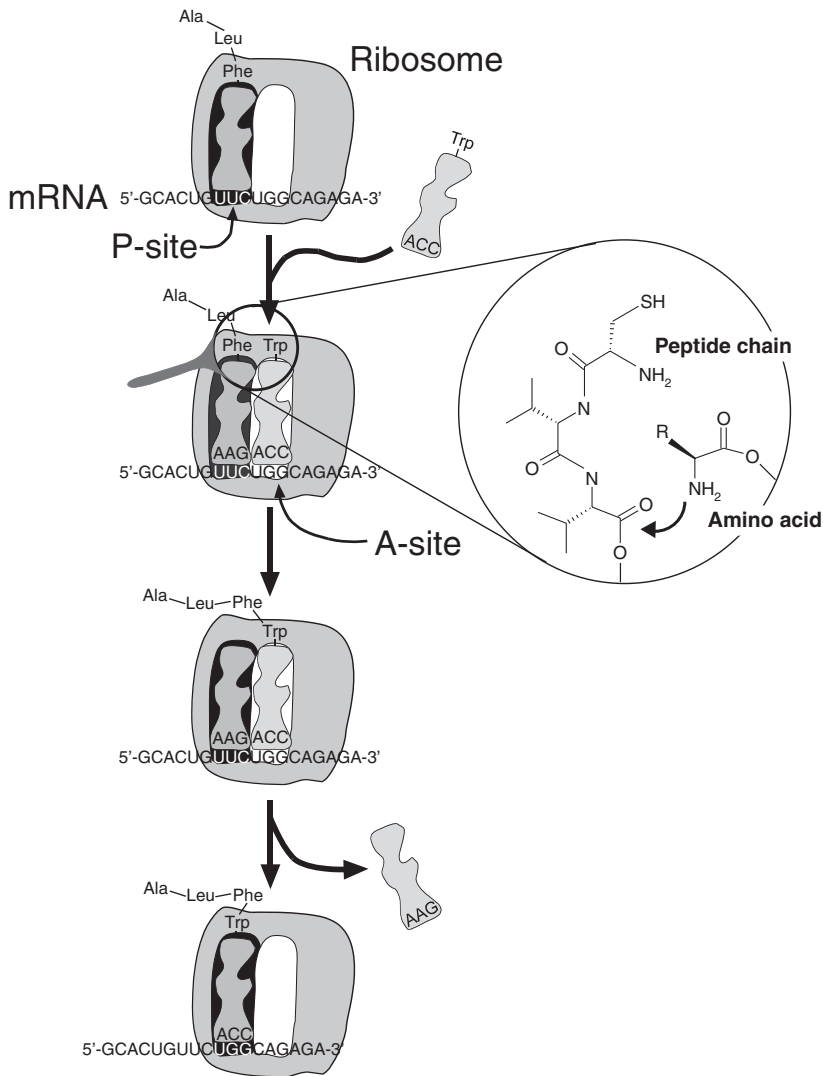


Figure 1.5 Schematic overview of translation.

An mRNA is bound to a ribosome, in this example with the codon UUC exposed to the P-site and the codon UGG to the A-site. The peptidyl tRNA is already bound to the P-site (the last amino acid added to the chain is phenylalanine). In the first step, the tRNA with the anticodon to UGG, which is specific for tryptophan, binds to the A-site. The free amino

group of tryptophan then reacts with the carboxyl group of the peptide chain bound to the tRNA in the P-site (step 2). The polypeptide chain is elongated by the formation of a new peptide bond, and the polypeptide chain is now attached to the tRNA at the A-site. This leads to the release of the free tRNA at the P-site (step 3), and translocation of the peptidyl tRNA from the A-site to the P-site together with movement of the ribosome exactly three nucleotides further along the mRNA.

1.2.4

Summary

Genetic information is made up of the sequence of bases in the DNA. This information determines the composition, structure, and above all the function of the cellular proteins. In both prokaryotes and eukaryotes a large and complex cellular machinery is responsible for the synthesis of proteins, which involves in a first step the transcription of DNA into mRNA, and in a second step the translation of mRNA into proteins. These processes have been briefly outlined in this chapter. However, the regulation of transcription and also translation, the mechanisms, which increase or decrease gene expression, allow the differentiation of cells in a multicellular body as well as the adjustment of a creature to changing environmental conditions. Also, these mechanisms provide insight into physiological and pathophysiological metabolic processes and may also serve as drug targets for drugs in the treatment of disease. The regulation of transcription and translation will be dealt with in detail in later chapters.

1.3

Regulation of transcription

1.3.1

Introduction

Gene expression in eukaryotic cells is be regulated at various stages on the path from DNA to the active protein. At least eight key points of regulation of gene expression have been identified up to now (Alberts *et al.*, 1998; Carey & Smale, 2000). These are (i) transcription initiation, (ii) transcript elongation, (iii) RNA processing (*e.g.*, splicing, polyadenylation), (iv) RNA transport from the nucleus to the cytoplasm, (v) RNA stability (*e.g.*, by additional polyadenylation), (vi) translation of mRNA into protein, (vii) degradation of proteins, and (viii) post-translational modification of proteins (figure 1.6). The following sections describe the principles of the regulation of gene expression at each of these levels.

1.3.2

mRNA expression profiles – the transcriptome

Transcription is the first and main level of regulation of gene expression. Protein synthesis rates generally correlate with transcription rates except for a small number of genes for which translation is regulated at a post-transcriptional level. Variations in the ensemble of the mRNA molecules of a cell, the transcriptome, is largely reflected in equivalent variations in the population of the cell proteins, the proteome (the term ‘proteome’ was invented by Marc Wilkins and Keith Williams at the Macquarie University Center for Analytical Biotechnology in Australia in 1995). However, the amount of a specific mRNA is usually not directly related to

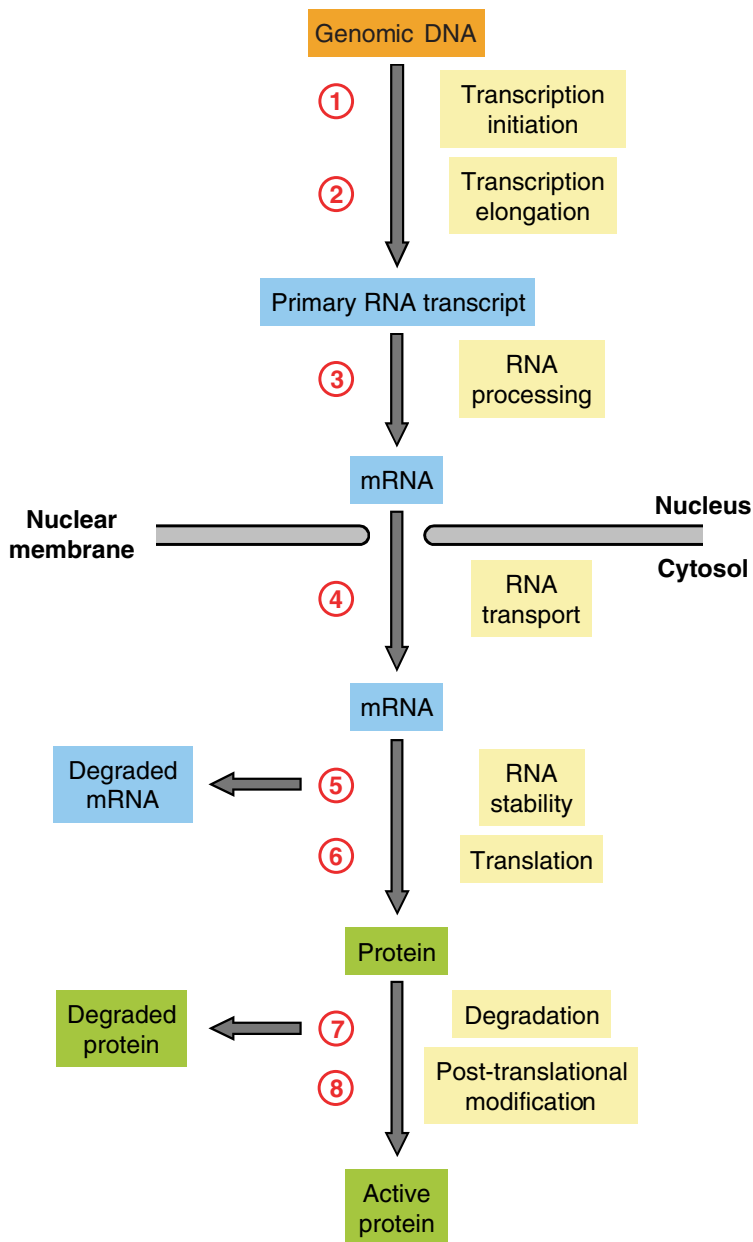


Figure 1.6 Levels of modulation and control of gene expression and protein activity in eukaryotic cells. The figure is adapted and modified from Alberts *et al.* (1998).

the amount of the corresponding protein in the cell because the translational efficiency of a given mRNA is not a constant and because the half-life of mRNAs varies widely from one mRNA species to another. Transcriptome analysis is now possible thanks to the knowledge of gene sequences and the development of different technologies such as serial analysis of gene expression (SAGE), DNA microarray technology, and real-time RT-PCR (Bustin, 2000; Chee *et al.*, 1996; DeRisi *et al.*, 1997; Velculescu *et al.*, 1995).

Transcriptome analysis may be used to assign a function to genes. The transcriptional state of the cell in a given condition provides a signature that can be compared to that of cells in other growth or physiological conditions (Eisen *et al.*, 1998; Hughes *et al.*, 2000; Spellmann *et al.*, 1998). These signatures can be used to group genes on the basis of their response to a stimulus. Genes that belong to a common pathway are often regulated in a similar way. Thus similarities in expression patterns may be used to assign function to unknown genes.

Many studies using DNA microarrays were performed in the *Saccharomyces cerevisiae* model system. This organism was used since it was the first eukaryotic genome to be sequenced, it is highly amenable to genetic analysis. Two typical examples of the use of transcriptome analysis to investigate cell function were the analysis of the cell cycle-regulated genes and of the transcriptional response of the yeast to various stresses. In the former example, yeast cells were synchronised at different points along the cell cycle and then released and allowed to resume progression within the cycle (Cho *et al.*, 1998; Spellmann *et al.*, 1998). About 15 percent of the yeast genes showed waves of transcription that peaked at particular times during the cell cycle. In this way numerous new genes were implicated in the cell cycle. The response of the yeast transcriptome to such stressors as heat or cold, or to chemical challenges such as diamide or hydrogen peroxide, was also investigated (Causton *et al.*, 2001; Gasch *et al.*, 2000). Strikingly, all these stressors were found to repress strongly and co-ordinately the transcription of around three hundred genes which encode proteins necessary for the constitution of the translation machinery. These included ribosomal genes, genes encoding the subunits of RNA polymerase I and III, proteins involved in ribosomal RNA (rRNA) and transfer RNA (tRNA) maturation, as well as factors involved in the initiation and elongation of translation. A number of genes of unknown function were also repressed by these stressors, at least some of which may encode proteins required for the constitution of the translation apparatus.

Compendia of transcription profiles may be accumulated for analysis of transcription profiles (Hughes *et al.*, 2000). These profiles observed in various growth conditions, under the influence of compounds or drugs, and in different mutants can be arranged or clustered according to their transcriptional signature.

Transcriptome analysis provides a powerful method to evaluate the side effects of a drug. Under ideal circumstances, a drug should not perturb pathways other than the one containing its target. Thus, in principle, a drug might be concluded to be free of side-effects if deletion of the drug target abolishes the transcriptional response to the drug (Marton *et al.*, 1998). A future application of transcriptome anal-

ysis might be to tailor treatments to obtain the best response and minimise the adverse effects of drugs in individual patients.

Transcriptome analysis may also provide an important contribution to medicine, particularly in understanding biology of cancer and improving the diagnosis and treatment of this condition. Current classification methods based on morphology often fail to distinguish between tumors that may have a very different prognosis. Microarray experiments of particular lymphoid leukaemia, for example, have been able to distinguish subtypes with widely different prognosis where classical morphology failed to do so (Alizadeh *et al.*, 2000). Moreover, the analysis indicated that the two kinds of leukemia originated from different cell types to which they had a closer resemblance than they had to each other. Being able to more finely classify different types of cancers and assess more precisely their sensitivity to drugs will help design more powerful treatments with fewer side effects.

An area of great interest concerns the use of mathematical tools to analyse patterns of gene expression and thus to identify the regulatory networks linking together the transcription of groups of genes (see also chapter 7). This technique has been proved in principle in the study of mutants of transcription factors in microorganisms. However, it is difficult to ascertain if the effect of a mutation is direct or indirect when the only data to hand is the pattern of gene transcription. One computational solution to this dilemma is to search for potential DNA-binding sites for transcription factors in the promoter regions of co-regulated genes (Spellmann *et al.*, 1998). Alternatively, one can use the genes that respond to a particular factor as a training set in order to define the sequence of the binding site for the transcription factor and then look for the presence of this newly defined sequence in the promoter regions of the regulated genes.

A more powerful approach has been devised recently that combines the chromatin immunoprecipitation approach (ChIP) with microarray technology (Iyer *et al.*, 2001; Lieb *et al.*, 2001; Ren *et al.*, 2000). Proteins bound to DNA are cross-linked with formaldehyde *in vivo* and the factor of interest is immunoprecipitated after random shearing of DNA by sonication. The cross-link is disrupted and the DNA that has been enriched by immunoprecipitation is amplified, labelled and hybridised to microarrays bearing the complete genome of interest in competition with chromatin which has been immunoprecipitated from a mutant in which the factor of interest is unable to bind. This procedure will reveal the binding site for the factor along the genome. In addition, the 'ChIP on chip' data can now be compared with transcriptome data in order to reveal which of the targets of the factor are directly responsive to its presence and which genes are regulated indirectly. A striking example of the possibilities of this approach was provided by a study of the regulators of the different phases of the yeast cell cycle (Simon *et al.*, 2001). In this study, it was found that the regulators of a given phase of the cell cycle not only regulated transcription of the genes expressed specifically in that phase but also regulated the regulator of the next phase in the cycle. The organisation of the cell cycle was thus mirrored in a cycle of regulators.

In 2001, two versions of the draft of the human genome were published (Lander *et al.*, 2001; Venter *et al.*, 2001). The two projects used different algorithms to predict the number of genes and arrived at roughly similar figures in the 30,000 range. However, comparison of the two sets of putative genes revealed that only half of the predicted genes were common to both projects and even in those cases, intron predictions varied (Hogenesch *et al.*, 2001). This fact shows how important it is to verify by experiment the presence and structure of the predicted genes. The gene organisation together with the sites of initiation and termination of transcription can be obtained using tiled oligonucleotide probe microarrays as has been done for a 150 kilobase pairs segment of human chromosome 22 (Shoemaker *et al.*, 2001). Microarrays can thus also be used as a genome annotation tool.

1.3.3

Protein expression profiles – the proteome

As stated previously, most of the cellular regulation of gene expression is exerted at the transcriptional level. However, in some cases, as with the yeast *GCN4* gene, a transcriptional activator, or *CPA1* gene, which encodes the glutaminase subunit of the arginine pathway carbamoyl-phosphate synthetase, protein amounts do not reflect mRNA levels (McCarthy, 1998; Sachs & Buratowski, 1997). Regulation can be effected at several steps after transcription. For example, the stability of the mRNA, its translation or even the stability of the protein may be regulated. Proteins may also be post-transcriptionally modified by, for example, glycosylation, phosphorylation, methylation, sulphation or ubiquitination. Such modifications cannot be inferred from the sequence of the proteins but are very important because they can greatly affect the activity of the protein. Again, transcriptome analysis cannot reveal this aspect of cell biology.

Knowing where a protein is located in the cell can help tremendously in determining what it does. It is possible to fractionate organelles such as mitochondria, chloroplasts, the Golgi, vacuoles or the nucleus, and to describe the complete set of protein contained within them. Proteome analysis coupled with subcellular fractionation can therefore provide answers to important biological questions. Traditionally, proteome analysis has relied on two-dimensional separation of proteins. In a first dimension, the proteins are focused according to their isoelectric point. They are then separated in the second dimension by their apparent molecular mass. With the recent developments of mass spectrometry together with the advent of the sequence of genomes, it has been possible to devise methods that allow identification of hundreds of proteins on two-dimensional gels (Pandey & Mann, 2000). These methods are also able to detect some protein modifications (see chapter 5).

Although proteome analysis using two-dimensional gels can at best reveal the 500 to 1,000 most abundant proteins present within a sample, there are many situations where it cannot be replaced by transcriptome analysis. This is true, for example, of samples which are devoid of mRNAs, for example body fluids such as cerebrospinal fluid, lymph or plasma. In addition, if one is interested only in the most abundant proteins, two-dimensional electrophoresis may be a

simpler and more reliable means to measure expression ratios than microarray analysis. This is particularly true of the study of protein expression in microbes where spot identifications is simpler due to the lower complexity of the protein samples (Maillet *et al.*, 1996).

The knowledge of the genome sequence of eukaryotes is an important resource. However, as stated above, particularly in multicellular organisms the gene structure is difficult to predict from pure sequence information. Moreover, alternative transcription initiation or splice sites can be used in different tissues and during development. Hence, the automatic annotation of genomic sequences cannot be considered as valid unless experimental data support it. Automated proteomic analysis, coupling two-dimensional separation of proteins with the mass spectrometric identification of the spots, can provide a reliable genome annotation method (Pandey & Mann 2000).

1.3.4

Interaction between genes and proteins – the interactome

Interactions between proteins play important roles in the cell biology. Numerous proteins participate in complexes and the knowledge of the interaction in which they are engaged is essential to understand how these machines function. Some of the interactions are transient, as is the case between a protein kinase and its substrate. Another good example of the importance of protein/protein interactions is in the transcription of messenger RNA in eukaryotes and its precise regulation. This process requires more than 70 polypeptides that belong to a dozen molecular machines also called transcription factors (Lee & Young, 2000) (see below). The arrival and departure of these machines from transcribed genes is the result of a precise choreography in which ordered protein contacts culminate in the initiation of transcription by RNA polymerase II. The structure of this twelve-subunit enzyme has recently been solved (Cramer *et al.*, 2000 and 2001; Gnatt *et al.*, 2001). Crystallographic analysis cannot however be used as a routine method to solve the structure of complexes. Other techniques are thus required that will ideally give a precise description of the interactome, cataloguing the identity of the interacting polypeptide, the extension of the interaction domains, and the timing and sub-cellular context in which the interactions take place.

The yeast two-hybrid method and its variants is to date the method of choice to screen for protein/protein interactions at the proteome level. This situation may change in the future with the advent of so-called protein chips. It is based on the observation that the DNA-binding domain and the activation domain of a eukaryotic transcription activator when fused to two interacting proteins will activate the transcription of a reporter gene in yeast. Thus, one can ask if two proteins interact by fusing them to the activation domain and the DNA-binding domain of a transcription factor respectively and by then following the expression of a reporter gene when they are co-expressed in yeast (Fields & Song, 1989). The method can be used to test individual proteins in a pairwise manner or to screen DNA libraries (Legrain *et al.*, 2001; Vidal & Legrain, 1999). These libraries are made of full-length

complementary DNAs (cDNAs) or of random DNA fragments fused to the activation domain of the transcription activator. Screening the whole set of proteins of an organism with such libraries is extremely labour intensive. Nevertheless, the procedure can be applied to small proteomes such as those of viruses (Bartel *et al.*, 1996; Flajolet *et al.*, 2000), to illuminate specific parts of complex proteomes (Wallhout *et al.*, 2000) or to explore proteomes of intermediate complexity like those of bacteria at a lower resolution (Rain *et al.*, 2001). In addition, the use of DNA fragments libraries provides a map of the interaction domains. A systematic study of the yeast RNA polymerase III complex disclosed a number of contacts that were latter described in much more details by the crystallographic analysis of the yeast RNA polymerase II indicating that the method can also be successfully applied to investigate the structure of complexes (Flores *et al.*, 1999).

The yeast two-hybrid system has been used to explore the yeast proteome by testing individual fusions of almost all the proteins fused to the activation domain or the DNA-binding domain of the *Gal4p* activator (*Gal4p* is one of the best described yeast transcription activators; it is required for the activation of transcription of the genes involved in the utilisation of galactose in that organism) (Ito *et al.*, 2000 and 2001; Uetz *et al.*, 2000). Three different medium or high-throughput studies showed few overlaps in their results. The disconcerting result may indicate that the screening procedures were not performed at saturation level. A more disturbing implication is that a substantial proportion of the reported interaction may in fact be false-positive. This situation may, for example, result from the fact that, in the two-hybrid system, proteins that are normally segregated in different sub-cellular compartments are co-expressed in the nucleus. In support of the general value of the yeast two-hybrid system, however, it has been found through bioinformatic analysis of the screening data that there is a general correlation between co-expression of genes and interaction between proteins (Ge *et al.*, 2001). Moreover, as expected, proteins interact in the yeast two-hybrid system more often with partners belonging to the same functional category than with partners from other categories. Despite its drawbacks, the yeast two-hybrid system is currently the only procedure able to provide even a glimpse at the interactome in an *in vivo* context for any organism.

Knowledge of the interaction partner of a given protein is valuable information. However, the two-hybrid system does not allow the identification of members of a protein complex since it usually probes for direct interactions. Recently, it was shown that coupling affinity purification of proteins followed by mass spectrometry after gel electrophoresis separation can be used at a proteomic scale for complex identification in yeast (Gavin *et al.*, 2002; Ho *et al.*, 2002). The purification procedures used either a single immunopurification of *in vivo* FLAG-tagged protein (FLAG is an eight amino acid marker peptide) or a tandem affinity purification method to identify hundreds of protein complexes. Though very powerful, this method is not applicable to any organism since it requires the insertion of a DNA sequence at the 3'-end of the gene coding the tagged protein in order to express the gene under its own promoter. Indeed, plasmid borne expression might bias the results. Nevertheless, the method can be used to query the existence of

complexes containing conserved proteins or more punctually to ask to which complex a particular protein belongs.

Advances in the production of recombinant proteins have allowed the production of extensive protein arrays (Zhu *et al.*, 2001). These arrays will certainly be used to investigate protein/protein interactions. However, the specificity of the *in vitro* tests that will use protein arrays and how it compares with the yeast two-hybrid system is not known presently and awaits further experiments.

1.3.5

The transcription machinery and core promoters

In eukaryotes, three distinct RNA polymerases are responsible for the transcription of the genome (see the following citations for reviews; Geiduschek & Kassavetis, 2001; Lee & Young, 2000; Paule, 1998; Weinzierl, 1999; White, 1998;). RNA polymerase I transcribes a single essential RNA, the ribosomal rRNA precursor (35S in yeast, 45S in higher eukaryotes). This RNA is transcribed from large clusters of tandem arrays of the genes that code for these RNA (rDNA) and that are located the nucleolus (or nucleoli). After transcription, the rRNA precursor is matured by the addition of methyl groups and pseudo-uridylation and processed into the two large ribosomal RNAs (18S and 28S in yeast) and the 5.8S rRNA. rRNA precursor transcription is extremely active since it represents 65 to 85 percent of all RNA synthesis in the cell. rRNA maturation and processing take place in the nucleolus and some of the steps of the pathway occur concurrently with the assembly of the ribosome. RNA polymerase II not only transcribes messenger RNAs, which are the subject of most of the studies on gene expression, but also a number of small stable RNAs like small nuclear RNAs (snRNAs) and most spliceosomal RNAs. These RNAs, on the contrary to mRNAs that are translated, are assembled in ribonucleoproteins and are responsible for the maturation of other RNAs. Finally, RNA polymerase III transcribes transfer RNA (tRNA) genes, the 5S ribosomal RNA and a few other stable RNAs like U6 spliceosomal RNA and some viral RNAs like VA from the adenovirus.

Transcription can be divided into steps. If one considers a gene that has never been transcribed, the first step is the recognition of the promoter by the transcription factors. While, for polymerase I and for most of polymerase III-transcribed genes, a set of basal factors is all that is required, polymerase II needs a very large number of gene-specific factors in addition to the basal transcription factors. The gene-specific transcription factors do not always communicate directly with the basal transcription machinery. Sometimes, they do so through the intermediate of co-activators that can be different complexes like chromatin remodelling machines or chromatin acetylation complexes (Jones & Kadonaga, 2000; Wolffe & Hayes, 1999). The complex of transcription factors with promoter DNA constitutes the pre-initiation complex. Each of the three RNA polymerases recognises its cognate pre-initiation complex and associates with it to form the initiation complex. The first step towards transcription initiation is the formation of the open complex that results in the separation of the two DNA strands around the transcription in-

initiation site. The RNA polymerase then transcribes, without leaving the promoter, short RNAs (three to ten nucleotides long) that are released. It then isomerises to enter the processive phase of transcription elongation. During the early phases of transcription, some of the basal transcription factors break loose from the promoter. On the other hand, the elongation complexes, formed of the DNA template, the transcribing RNA polymerase and the RNA are extremely stable and can be biochemically purified. Transcription elongation is not a monotonous process. Portions of a given gene are transcribed faster than others. RNA polymerases may even halt at specific locations along the DNA. Transcription elongation factors are sometimes required to promote transcription through the pause sites. At the end of the transcription unit the RNA polymerase will recognise the termination signal, release the RNA and separate from the template. After RNA release, the enzyme is recycled. Interestingly, a recycled RNA polymerase initiates preferentially on the template it has just transcribed and does so, at least *in vitro*, much faster than the first initiation (Dieci & Sentenac, 1996). It should be stressed that each of the steps during the transcription cycle can be regulated and that each of the numerous proteins participating in the process can be the target of activators or repressors.

RNA polymerases are the complex enzymes that synthesise RNA. In eukaryotes they are comprised of twelve subunits for polymerase II, 14 for polymerase I and 17 for polymerase III (Carles & Riva, 1998; Chédin *et al.*, 1998; Woychik, 1998). These subunits are conserved through evolution to the point that yeast subunits can, in many cases, be substituted by human subunits *in vivo* (Shpakovski *et al.*, 1995). The subunits can be subdivided into three categories: (i) Five small subunits are shared by the three RNA polymerases. (ii) Five subunits are homologous in the three enzymes. Among these subunits, one finds the two large subunits that are the hallmark of all multi-subunit RNA polymerases, including the eubacterial and archaeobacterial ones, and which form the catalytic site of the protein. (iii) The other subunits are specific to each RNA polymerase class and may or may not be conserved through evolution. Some of these subunits interact with the basal transcription factors during initiation complex formation and are thus responsible for the determination of class specificity of transcription.

As mentioned above, most of the transcriptional regulation occurring in the cell pertains to the synthesis of mRNAs by polymerase II. Nevertheless, global regulation of transcription by polymerases I and III results in the adaptation of the cell translational capacity to the growth rate. The basal polymerases I and III transcription factors will therefore be described briefly before introducing those factors that are needed for mRNA transcription and regulation.

Polymerase I pre-initiation complex formation in eukaryotes requires two multi-subunit factors that bind to the two promoter elements, the upstream control element (UCE; also termed upstream activator sequence, UAS) and the control element (CE). Though the subunit composition of the polymerase I factors and the sequences they bind to are not conserved from yeast to higher eukaryotes. In all cases where it has been investigated, the TATA box-binding protein (TBP) is present in the pre-initiation complex. Polymerase I recognises the pre-initiation com-

plex only when it is associated with the conserved *Rrn3p* transcription factor through an interaction with the polymerase I A43 subunit. It has been documented that most of polymerase I transcription regulation is mediated through the association of polymerase I with *Rrn3p* which happens when the cell is actively growing and which is prevented when the cell is quiescent.

Polymerase III RNA synthesis is epitomised by tRNA gene transcription that requires two transcription factors (TF), TFIIC and TFIIBs (TF), TFIIC (Chédin *et al.*, 1998; Geiduschek & Kassavetis, 2001). While the number of subunits varies between organisms in TFIIC, TFIIB consists of three conserved polypeptides, one of which is TATA box-binding protein (TBP) (TF, TFIIC). TFIIC is first recruited to the tRNA gene through interactions with two intragenic sequences termed the A and B-box. Then, TFIIB joins in, binding upstream of the transcription start site. Polymerase III recognises the TFIIB-related factor (BRF) subunit of TFIIB and forms the initiation complex. 5S RNA requires TFIIA in addition to TFIIC and TFIIB for transcription that can be thus considered as a specific transcription factor. U6 snRNA transcription details varies from organism to organism and will not be considered here. Biochemical studies suggested that the initiation complex is assembled sequentially. However, recent studies in human suggested that polymerase III, like polymerase II (see below) can be purified in association with its transcription factors and that the initiation complex can assemble in one step. *In vivo*, the arrival of the different components on the class III promoters has not yet been investigated.

Polymerase II requires a large set of basal transcription factors and co-activators comprising more than 50 polypeptides for the transcription of class II genes (Hampsey, 1998; Lee & Young 1998; Lemon & Tjian, 2000). This paragraph will not describe of the co-activators that are required to alter the chromatin structure. The basal transcription machinery comprises six general transcription factors (TFIIA, B, D, E, F and H) made of one to fourteen polypeptides. Together, they are sufficient to assemble a polymerase II pre-initiation complex but not to support transcription of chromatin templates. Order of addition studies have indicated that *in vitro* on naked DNA templates, the first basal factor to bind the promoter DNA is TFIID, which thus has a central role in pre-initiation complex formation. TFIID is composed of TBP together with 13 TBP-associated factors (TAF_{II}). TBP therefore plays a central role since it is present in pre-initiation complexes of all three RNA polymerases. However, recently, it has been found in multicellular eukaryotes that homologous proteins in specific situations can replace TBP (Dantonel *et al.*, 1999). After the formation of the TFIID/DNA or D complex, TFIIA and TFIIB join in to form the DAB complex which is an intermediate complex in the formation of the polymerase II initiation complex composed of DNA/TFIID/TFIIA/TFIIB. The crystallographic structure of TBP, TFIIA and TFIIB bound to DNA has been solved. The DNA is severely kinked by TBP binding to the minor groove. The DAB complex then recruits a polymerase II/TFIIF complex to form the DABF/polymerase II complex (the DABF complex comprises DAB plus TFIIF). The DABF/polymerase II complex is then recognised by TFIIE and then TFIIH, which has a dual role in transcription initiation and DNA mismatch repair (de

Laat *et al.*, 1999), to complete the formation of the polymerase II pre-initiation complex.

The largest subunit of polymerase II has, compared to those of polymerase I and III, is a carboxy-terminal extension consisting of more or less perfect repetitions of the YSPTSPT amino acid motif. The number of repetitions of the motif, which are organised in tandem repeats, varies from 25 in yeast to 52 in man and its sequence is conserved in eukaryotes. This carboxy-terminal domain (CTD) is extensively phosphorylated by three different cyclin-dependent kinases activated by divergent cyclins. One CTD phosphatase has been described. During the transcription cycle the phosphorylation state of the CTD varies with hypo-phosphorylation during transcription initiation, hyper-phosphorylation being the hallmark of the elongating enzyme. This modification is effected by the general transcription factor TFIIF.

Though some of the specific transcription factors may act by recruitment of TFIID, it has been observed that a CTD-binding complex, the mediator of transcription activation, is the main target of gene specific activators (Myers & Kornberg, 2000). This complex, comprising about 24 subunits, also interacts with polymerase II *via* other parts of the enzyme as documented by electron microscopic studies. The mediator is viewed as a co-activator since it links the gene-specific factors to the basal transcription machinery and relieves repression of transcription by chromatin. Though individual mediator subunit sequences are not very well conserved, the general structure of mediators is. Moreover, in multicellular eukaryotes, different mediators, responding to different gene specific activators, have been purified biochemically. These mediator forms have at least a core of common subunits. It is not clear yet if these complexes coexist in the cell or are the result of different purification protocols.

Once assembled, the initiation complex will undergo a number of changes that will allow transcription elongation to begin. The first step is the separation of the two DNA strands around the transcription initiation site to form the open complex. This step requires TFIIE and the ATP-dependent helicase activity of TFIIF. Initially, polymerase II repeatedly synthesises short RNAs, three to eight nucleotides long, without moving along the DNA template (as is also the case during transcription initiation by the other RNA polymerases). Then, the CTD is phosphorylated in an ATP-dependent manner by TFIIF and enters elongation. Of all the factors, the only one that remains associated with the promoter is TFIID. The other factors dissociate from the promoter during the transition from initiation to elongation, TFIIB is released after transcription of the first or second nucleotide, TFIIE separates from the promoter when polymerase II reaches nucleotide ten and TFIIF is released within a window of 30 and 40 nucleotides after the initiating nucleotide. TFIIF remains bound to polymerase II (Yudkovsky *et al.*, 2000).

Transcription elongation is not a monotonous process (Conaway *et al.*, 2000). The RNA polymerase can pause or be halted at certain DNA sequences or the progression of the enzyme can be hindered by nucleosomes. Factors interact with the enzyme during elongation to help its progression. These factors can be divided into different classes. Elongation factors increase nucleotide incorporation speed and

suppress polymerase II pauses. Cleavage factors increase the intrinsic RNA cleavage activity of polymerase II ternary complexes and allow the complex to back up, giving it another chance to go through the pause or arrest site. A last class of elongation factors is made of complexes that help the RNA polymerase to transcribe through nucleosomes by increasing their movement.

Transcription is thus a complex process involving a huge number of proteins that intervene during the transcription cycle. It should be stressed again that each of these steps, be it initiation, elongation or termination, can be regulated. Therefore, transcription abundance variations can reflect changes in the rates of any of the transcription steps.

1.3.6

Regulatory promoters

Transcription of the great majority of eukaryotic genes is controlled by sequence-specific DNA-binding proteins. Eukaryotic DNA is usually packed into chromatin, blocking the access of RNA polymerases to the initiation sites of gene transcription. Therefore, binding of transcriptional activators to specific elements in regulatory promoters and enhancers is an essential prerequisite for transcription. Most genes are expressed only in certain tissues or during short developmental stages. The activation of most promoters depends on the co-operation of several regulatory proteins. This obligatory formation of activator complexes guarantees high specificity of transcription, even in large genomes such as that of human beings. A region of a few hundred base pairs around the core promoter generally contains most of the activator binding sites and is termed the regulatory promoter. Enhancers, by contrast, are regulatory elements that are located further upstream or, in some cases, downstream of the gene. Enhancers may also be present in introns (Carey & Smale, 2000). Detailed information on the mapping of eukaryotic promoters is available from the Eukaryotic Promoter Database (EPD) at <http://www.epd.isb-sib.ch/> (Swiss Institute for Experimental Cancer Research, ISREC; Epalinges sur Lausanne, Switzerland) (Perier *et al.*, 1998, 1999 and 2000; Praz *et al.*, 2002).

Several proteins are required to allow RNA polymerase II to bind efficiently to promoters. These proteins may be classified as follows (see figure 1.7 and section on transcription machinery and core promoters). (i) Basic transcription factors which are needed at every promoter. (ii) Trans activators, proteins that bind to enhancers and upstream activator sequences. Some trans activators react to signal molecules and thereby activate or inhibit transcription in response to changes within the intracellular environment. DNA loop formation is supported by unspecific binding of non-histone proteins. These so called 'high mobility group' proteins play an essential role in chromatin remodelling and transcriptional activation. (iii) Co-activators such as transcription factor II D (TFIID), which do not bind DNA, but indirectly support the interaction of trans activators and the RNA polymerase II complex. TFIID, the best studied co-activator, is a huge complex of at least ten proteins (TATA box-binding protein, TBP, and TATA box-binding protein-associated factors). Some TATA box-binding protein-associated factors resem-

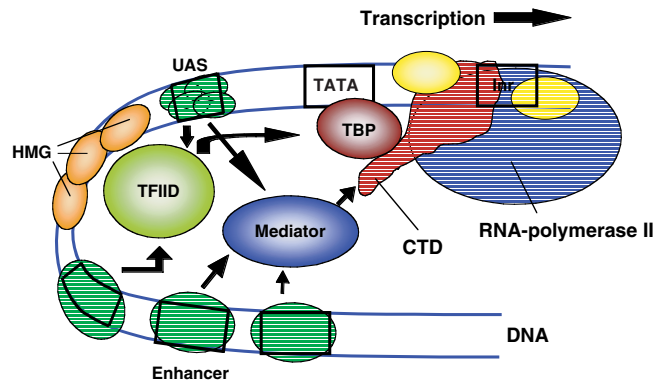


Figure 1.7 The regulatory promoter of eukaryotes. RNA polymerase II (translucent blue) and the basic transcription factors build a pre-initiation complex at the TATA box and the eukaryotic initiator sequence (Inr) of each promoter. Complex formation is supported by trans activators (green translucent) which have an effect on TFIID (green) and the mediator (blue). Unspecific bound high-mobility group

proteins (HMG) are essential for chromatin remodelling (orange). The carboxy terminal domain (CTD) of the RNA polymerase II is important for the interactions with protein complexes (arrows) such as, for example, the mediator. Histone acetylating proteins are not shown; upstream activator sequence is abbreviated as UAS and TBP is the abbreviation for TATA box-binding protein.

ble histones and are probably involved in displacement of nucleosomes from DNA (Carey & Smale, 2000).

1.3.7

Enhancers

Studies of viral enhancers have revealed some general properties of enhancer mechanisms. Different activators cooperate during promoter activation. Utilisation of cell-type specific regulators guarantees the general functionality of enhancers. Since enhancers are distant from the promoter, activation is independent of the exact orientation and positioning of the enhancer. Enhancers are composed of modules with multiple binding sites. The multimerisation of such modules augments the activation capacity of the enhancer.

The exact mechanisms of enhancer-dependent transcriptional activation is not known. One model of enhancer-dependent transcriptional activation was that enhancers transmit changes in the DNA structure to the core promoter. Another hypothesis postulated that enhancers act as binding sites for RNA polymerase II which then moves along the DNA towards the core promoter. Both models were disapproved by experiments that demonstrated sustained activation or transcription after protein bridging of the enhancer and the core promoter. The binding of activators to the enhancers probably lead to chromatin remodelling and loop formation. This enables the enhancer complex to interact with the core promoter. Stabilisation of the transcription complex may be the key feature of enhancer-

mediated transcriptional activation. The looping model is also supported from an energetic point of view. Michael R. Botchan and his colleagues have produced visual evidence of this model of enhancer action. They created an artificial DNA molecule with several promoter sites for the transcription factor SP1 and several enhancer sites, the latter being 800 base pairs from one end of the construct. The enhancers were bound by the enhancer-binding protein E2. The SP1 sites and the enhancers were separated by 1,860 base pairs. When a mixture of SP1 and E2 proteins was added to the DNA molecules, the electron microscope picture showed the DNA drawn into loops with 'tails' of approximately 300 and 800 base pairs. At the neck of each loop were two distinguishable accretions of material, representing SP1, and E2 molecules respectively. Artificial DNA molecules lacking the promoter or enhancer sites, or containing a mutant promoter or enhancer, failed to form loops when mixed with the two proteins.

Some enhancers are activated by cooperative binding of a combination of regulators. The impact of cooperative binding on gene expression patterns was studied in several experiments (for review see Carey, 1995; Ernst & Smale, 1995; Grosschedl *et al.*, 1994). It is not clear if enhancer-mediated activation produces a general increase in the transcription rate (rate model) or if the transcription is stabilised by the enhancer linked to a particular gene (binary model) (Bulger & Groudine, 1999; Fiering *et al.*, 2000; Molet *et al.*, 2001). Most experimental data are consistent with both models.

1.3.8

Locus control regions

A number of eukaryotic gene loci are regulated by locus control regions (LCRs). These elements are defined by their ability to confer high-level tissue-specific transcription to linked genes, as well as by their capacity to overcome position effects in transgenic assays. Locus control regions modulate transcription by a chromatin opening activity that acts through an extended DNA region (*e.g.*, over 100 kilobases; Bulger & Groudine, 1999; Engel & Tanimoto, 2000), distinguishing them functionally from classical enhancers. Locus control region-mediated changes in chromatin structure renders the DNA hypersensitive to nuclease activity. Each locus control region can include several hypersensitive sites, each representing a module with a distinct function. It is not known how locus control regions work. They seem to maintain accessibility of a region to transcription factors (Grosfeld, 1999). It has been suggested that locus control regions loop out and interact with regulatory promoters and enhancers. Locus control regions characteristically hyperacetylate and open chromatin over long distances. The acetylation of histones starts at the locus control region and moves along the chromatin (Elefant *et al.*, 2000). The expression level is proportional to the copy number of the locus control region. The duration of contact of the locus control region with the regulatory promoter or enhancer of the linked gene controls the period of transcription. Locus control regions function even in the presence of generally inactivating centromeric heterochromatin.

Insights into the mechanisms of locus control region-mediated transcription activation came from studies of the well characterised multigene clusters such as the human and mouse β -globin loci. Due to its compact organisation, the β -globin locus has become a model system for locus control regions. In transgenic mice experiments, high levels of expression were only observed when the constructs included the locus control region-containing region from the distal end of the globin locus (Bender *et al.*, 2000). The expression of β -globin is independent of the integration site. Although deletions of the human or mouse β -globin locus control regions significantly reduced transcription, this does not result in a closed chromatin structure of the endogenous β -globin locus. These results suggest a sequential activation model in which promoter remodelling and gene hyperacetylation lead to low-level basal transcription and precede locus control region-mediated high-level activated transcription (Schübeler *et al.*, 2001). In many systems, histone acetylation has been associated with chromatin opening and gene regulation. Local hyperacetylation is generally linked to transcriptional activation (Chen *et al.*, 1999). In some cases, histone-acetyl transferase (HAT) recruitment for a promoter has been shown to be the critical and rate limiting step for transcription and to occur before polymerase initiation (Agalioti *et al.*, 2000; Shang *et al.*, 2000). The locus control region may be necessary for the modification or the recruitment of new components for the remodelled and hyperacetylated promoter or for additional chromatin modifications.

1.3.9

Matrix attachment regions

Since locus control regions can activate gene expression over long distances, negative regulating elements must exist in order to repress locus control region function on adjacent loci. Two different classes of negative control elements have been described: Matrix attachment regions (MARs) and insulators. Matrix attachment regions are DNA fragments capable of associating with nuclear structures that remain associated even after stringent extraction and washing under high salt conditions or stripping with detergents. Chromosomes are thought to be incorporated into looped structures with the loops fastened to the intranuclear matrix (Gasser & Laemmli, 1987; Phi-Van & Strätling, 1990). It has been suggested that these loops may separate each chromosome into regulatory domains. Matrix attachment regions are AT-rich sequences often located in the vicinity of transcription units or enhancers. Several unique proteins have been found in association with matrix attachment regions (Hart & Laemmli, 1998).

1.3.10

Insulators

Most insulators or boundary elements have been found in the vicinity of junctions between decondensed and condensed chromatin, which presumably correspond to active and inactive loci, respectively. The insulators appear to provide a boundary

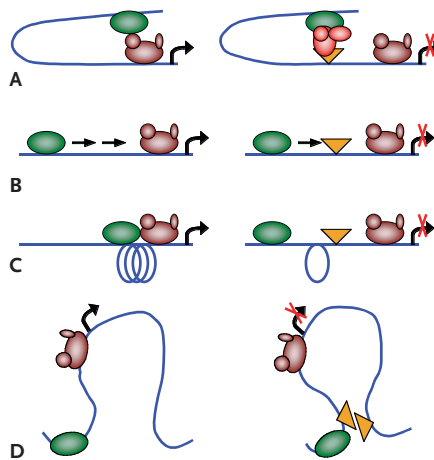


Figure 1.8 Model of enhancer blockade by insulators. **A)** The enhancer (green) interacts by looping out to interact with the promoter complex (brown). The insulator (red) traps the enhancer *via* mimicry of the promoter. **B)** The insulator acts as a molecular barrier blocking the propagation of the enhancer signal to the promoter (black arrows). **C)** The insulator blocks the looping necessary for the interaction of the enhancer with the promoter. **D)** The interaction of two insulator complexes results in DNA looping. Activation of the promoter is possible only by an enhancer located within the same loop.

between accessible and inaccessible chromatin. For repression, the insulator must be located between the enhancer and the gene. Several models of insulator function have been described (figure 1.8). One model is based on the idea that the insulators trap enhancer activity and eventually disturb the loop forming proteins (Bell *et al.*, 2000). A second model proposes the insulator as a molecular barrier for a propagating enhancer signal. Alternatively, an insulator may block factors needed for the loop formation of chromatin between enhancer and promoter. It is still unclear, if the interaction of two insulators could induce a chromatin loop, which would hold some genes in an inactive core compartment.

The highly conserved ubiquitous CCCTC-binding factor (CTCF) activates gene expression. CTCF binds at an insulator and blocks enhancers. CTCF is the first factor that has been isolated as being essential for gene separation in vertebrates (Filippova *et al.*, 1996). All insulators discovered to date in vertebrates function only when bound by CTCF. CTCF has eleven zinc finger domains. The domains for DNA binding, recruitment of co-repressors and repression of transcription have been identified (Lutz *et al.*, 2000). CTCF binds to different DNA elements – a common consensus sequence has not been identified. Locus control regions, enhancer induced acetylation and opening of chromatin proceeds until an insulator has been reached. Insulator-bound CTCF recruits histone deacetylases, their activity counteracts the acetylation and stops extension of the enhancer signal.

1.3.11

RIDGES – Regions of increased gene expression

Most chromosomes contain regions of increased gene expression or RIDGES, in which transcription per unit length is seven times that of the genomic average and 20 to 200 times that of weakly expressed regions (Caron *et al.*, 2001). On chromosome 6, for example, the regions of increased gene expression correspond to the major histocompatibility locus (MHC). Regions of increased gene expression are

not randomly distributed within the genome, but most likely represent a higher order structure. Many of the regions of increased gene expression have a high gene density. For 50 to 60 percent of the regions of increased gene expression, a correlation between gene expression and density of the genes is found. The remaining 40 to 50 percent are not gene dense. These regions of increased gene expression preferentially map to telomeres, which is remarkable in view of the phenomenon of telomeric silencing that has been observed in yeast (Velculescu *et al.*, 1997).

1.3.12

Enhanceosomes

The term enhanceosome has been assigned to functionally related groups of DNA-binding proteins that act in concert to enhance transcriptional activation upon the presence of the transcription machinery. Enhanceosomes therefore represent another potential mechanism for fine tuning of transcriptional activation. Experimental evidence for the enhanceosome theory has first been provided by Thanos & Maniatis (1995). Subsequent studies suggested that the proteins involved in enhanceosome formation presumably form a stereospecific assembly of surface structures enabling a more efficient binding of the transcription apparatus (Kim & Maniatis, 1997). One of the best characterised mechanisms of enhanceosome action is the activation of the interferon- β gene in response to virus infection (Yie *et al.*, 1999). The first step in interferon- β enhanceosome assembly is recruitment and cooperative binding of activating factors (nuclear factor κ B, NF- κ B; interferon regulatory factors, IRFs; activating transcription factor, ATF2, homodimers and ATF2/c-Jun heterodimers) orchestrated by the architectural protein HMG1Y (high mobility group protein isoforms I and Y). Subsequent changes in the three-dimensional structure form an 'enhanceosome pocket' that facilitates recruitment and binding of the basal transcription machinery. This in turn leads to an increased rate of pre-initiation complex formation. The DNA-binding affinity of the factors is also regulated by the acetylation of histones, nucleosomes and the proteins of the enhanceosome complex, and by the cAMP response element-binding protein (CREB)-binding protein (CREBBP). The current view is that the enhanceosome complex is also involved in regulating chromatin-modifying factors and that specific gene activation is dependent on the presence of all enhanceosome components (Merika & Thanos, 2001).

1.3.13

Chromatin

The eukaryotic genome is organised into chromosomes consisting of nucleosomes, which are independent functional domains. Nucleosomes are the basic units of chromosome architecture. The nucleosomal structure is regulated by the temporal and spatial status of the cell, by the ATP-dependent chromatin-remodelling SWI/SNF-like protein complex and by histone acetyltransferases (for review, see Fry &

Peterson, 2001; Kingston *et al.*, 1996). Recent experimental data suggest an important role in tumourigenesis for chromatin-remodelling proteins BRM (*brahma*) and BRG1 (*brahma*-related gene 1), which are two ATPase subunits of the SWI/SNF-like complex. In addition to different chromatin and nucleosomal structures, the DNA molecule itself may have different conformations. A dramatic structural transition is made when the right-handed B-DNA double helix is converted to a left-handed Z-DNA double helix (Wang *et al.*, 1979). The energetically unfavourable Z-DNA conformation is stabilised by negative supercoiling, which occurs in DNA wrapped around nucleosomes. Negative supercoiling may also occur during transcription just behind the moving RNA polymerase (Liu & Wang, 1987). Z-DNA structures have also been detected in the promoter region of transcribed mammalian genes, suggesting that DNA conformation is an important feature of transcriptional regulation (Herbert & Rich, 1999).

1.3.14

Silencer elements

Since each cell expresses only a small fraction of the genes it contains, most genes must be kept silent. Some genes must also be rapidly switched on and off. Transcriptional repression in eukaryotes is achieved by silencers. Two types of silencers have been described, classical silencer elements and negative regulatory elements, NREs (for review, see Ogbourne & Antalis, 1998). Classical silencer elements are DNA motifs which interfere with transcription factor assembly of the general machinery (active repression mechanism), whereas negative regulatory elements physically inhibit the interaction of transcription factors with their specific DNA-binding sites, or interfere with other factors regulating transcriptional events (passive repression mechanism). Silencers have also been described as orientation and promoter-dependent and have been found in introns, in the 3'-untranslated regions and even in exons. Although the research on silencers and the identification of new silencer elements have been rapidly increasing over the past ten years, knowledge of the silencing mechanisms is still limited.

1.3.15

Transcription factors, repressors and co-repressors

Transcription factors, repressors and co-repressors are proteins that bind to promoter, enhancer or silencer regions in highly specific fashion in order to up or down-regulate gene expression. These elements are predominantly located in the cell nucleus, exist in a great number and variety (more than 2,000 are known today) and are essential factors of transcriptional regulation (for recent review, see Brivanlou & Darnell, 2002). *In vitro*, RNA polymerase II alone fails to find the correct transcriptional start site and to synthesise efficient amounts of RNA unless nuclear extracts are also present. The proteins involved in transcriptional regulation can roughly be classified into general, constitutively active nuclear factors and regulatory transcription factors. The general factors are involved in each transcriptional process, the

regulatory factors are responsible for example for the timing of expression of particular genes needed for embryonic development. Since transcriptional regulation not only reflects the basic processes of a cell (*e.g.*, metabolism, osmoregulation, cell division), but also tissue specific tasks and challenges (*e.g.*, production of clotting factors, immunoglobulins, repair of skin injuries), transcription factors are in continuous communication with the cell environment *via* cell surface sensors and receptors and the intracellular signalling system.

One of the most critical general transcription factors is TFIID. The association of TFIID with the promoter is one of the initial steps for the assembly of the RNA polymerase II holoenzyme. TFIID contains the TATA box-binding protein. The interaction of the TATA box-binding protein with the DNA leads to bending of the DNA double helix and to exposure of the minor groove followed by a cascade of interactions with TATA box-binding protein-associated factors and various regulatory and stabilising proteins. The assembly of the initiation complex requires at least 30 different transcription factors, which have been alphabetically designated as TFIIA, TFIIB, *etc.* The transcription factors themselves are activated and deactivated by phosphorylation. Another important general transcription factor is SP1, which binds to the major groove of GC boxes in the promoter region *via* a zinc finger structure. GC boxes are usually found in the promoters of housekeeping genes which frequently lack a TATA box. SP1 promotes the formation of the initiation complex by interaction with TATA box-binding protein-associated factor proteins and assists TFIID in finding the correct transcriptional start site. The CCAAT box transcription factor (CTF) is another example for a general transcription factor, which participates in the assembly of the initiation complex. A tissue-specific variant of the CCAAT box transcription factor is C/EBP (CCAAT/enhancer-binding protein), which is predominantly found in liver cells.

RNA polymerase I, responsible for the synthesis of rRNAs, and RNA polymerase III, responsible for the transcription of tRNAs, 5S rRNAs and a few other stable RNAs, are regulated by a distinct set of *general* transcription factors including upstream binding factor (UBF), TAF1B (SL1), TFIIB, TFIIC, TATA box-binding protein and a number of TATA box-binding protein-associated factors and other factors, which promote the formation of the polymerase-specific initiation complexes (Weinzierl *et al.*, 1993).

The regulatory (specific or conditional) transcription factors can be divided into developmental (embryonic development) and signal-dependent factors. Cellular signals may be classified into receptor-ligand (cell surface), internal and steroid receptor signals (Brivanlou & Darnell, 2002). Steroid and thyroid hormones are transcriptionally active molecules that circulate in the body and gain access to regulatory DNA elements not by signal transduction, but also by uptake into the cell and diffusion to the nucleus. The most predominant signal-dependent transcription factors are involved in cell proliferation and reside in the nucleus. Well-established resident nuclear factors are the oncogenes *Fos* and *Jun*, which form heterodimers with the transcription factor AP1. The genes coding for these factors are characterised by additional promoter elements that respond to growth factors and the downstream signalling pathways. The activation and synthesis of *Fos* seems also

to be determined by an auto-regulation process. Because serum is particularly rich in growth factors, a number of transactivating functions are initiated *in vitro* upon incubation of a cell culture with serum-containing media. These events include transactivation of serum-responsive elements (SREs) contained in certain gene promoters by serum-responsive factors (SRFs) and other cooperating factors such as the oncogene ELK1 and the p62 subunit of TFIIH. Another important transcription factor involved in cell proliferation is E2F which interacts with the product of the retinoblastoma gene in controlling the cell cycle. Another class of receptor-mediated, signal-dependent transcription factors has been designated as latent cytoplasmic factors that include the REL/NF- κ B family and the proteins of the *Sma* (a protein that is suggested to be involved in the transforming growth factor β [TGF- β], signalling pathway), *SMAD* (*Mad* [mothers against decapentaplegic]-related protein), and *STAT* (signal transducer and activator of transcription) pathways, responding to stress, immunological challenges or growth factors.

Since most of the genes of eukaryotic cells are silent, mechanisms of negative transcriptional regulation are also required. The cell must be able to rapidly switch on and off the transcription of genes. This is achieved by activation and deactivation of transcription factors *via* phosphorylation and dephosphorylation, by the interaction of repressors with silencer motifs of the DNA and inhibitors. Similar to the activating transcription factors, there are global (*e.g.*, B-TFIID transcription factor-associated factor, BTAF1) and specific (*e.g.*, neuron restrictive silencer factor, NRSF, also called repressor element-1, RE1, silencing transcription factor, REST) repressors and co-repressors which interact with silencer elements. Well established examples of inhibitory elements are I κ B proteins that block NF κ B, the MAX/MAX (oncogene MYC-associated factor X) dimer that blocks the MYC/MAX heterodimer, the cAMP response element modulator (CREM), which is an alternative splicing product of the cAMP response element (CRE) gene lacking the transactivating domain, or the CCAAT displacement protein (CDP).

In summary, the number of newly identified classes and families of transcription factors and other regulatory elements has been rapidly increasing over the past ten years. The Human Genome Project has contributed a great number of new targets and potential functions. A complete list and review of all factors would greatly exceed the limits of this book chapter. The functional interrelationships between the different classes and families of transcription factors is far more complicated than was originally thought (Lemon & Tijan, 1987). Today, we understand the communication between single cells or tissues and the whole organism on the one hand side and the cross talk between the cell surface and its protein machinery on the other side as complex interactions on several levels resembling a neuronal network (figure 1.9).

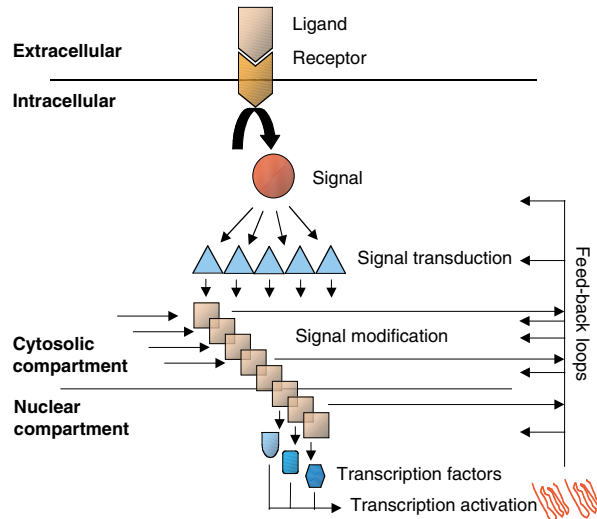


Figure 1.9 Schematic model of signalling pathways between extracellular signal receptors and intracellular signal response. The organisation of signal modification on several levels of signal transduction resembles data management in neuronal networks.

1.3.16

Epigenetics

Recent collected data from cloning experiments in mammalian cells have shown that the gene expression profile of a nucleus is not determined solely by its exposure to a particular set of transcription factors. In other words, heritable changes in gene function must exist that cannot be explained by changes in the DNA sequence (Russo *et al.*, 1996). This field on non-sequence-related heritable genetic regulation has been termed 'epigenetics'. An important mechanism of epigenetic regulation is methylation of GC-rich DNA regions. GC-rich regions are primarily located upstream of genes (called CpG islands). Methylation of the CpG islands upstream of a gene silences that gene, a mechanism which is essential for embryonic development, for so-called imprinting and for chromosome X inactivation. In most cases, the methylation pattern of a cell is transmitted to the daughter cells during cell division. The mechanisms governing DNA methylation patterns remain to be elucidated (for recent review, see Bird, 2002). Another example of epigenetic regulation is the protein complexes of the polycomb/trithorax group (Pc-G/trx-G) that have been identified in *Drosophila*. Pc-G/trx-G is a multiprotein complex that interacts with specific regions of the DNA leading to an arrested state of gene expression in early embryogenesis that is stable throughout development. The current understanding is that both CpG methylation and the Pc-G/trx-G complex serve as an epigenetic memory of transcriptional activity during development. The mechanisms by which these patterns are generated are not yet understood.

1.3.17

Summary and conclusions

Up to recently, there was an understandable emphasis in the biological scientific community on determining the sequence of the genetic code. Now, in the post-genomic era, the focus has shifted to understanding how the blueprint contained in the genes is read, first to generate the organism during embryogenesis, and then to control cell fate and function in the finished product. This chapter has attempted to provide a short review of the processes involved in regulating gene transcription. We have seen that control of transcription is the main, but, as we will see in the next sections, by no means the only way in which gene expression is regulated. We will also see that the population of protein molecules in a cell generally, but not always reflects the population of mRNA molecules. For this reason, a comprehensive understanding of cellular function requires knowledge not only of the transcriptome, but also of the proteome and of how proteome and transcriptome interact. We have learned that the expression of a particular gene is regulated not only by the interaction of its direct upstream promoter with a dizzying array of transcription factors and repressors, but also by a complex interaction of often distant regulatory regions including enhancers, locus control regions, matrix attachment regions, insulators and regions of increased gene expressions. The packing of the DNA in chromatin, interaction with a range of proteins including histones and enhanceosome complexes, the conformation of the DNA and methylation of DNA add a further layer of complexity to regulating when, where and with what intensity a particular gene is expressed. The next sections of this chapter will further extend the complex interaction to the post-transcriptional and post-translational level.

In the past, many breakthroughs in biology have been achieved by studying simple model systems. The classic studies of the *Lac* operon by François Jacob, Jacques Monod and their co-workers are a good example of this (Jacob & Monod, 1960 and 1961). Now, however, the paradigm may be changing. It is clear that answering at least some of the questions in the new biology will require a willingness to analyse systems of great complexity and to look for patterns rather than single data points. Increasingly, the whole is being seen to be much more than the sum of the parts.

1.4**Post-transcriptional regulation**

1.4.1

Introduction

At first glance, it is surprising that the expression of genes is modulated not only by controlling the rate of transcription initiation, although it could be argued that this is the most economical mechanism for a cell. Thus, little energy would be expended if the gene is not transcribed into RNA when the gene product is not

needed. However, although transcription initiation is a principal mode of regulating many, if not most, genes, regulation at several other points in the expression pathway may also occur. Transcription elongation, mRNA stability, pre-mRNA-splicing, mRNA transport, polyadenylation, and translation are such alternative modes of regulating gene expression. It should be remembered that the determination of an expression profile only provides information on the steady-state mRNA abundance of a particular mRNA and provides no clues on how the expression of that gene is modulated (Carey & Smale, 2000).

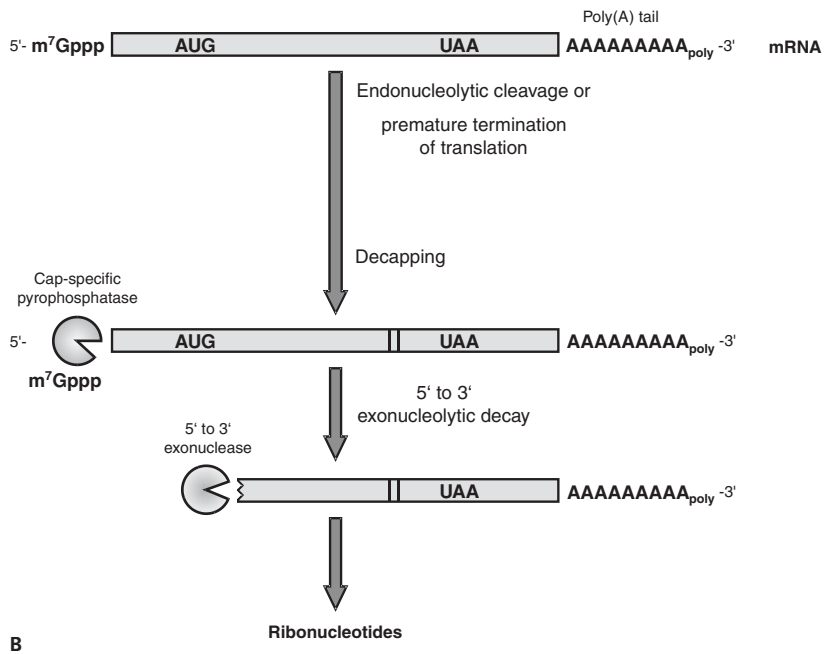
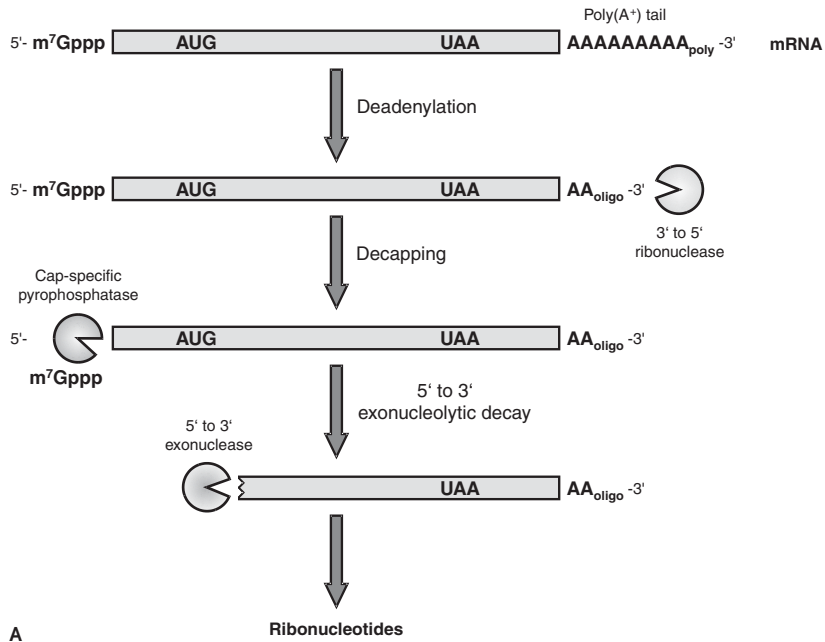
1.4.2

Regulation of RNA stability and degradation

It is obvious that the steady-state abundance of every mRNA is dependent on its rate of synthesis and decay (Caponigro & Parker, 1996; Jacobson & Peltz, 1996; Ross, 1995; Sachs, 1993). The stabilities of mRNAs in mammalian cells vary widely, with approximate half-lives ranging from a few minutes (*e.g.*, *c-fos* mRNA) to several hours (*e.g.*, β -actin mRNA) (Ross, 1995). Moreover, the mRNA stability depends not only on the chemical composition of the RNA molecule, but also on the activity of several different mRNA degradation pathways.

Details of the mRNA degradation pathways within eukaryotic cells are lacking. It is also not known how many different pathways for controlled mRNA degradation exist. Recent studies performed primarily in *Saccharomyces cerevisiae* have identified the so-called deadenylation-dependent mRNA degradation pathway (figure 1.10.A) (Caponigro & Parker, 1996; Jacobson & Peltz, 1996), one of the two degradation pathways described so far. This pathway seems to be conserved in mammalian cells and may be responsible for the passive and regulated degradation of many mRNAs. The deadenylation-independent pathway (figure 1.10.B) is a second mechanism of mRNA degradation in mammalian cells, which may be quite similar to the deadenylation-dependent pathway following the event that initiates degradation (Carey & Smale, 2000).

Determinants of intrinsic mRNA stability include primary mRNA sequences, mRNA secondary structures, and proteins binding to the mRNA that alter the rate of poly(A) shortening, cap removal, or the 5' to 3' exonuclease activity (Caponigro & Parker, 1996; Rajagopalan & Malter, 1997; Ross, 1995). A well-studied common determinant is the AU-rich response element (ARE), that contributes to the short half-lives of a number of early response mRNAs (Chen & Shyu, 1995; Jacobson & Peltz, 1996; Ross, 1995; Wilson & Brewer, 1999). These elements often contain the sequence AUUUA and are typically found in multiple copies in the 3'-untranslated regions of relevant mRNAs. Although it has been shown that these elements contribute to mRNA stability *in vitro*, the precise nature of the AU-rich response elements and the proteins binding to these sequences are not known. One well-studied mammalian gene containing such AU-rich response elements is *c-Fos*, degradation of which is consistently preceded by the rate of poly(A) tail shortening directly influencing the deadenylation-dependent pathway (Carey & Smale, 2000).



◀ **Figure 1.10 Messenger RNA degradation pathways.**

A) The deadenylation-dependent pathway, and **B)** the deadenylation-independent pathway. The figure is adapted from Carey & Smale (2000). An early step in the deadenylation-dependent pathway is the shortening of the poly(A) tail by 3' to 5' ribonucleases (deadenylation) which appears to take place in the cytoplasm. The poly(A) nuclease (PAN) in yeast and the deadenylating nuclease (DAN) in mammalian cells may be involved in this process (Korner & Wahle, 1997). After the poly(A) tail is shortened, the 7-methyl-guanosine cap structure at the

5'-end of the mRNA is removed by a cap-specific pyrophosphatase (decapping). This removal makes the mRNA susceptible to cytoplasmic 5' to 3' exonucleases completing the degradation process (5' to 3' exonucleolytic decay) (Carey & Smale, 2000). Compared to the deadenylation-dependent pathway, the deadenylation-independent pathway initiates with an endonucleolytic cleavage of mRNA or a premature termination of transcription which afterwards promotes degradation similar to that of the deadenylation-dependent degradation pathway (Carey & Smale, 2000).

Two well-characterised regulatory mechanisms based on mRNA stabilities concern regulation of levels of mRNA for α -tubulin and β -tubulin (Carey & Smale, 2000; Ross, 1995; Theodorakis & Cleveland, 1993) and the regulation of ferritin expression (Carey & Smale, 2000; Hentze & Kühn, 1996). Details of both systems are described in detail elsewhere (Carey & Smale, 2000).

From an experimental point of view, there is no single experiment that definitely determines if a gene is regulated at the level of transcription initiation or mRNA stability. Typically, nuclear run-on experiments are combined with standard transfection procedures to distinguish between the two modes of regulations and to determine their relative importance (Carey & Smale, 2000).

The most important unanswered question concerning mRNA degradation is whether most or all events in a cell that promote mRNA degradation feed into a single pathway, or whether several parallel degradation pathways exist (Wilson & Brewer, 1999). Clearly, there are several mechanisms for initiating degradation as mentioned above. Further studies are necessary to determine if the diversity of initiating events is due to the existence of several mRNA degradation pathways, or whether the various initiating events converge on one or a few general pathways.

1.4.3

Regulation of transcription elongation

1.4.3.1 Introduction

Usually, transcription is regulated by the initiation event. Nevertheless, a large number of genes are known to be regulated during elongation of transcription, a process that is catalysed by RNA polymerases in a highly processive manner following initiation and promoter escape (Bentley, 1995; Greenblatt *et al.*, 1993; Reines *et al.*, 1996; Shilatifard *et al.*, 1997; Uptain *et al.*, 1997). Although promoter escape is an equally attractive target for regulation, in eukaryotes, this event cannot be measured unambiguously *in vivo* or separated from initiation and elongation (Carey & Smale, 2000).

The process of transcription elongation requires a stable ternary complex consisting of RNA polymerase, DNA template, and nascent transcript (Nudler, 1999; Up-

tain, 1997) and takes place within a DNA ‘bubble’ that is formed by unwinding and reannealing of the double-stranded template DNA as the RNA polymerase moves along the gene (see sections above on basics of transcription and translation in the cell). Based on studies of the RNA polymerase of *Escherichia coli*, the movement of RNA polymerase along the DNA is thought to depend on the contact of the enzyme with the DNA at three sites. (i) A contact with the double-stranded DNA that precedes the transcription bubble based on strong, non-ionic interactions. (ii) A relatively weak contact with the RNA/DNA hybrid immediately adjacent to the site of ribonucleotide addition. (iii) A contact with the nascent transcript upstream of the RNA/DNA hybrid binding site. Each binding site contributes to the stability of the complex and restricts movement in the vicinity of the other binding sites. On the other hand, the presence of three binding sites allows each to be sufficiently flexible to allow movement following ribonucleotide addition. Thus, this structure provides the stability and flexibility that are essential for an efficient elongation process (Carey & Smale, 2000).

Elongation of transcription is a complicated and dynamic process that can be influenced by several factors and determinants. It is therefore not surprising that the precise nature of the regulation mechanisms and its influencing determinants has not been established. The following sections present some examples of mechanisms regulating the process of transcription elongation.

1.4.3.2 Regulation of transcription elongation in prokaryotes

In prokaryotes, several different mechanisms and numerous determinants of elongation rate have been identified (Carey & Smale, 2000). Two classic examples are well-characterised in the model organism *Escherichia coli*.

The first example is the *Escherichia coli* tryptophan (*trp*) operon, which is expressed at high levels in the absence of tryptophan and transcribed less frequently as the concentration of tryptophan increases. Transcription of the tryptophan operon is regulated in part at the level of transcription initiation as well as at the level of elongation. The latter process mediating this regulation is called attenuation (Yanofsky, 1998).

For the understanding of the following paragraphs, it is important to define the terms RNA polymerase pausing, arrest and termination. According to Uptain *et al.* (1997), pausing of the RNA polymerase is defined as “a reversible cessation of elongation for a defined period of time”, whereas arrest is “similar to pausing, but arrested RNA polymerase molecules cannot continue unless supplementary factors are added”. Termination means the end of the transcription elongation mediated by dissociation of the transcription complex. It is noteworthy that it is often difficult to distinguish between pausing, arrest, and termination of elongation. This has led to the use of a more generic term, an elongation block (Carey & Smale, 2000).

During transcription of the tryptophan operon, the presence of tryptophan causes elongating RNA polymerase molecules to pause frequently after synthesising a 130-nucleotide leader transcript. If tryptophan is absent, then synthesis of the

entire 7000-nucleotide operon is more likely. The leader RNA encodes a short tryptophan-rich peptide and consists of four intra-complementary regions of which two are able to form a hairpin between each other. A hairpin between region 3 and 4 is formed in the presence of tryptophan and signals transcription termination. The hairpin formation between region 3 and 4 is promoted by the ribosome that efficiently synthesises the leader peptide following closely behind the RNA polymerase and therefore sterically preventing region 3 from instead forming a hairpin with region 2. In the absence of tryptophan, or in the presence of low concentrations, translation is less efficient due to the low tryptophan concentrations. The reduced rate of translation, associated with a low rate of bound ribosome and a therefore reduced sterically prevention of 3:4-hairpin formation, allows region 3 to pair with region 2, rather than region 4. This prevents the transcriptional arrest and thereby promotes transcription of the entire operon (Carey & Smale, 2000).

The second example involves the bacteriophage λ Q protein, which is required for efficient elongation through the λ late genes (Greenblatt *et al.*, 1993; Uptain *et al.*, 1997). *Escherichia coli* RNA polymerase pauses after transcribing 16 or 17 nucleotides of the late gene if the Q protein is absent. Even if the RNA polymerase is released from this pause site, it is susceptible to pausing at several other locations within the late gene. A sequence-specific interaction between the polymerase-associated sigma factor and the non-template strand of DNA in the transcription bubble seem to be responsible for the pause in transcription elongation. The Q protein is released from the pause site and facilitates elongation by binding to the RNA polymerase and presumably altering the conformation of this enzyme. The resulting complex of Q protein and RNA polymerase does not bind to pause sites within the late gene. Moreover, the Q protein binds exclusively to the RNA polymerase after the sigma factor promotes pausing at the proximal promoter site. Thus, the activity of the initiation factor is directly linked to the elongation competence of the RNA polymerase (Carey & Smale, 2000).

1.4.3.3 Regulation of transcription elongation in eukaryotes

The mechanism by which elongation of transcription is regulated in eukaryotes is not well understood. Current data suggest a process similar to that of the Q protein in *Escherichia coli*. Several studies have revealed a link between the relief of pausing and control elements within the promoter of eukaryotic genes but no homologue of the prokaryotic Q protein has yet been identified.

Much of our knowledge in this field has emerged from studies of the heat-shock, *c-myc*, and human immunodeficiency virus (HIV) genes (Bentley, 1995; Carey & Smale, 2000; Greenblatt *et al.*, 1993; Lis & Wu, 1993). These genes have two characteristics in common that may be general features of many eukaryotic genes that are regulated in part at the level of elongation. First, they all are subject to rapid induction of expression. Second, they all appear to promote pausing of RNA polymerase II within 20 to 60 base pairs of the transcription initiation site, similar to the location of the pause site controlled by the Q protein in *Escherichia coli* (Carey & Smale, 2000).

The *Drosophila hsp70* gene provides a prime example of the regulation of transcription elongation in eukaryotes (Lis & Wu, 1993; Uptain *et al.*, 1997). Figure 1.11 illustrates the *in vivo* regulation mechanism that is suggested for the *hsp70* promoter. Details of the mechanism are described in the legend to figure 1.11.

The human immunodeficiency virus-1 and *c-myc* genes appear to be regulated in a similar fashion to the *Drosophila hsp70* gene (Bentley, 1995; Carey & Smale, 2000; Greenblatt *et al.*, 1993; Jones, 1995; Uptain *et al.*, 1997). The precise regulation mechanisms of elongation of these genes are not precisely understood. The interested reader is referred to the model described in Carey & Smale (2000).

According to Carey & Smale (2000) several general elongation stimulatory factors may also contribute to the regulation of transcription elongation such as SII-type elongation factor (also termed transcription elongation factor A1, TCEA1, or TFIIS), TFIIF, elongin (a transcriptionally active component and regulatory subunits of the SIII complex consisting of elongin A, B, and C; also termed transcription elongation factor B1, TCEB1), eleven nineteen lysine-rich leukemia (ELL) gene, and TFIIF (Park *et al.*, 1994; Reines *et al.*, 1996; Shilatifard *et al.*, 1997; Uptain *et al.*, 1997). Although the precise mechanisms of regulating transcription elongation mediated by these proteins remain to be elucidated, it is known that they can (i) help an arrested RNA polymerase resume elongation, (ii) suppress RNA polymerase pausing, or (iii) modify the RNA polymerase by phosphorylation (Carey & Smale, 2000).

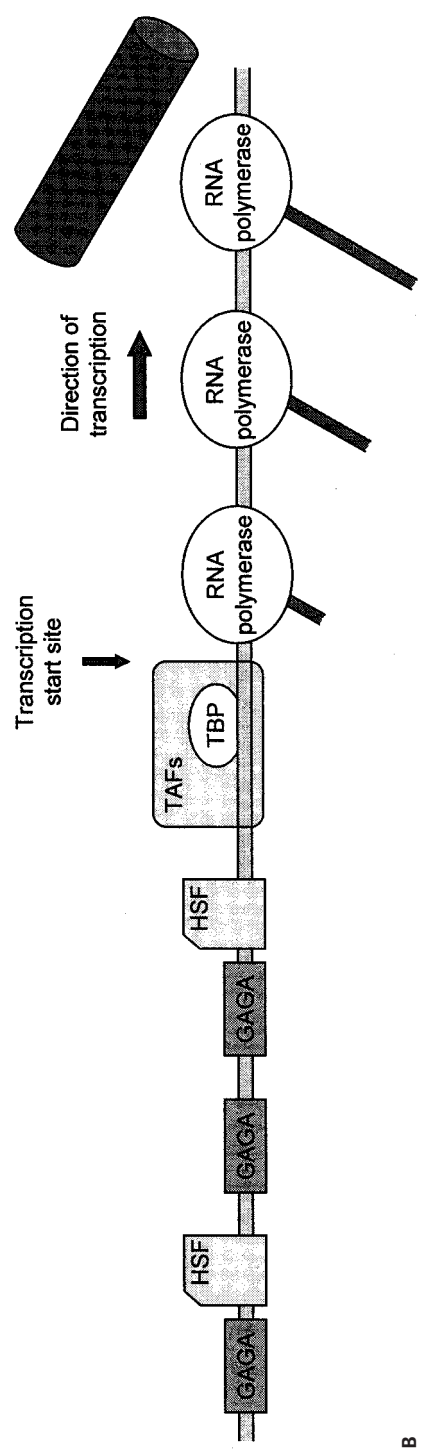
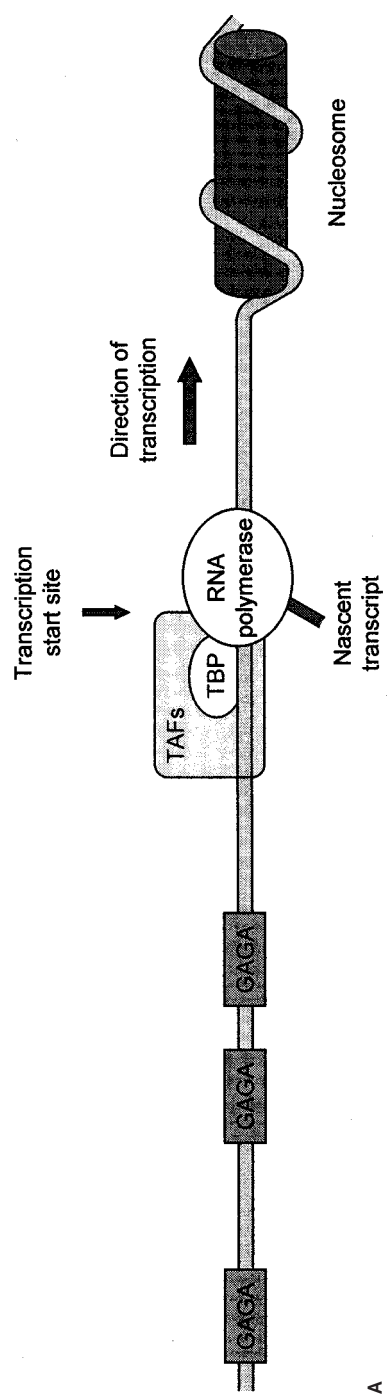
1.4.3.4 Conclusions

Further studies will be needed to determine the factors modulating the transcription elongation in prokaryotes and, in particular, in eukaryotes. Moreover, much more work is required to characterise the mechanisms by which such determinants function within cells.

In some ways, it is more difficult to distinguish between regulation of initiation and elongation than between initiation and RNA stability. In addition, not only would it require technically demanding experiments to distinguish between

Figure 1.11 The schematic architecture of the *Drosophila hsp70* promoter responsible for mediating the regulation of transcription elongation. A) Architecture of the promoter *in vivo* before heat shock, and B) after heat shock. Prior to heat shock and induction of *hsp70* expression the so-called GAGA transcription factor binds to multiple binding sites (GAGA) located in the core promoter of the *hsp70* gene to facilitate the initiation of transcription by RNA polymerase II. Under these circumstances, the RNA polymerase pauses after synthesising a transcript of only 20 to 40 nucleotides. Upon heat shock, a heat shock factor (HSF) is induced to bind the

promoter of the *hsp70* gene. This releases the paused RNA polymerase and allows the enzyme to synthesise the full-length transcript (Lis & Wu, 1993). In addition, further rounds of initiation are partly stimulated by the presence of heat shock factor. This is mediated by stimulating the assembly of a functional pre-initiation complex. Both processes result in a rapid induction of *hsp70* expression (Carey & Smale, 2000). The following abbreviations are used in the scheme: TBP, TATA box-binding protein; TAFs, TBP-associated factors. The figure is adapted from Carey & Smale (2000). ►



these two regulation mechanisms, this knowledge is also of only limited use as the precise mechanism of initiation is still unclear and the regulatory elements and interacting proteins remained to be identified.

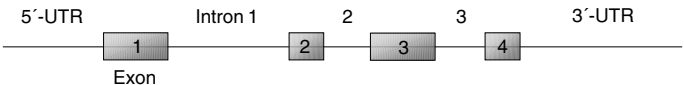
The technique most frequently used for the purpose of distinguishing between regulation of initiation and elongation is a modified nuclear run-on assay (see, *e.g.*, Blau *et al.*, 1996, and chapter 3 of this book). For more detailed information and strategies to distinguish between regulation of elongation and regulation of initiation, the reader is referred to Carey & Smale (2000).

1.4.4

Differential/alternative pre-mRNA splicing

Alternative splicing (see figure 1.12) is a common mechanism that enables the cell to generate different proteins from a single gene (for reviews, see Misteli, 2000; Modrek & Lee, 2002 and Smith & Valcarel, 2000). The primary transcript, which contains a copy of the entire genomic DNA of a gene, is tailored by the splicing apparatus into different mRNA species by skipping or omitting one or more exons and by using alternative splice sites. The data from the Human Genome Project suggest a total number of approximately 30,000 genes, whereas the number of protein species has been estimated to exceed 100,000. Recent data indicate that 40 to 60 percent of human genes have alternative splice forms. Alternative splicing is therefore thought to be the major reason for the numerical difference between the total numbers of genes and proteins (for more information, see, for example, the

Primary RNA



Alternative splicing



Splice variants of mature mRNA

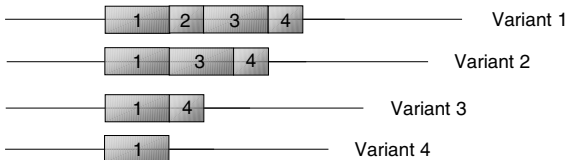


Figure 1.12 Schematic model of alternative splicing. Different splicing reveals in a great variety of mRNAs and protein subspecies encoded by a single gene.

Alternative Splicing Database at <http://cbcg.nersc.gov/asdb>; Gelfand *et al.*, 1999, and Dralyuk *et al.*, 2000). Most of the splice variants have recently been identified by alignment of expressed sequence tags to genomic data or mRNA. Although the most recent data on splice variants are based on information from the Human Genome Project, the data are largely incomplete due to fragmentation of the DNA data and false-negatives within the bioinformatics software. Cataloguing for alternative splice forms will be an important project for the future. Important information may also come from studying human diseases. Up to 15 percent of monogenic disorders may be caused by aberrant splicing.

1.4.5

Trans-RNA splicing

Maturation of mRNA in eukaryotes typically involves the removal of introns from long precursor molecules by the process of RNA splicing. An unusual form of RNA splicing in which separate precursor transcripts contribute sequences to the mature mRNA through intermolecular reactions has also been documented in several organisms. This phenomenon of pre-mRNA *trans*-splicing (Agabian, 1990), also called *trans*-RNA splicing, has been divided into two categories. The 'spliced leader' type, found both in protozoans such as trypanosomes and in lower invertebrates such as nematodes, results in the addition of a short, capped 5'-noncoding sequence to the mRNA (Blaxter & Liu, 1996; Lee *et al.*, 1997; Nilsen, 1992 and 1993). The 'discontinuous group II intron' form³⁶

trans-RNA splicing|discontinuous group II intron form^{4>} of *trans*-splicing, found in chloroplasts of plants and algae, and in plant mitochondria, involves the joining of independently transcribed coding sequences, presumably through interactions between 'intronic' RNA pieces (Bonen, 1993; Kohchi, 1988). Both categories of *trans*-splicing are mechanistically similar to conventional nuclear pre-mRNA *cis*-splicing.

Trans-RNA splicing is an essential step in pre-mRNA maturation in a variety of lower eukaryotic organisms. However, this processing pathway is absent in mammals, insects, yeast and plants, with the exception of some chloroplasts and mitochondria (Nilsen, 2001). The process of *trans*-splicing is thought to be a process for accelerating the evolution of new proteins. Only a few cases have been reported in which exons have been re-assembled using *trans*-splicing (Alberts *et al.*, 1998). It is therefore possible that *trans*-splicing is a remnant of a more complex mechanism that was in common use in primordial cells. The patchy phylogenetic distribution of *trans*-splicing is consistent with either 'multiple gain' or 'multiple loss' evolutionary scenarios. Although it is still unclear if this unusual type of splicing is an ancestral or an acquired trait, *trans*-splicing seems not to be involved in the transcriptional or post-transcriptional regulation of gene expression.

1.4.6

Regulation of mRNA transport

In addition to the other regulatory mechanisms described here, nucleocytoplasmic transport of mRNAs contributes to the mRNA abundance of many genes (Alberts *et al.*, 1998; Carey & Smale, 2000; Gorlich & Mattaj, 1996; Mattaj & Englmeier, 1998). In contrast to investigations to distinguish between a regulation of mRNA stability and transcription elongation, which are technically demanding, nucleocytoplasmic mRNA transport is readily accessible by northern blot analyses of cytoplasmic and nuclear RNA fractions (Carey & Smale, 2000).

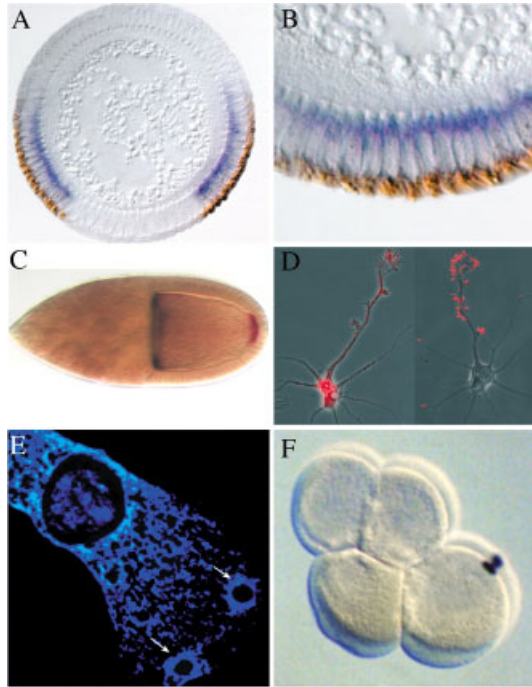
The regulation of mRNA transport of transcripts of the human immunodeficiency virus 1 (HIV-1) is an established example for that mechanism (Carey & Smale, 2000; Cullen, 1992). Primary transcripts of the human immunodeficiency virus 1 are efficiently spliced at an early stage of the viral life cycle. Desired proteins are encoded by the mature, processed transcripts. In the later viral life cycle, proteins are needed that are encoded by the unspliced transcript. The transport of these unspliced mRNAs to the cytoplasm is facilitated by the *Rev* protein that is encoded by the virus. *Rev* binds to a specific control element, the *Rev*-responsive element, that is located near to one of the splice acceptor sites of the HIV-1 pre-mRNAs. The interaction between *Rev* and RNA interferes with splicing and promotes the transport of the precursor transcript to the cytoplasm. It is suggested that *Rev* contains a leucine-rich nuclear export signal. This signal redirects the *Rev*/RNA complexes away from the typical mRNA processing pathway and toward a nuclear export pathway. The nuclear export receptor *CRM1* (chromosome region maintenance 1; a nuclear protein essential for proliferation and chromosome region maintenance, also called exportin 1) is involved in this export process (Fornerod *et al.*, 1997; Ohno *et al.*, 1998). It is not clear if any mRNAs are regulated by a similar mechanism.

1.4.7

Directed intracellular mRNA localisation

Following processing in the nucleus, mRNAs are transported into the cytoplasm where they are usually translated directly into proteins by ribosomes. It has recently emerged, however, that some mRNAs are targeted to defined locations in the cytoplasm before being translated. This raises the important point that mRNA localisation is often coupled to translational control. Indeed, proteins can be restricted to one region of a cell by localisation of their transcripts only if their mRNAs are not translated before or during this process of localisation (Palacios & Johnston, 2001). Many localised mRNAs are found in oocytes and early embryos (Johnstone & Lasko, 2001), where they function as localised determinants that control axis formation and development of the germline, and in neurons, where it may be involved in learning and memory (Palacios & St. Johnston, 2001) (see figure 1.13).

Figure 1.13 Examples of cytoplasmic localised mRNAs. **A)** and **B)** *brinker* mRNA (blue) localises basally in the *Drosophila* blastoderm embryo, whereas *short gastrulation* mRNA (brown) is concentrated apically. **C)** *bicoid* (black) and *oskar* (red) mRNAs localise to the anterior and posterior poles of the *Drosophila* oocyte. **D)** β -actin mRNA (left) and protein (right) localisation in the axonal growth cone of cultured neurons. **E)** Recruitment of total RNA (blue) to focal adhesion complexes after cell binding to extracellular matrix-coated beads (arrows). **F)** Localisation of *HrZF-1* mRNA to the posterior-vegetal cytoplasm in an eight-cell embryo of the ascidian *Halocynthia roretzi*. The figure is reproduced from Palacios & St. Johnston (2001) with permission from *Annual Review of Cell and Developmental Biology*.



The first localised mRNA was discovered almost 20 years ago when it was shown that β -actin mRNA accumulates in the myoplasm of Ascidian eggs (Jeffery *et al.*, 1983; Lee *et al.*, 1999). The mRNA encoding β -actin is restricted to the actin filament-rich cell cortex of mammalian fibroblasts. The reason for this restriction is probably that it is more efficient to localise an mRNA at a place in the cell at which the encoded protein is required than to move and compartmentalise the protein encoded by that mRNA.

In recent years, dramatic improvements in the resolution and sensitivity of *in situ* hybridisation techniques have led to the identification of a large number of localised mRNAs in many different animal, yeast and plant cells (Bouget *et al.*, 1996; Long *et al.*, 1997; Takizawa *et al.*, 1997). Indeed, several cells have been shown to localise multiple mRNAs to different positions. At least 20 mRNAs accumulate at either the animal or vegetal pole of the *Xenopus* egg, whereas *bicoid*, *oskar*, and *gurken* mRNAs localise to three distinct positions within the *Drosophila* oocyte (St. Johnston, 1994; Mowry & Cote, 1999). Furthermore, about ten percent of randomly chosen cDNAs recognise transcripts that localise within *Drosophila* germline cysts (Dubowy & Macdonald, 1998).

The importance of mRNA localisation for targeting proteins to a particular location in the cells depends on whether there is an alternative pathway for localising the protein to where it is required. Messenger RNA localisation is often partially redundant, as the encoded polypeptide may contain targeting signals that direct it to the same destination. In several cases, mRNA localisation is thought

not to be essential and seems to provide a backup mechanism for protein localisation (Palacios & St. Johnston, 2001). However, many instances also exist where protein localisation is achieved exclusively through the localisation of the corresponding mRNA and where mRNA localisation is therefore essential for important biological functions. Examples are asymmetric division that gives rise to two daughter cells with different cell fates, localisation of the cytoplasmic determinants that control the development of the regions of an embryo, control of the direction of cell signalling, determination of cell polarity and synaptic plasticity or, more generally, exclusion of proteins from certain regions of a cell (Palacios & St. Johnston, 2001).

Several mechanisms for the restricted localisation of mRNA to certain areas of the cytoplasm have been suggested (Palacios & St. Johnston, 2001): (i) diffusion and local anchoring, (ii) localised degradation, (iii) localised synthesis, (iv) active transport, (v) actin-dependent transport, and (vi) multistep localisation.

Diffusion and local anchoring. The localisation of *nanos* mRNA in the *Drosophila* oocyte is thought to occur by general diffusion followed by trapping at the posterior pole by a localised anchor (Palacios & St. Johnston, 2001). Thus, targeting of *nanos* mRNA requires that components of the cytoplasm first localise to the posterior pole (Wang *et al.*, 1994). Although an active transport mechanism cannot be ruled out, this seems unlikely. Several other transcripts also localise to the pole plasma, and it is possible that this localization operates by the mechanism suggested for *nanos* mRNA (Jongens *et al.*, 1992; Nakamura *et al.*, 1996; Palacios & St. Johnston, 2001; Raff *et al.*, 1990).

Localised degradation. Maternal *hsp83* mRNA, which is uniformly distributed in the fertilised egg of *Drosophila*, is a typical example of an mRNA with restricted cytoplasmic localisation (Palacios & St. Johnston, 2001). *Hsp83* mRNA forms a contracting posterior gradient as the nuclear division cycles proceeds, and eventually becomes restricted to the pole cells when they form at the posterior of the embryo after nuclear cycle ten (Ding *et al.*, 1993). Even though the total amount of *hsp83* mRNA decreases dramatically during embryogenesis, the amount of transcript at the posterior pole remains constant, suggesting that localisation is achieved by RNA degradation and local protection at the posterior end (Bashirullah *et al.*, 1999). *Hsp83* mRNA is probably stabilised at the posterior pole by components of the pole plasma since it can be induced at the anterior pole in embryos if ectopic plasma from the posterior end is introduced at this pole.

Localised synthesis. Localised transcription plays a role in the localisation of *gurken* mRNA in the *Drosophila* oocyte (Neuman-Silberberg & Schüpbach, 1993). *Gurken* is the only gene transcribed in the oocyte nucleus that is transcriptionally quiescent and arrested in meiotic prophase. This probably contributes to the localisation of the mRNA to the apical side of the nucleus (Saunders & Cohen, 1999). The first steps in patterning occur at the syncytial blastoderm stage, when the embryo is still a single cell because there are no membranes between the adjacent nuclei (Palacios & St. Johnston, 2001).

Active transport. The mechanism that is most commonly invoked to explain intracellular mRNA localisation is active transport along the cytoskeleton, but this is

surprisingly difficult to demonstrate. In general, there are two ways to show that an mRNA is localised by active transport. First, one may observe the directed movement of the RNA *in vivo*, either by injecting labelled RNA or by tagging the transcript with a labelled protein (see chapter 6). However, this type of motion can also be generated by directed cytoplasmic flows that passively move the RNA with the rest of the cytoplasm, and this possibility needs to be ruled out before concluding that active transport is occurring. A second approach is to identify a motor protein that is required for the localisation of the RNA. Motor proteins can also contribute to mRNA localisation indirectly, for example, by transporting proteins that will anchor the RNA at the correct site or by generating cytoplasmic flows, and it is therefore also essential to prove that the motor is directly linked to the mRNA. The most convincing example is *Ash1* mRNA in yeast, which is transported along actin cables by the myosin motor *Myo4p* (reviewed by Chartrand *et al.*, 2001). Nevertheless, there is strong evidence that a number of mRNAs in higher organisms are localised by active transport, although the majority of these require microtubules rather than actin (Palacios & St. Johnston, 2001).

Actin-dependent transport. Unlike *Ash1* mRNA in yeast, it seems that most transcripts in higher eukaryotes that localise *via* active transport move along microtubules rather than actin. One possible exception is *prospero* mRNA, which moves from the apical to the basal side of *Drosophila* neuroblasts as they enter mitosis (Palacios & St. Johnston, 2001). The actin cytoskeleton does not show any obvious polarity in these cells, however, and it is possible that a myosin anchors the *prospero* mRNA complex to the cortical actin cytoskeleton rather than transporting it. β -actin mRNA localisation to the leading edge of motile fibroblasts is also actin-dependent, but nothing is known about the mechanism in this case (Hill & Gunning, 1993; Kislauskis *et al.*, 1993). The small number of transcripts whose localisation is actin-dependent may reflect the different stability and organisation of the two types of cytoskeleton in most cells. Since microtubules form stable and long-range structures in interphase, they are ideally suited for transport over long distances in polarised differentiated cell types, whereas actin filaments tend to be shorter and more dynamic. If mRNAs are actively transported along actin, this would probably be over short distances (Palacios & St. Johnston, 2001).

Multistep localisation. Some mRNAs are localised by more than one of the mechanisms discussed above because their localisation occurs either in several steps or by overlapping and partially redundant pathways (Palacios & St. Johnston, 2001). For example, the initial accumulation of *nanos* mRNA at the posterior of the *Drosophila* oocyte is mediated by local anchoring in the pole plasma, but this localised RNA is also stabilised after fertilisation, whereas the unlocalised transcripts are rapidly degraded (Bashirullah *et al.*, 1999).

Although mRNAs can be localised by a variety of mechanisms, two aspects of this process are universal. First, the transcript must contain *cis*-acting elements that direct its localisation. Second, these elements must be recognised by specific *trans*-acting factors that somehow link the RNA to one of the localisation mechanisms described above (Palacios & St. Johnston, 2001).

Cis-acting elements. The intracellular localisation of mRNAs is often coupled to their translational regulation. Messenger RNAs that are targeted to certain location within the cytoplasm often contain signal sequences (*cis*-acting elements) typically located in the 3'-untranslated region of the mRNA (Alberts *et al.*, 1998; Palacios & St. Johnston, 2001). An overview of the known *cis*-acting elements is given in Palacios & St. Johnston (2001). Unlike double-stranded DNA, single-stranded RNA molecules have the capacity to fold into elaborate secondary and tertiary structures, and this is likely to be important for the function of many of the larger and more complex localisation signals. Indeed, one reason why localisation elements almost always reside in 3'-untranslated regions may be that these regions of the mRNA can be structured without disrupting translation. The structure of only a few localisation elements has been investigated (Palacios & St. Johnston, 2001).

Trans-acting factors. Only a few *trans*-acting factors involved in mRNA localisation have been identified, in most cases, by biochemically purifying RNA-binding proteins that recognise *cis*-acting localisation elements. Heterogeneous nuclear riboproteins (hnRNPs) are a typical example of *trans*-acting factors involved in the intracellular mRNA localisation. Heterogeneous nuclear riboproteins are abundant nucleic acid-binding proteins that have a variety of functions in the nucleus. However, a subset of these proteins shuttle into the cytoplasm and can play important roles in the regulation of the translation, stability, and cellular localisation of specific cytoplasmic mRNAs (Shyu & Wilkinson, 2000). Current data implicate hnRNPs in cytoplasmic mRNA localisation, but it is unclear if they play a direct role in coupling the mRNAs to the transport machinery or if they rather act as RNA chaperones that fold the RNA so that it can be recognised by other factors (Palacios & St. Johnston, 2001). The role of heterogeneous nuclear riboproteins in mRNA localisation suggests that the assembly of specific RNA/protein complexes in the nucleus may determine the fate of the mRNA in the cytoplasm.

It should be recalled that the *cis*-acting elements that regulate translation often overlap with the localisation signals, and some of the *trans*-acting factors that bind to the 3'-untranslated regions of localised mRNAs regulate both protein translation and mRNA localisation (Palacios & St. Johnston, 2001).

While the inventory of localised transcripts increases year on year, progress in understanding the mechanisms of mRNA transport has been comparatively slow, largely because this is a difficult process to analyse. Unlike other intracellular trafficking events such as nucleocytoplasmic transport or protein secretion, the cytoplasmic localisation of mRNAs does not involve movement between membrane-bound compartments that can be purified by fractionation, and the specific mRNAs that localise to a particular site represent only a small fraction of the total mRNA in the cell (Palacios & St. Johnston, 2001). Therefore, it is still not possible to estimate what proportion of mRNAs are localised in any particular cell type and it is now clear that the intracellular localisation of specific mRNAs occurs in most if not all polarised cells and provides an important mechanism for targeting proteins to the cellular regions where they are required.

For detailed information on mRNA localisation and translation the reader is referred to Johnstone & Lasko (2001) and Palacios & St. Johnston (2001).

1.4.8

Regulation of mRNA polyadenylation

Although rare, sufficient examples for the regulation of mRNA abundance by polyadenylation have been reported and establish this process as a further regulatory mechanism of mRNA abundance within eukaryotic cells (Carey & Smale, 2000). Differential polyadenylation has, for example, been reported for the immunoglobulin (Ig) heavy-chain locus (Colgan & Manley, 1997). In mature B cells the immunoglobulin heavy-chain protein *IgM* is synthesised as a plasma membrane-binding protein that is called the μm form of the *IgM* protein. During differentiation of mature B cells into plasma cells, an *IgM* protein form, called μs , is produced with an alternative carboxy-terminal domain lacking the transmembrane domain of the μm protein form. The μs form of the *IgM* protein is secreted by the plasma cells. The carboxy terminus of the μs protein form is encoded by an exon that resides upstream of the exons encoding the transmembrane domain of the μm protein form. In mature B cells, the exon encoding the alternative carboxyl terminus of the μs form is excised by splicing of the pre-mRNA. During processing, the mRNA is polyadenylated downstream of the μm exons. By contrast, in plasma cells, a second polyadenylation signal between the μs and μm exons is activated leading to the generation of polyadenylated mRNAs containing the μs exons and lacking the μm exons. The μs polyadenylation signal seems to be relatively weak. It is activated only in plasma cells, in which a general polyadenylation factor, called *CstF-64*, is synthesised at a relatively high level (Colgan & Manley, 1997).

Additional studies have revealed that the length of the poly(A) tail of some mRNAs is controlled in a specific manner by the selective addition or removal of poly(A) oligonucleotides in the cytoplasm (Alberts *et al.*, 1998). mRNA stability is thought to be dependent on the length of the poly(A) tail (see figure 1.14) (Wickens, 1990). Therefore, mechanisms for modulating the length of the mRNA poly(A) tails may be of importance for the control of the translation process by regulating the mRNA stability in the cytoplasm.

1.4.9

Antisense RNA

The ability of antisense RNA to 'neutralise' mRNA was first described in *Caenorhabditis elegans* (Bastin *et al.*, 2001) in which it was demonstrated that the presence of RNA complementary to a given mRNA led to a loss of gene function. In addition, double-stranded antisense RNA was also found to exert a pronounced negative effect on translation. This phenomenon was termed post-transcriptional gene silencing and was found to be related to the presence of 'abnormal' or 'unusually high' amounts of RNA. Presumably, this system developed during evolution in order to resist viral RNA or to protect the cellular genome from damage by trans-

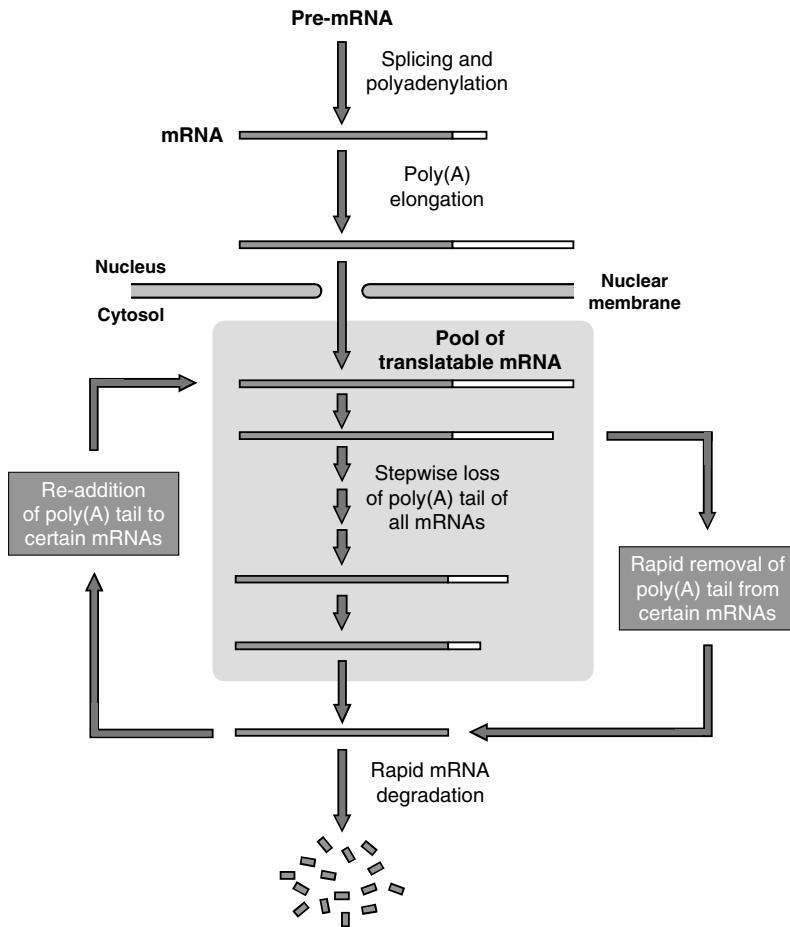


Figure 1.14 Control of polyadenylation and length of the poly(A) tail allows the modulation of mRNA stability and translation of mRNAs into proteins in eukaryotic cells. Most of the mRNAs that are translated into proteins have a poly(A) tail consisting of at least 30 nucleotides. The tail of certain mRNAs is modulated in length by elongation or truncation. This

modification of the length of poly(A) tails affects also the translation efficiency. It is thought that the large subunits of the ribosomes bind to proteins by interacting specifically with the long poly(A) tails of the mRNAs. This process enhances a new translation cycle using again the mRNA molecule as a template. The figure is adapted from Alberts *et al.* (1998).

positions (Bastin *et al.*, 2001; Cogoni & Macino, 2000). A simple mechanism by which antisense RNAs may operate is as follows. The foreign RNA is amplified by RNA polymerases and degraded into small pieces of 20 to 25 nucleotides. These RNA pieces bind to complementary mRNA sequences by Watson-Crick base pairing and tag them for RNase degradation (see the following references for a more detailed review of presumable mechanisms: Boshier & Labouesse,

2000; Grant, 1999; Hammond *et al.*, 2001). Recent approaches in therapeutic applications of nucleic acids (gene therapy) use antisense technology and post-transcriptional gene silencing by RNA interference in order to switch off gene activity. The expression of virtually any gene can be modified by using synthetic antisense oligonucleotides (Bosher & Labouesse, 2000; Grant, 1999; Hammond *et al.*, 2001). Several oncogene mRNA have been targeted by antisense oligonucleotides in cancer therapy. The results of the first clinical trials demonstrated a low toxicity, but only a moderate effect on tumour progression (for reviews, see Flaherty *et al.*, 2001; Golden *et al.*, 2002).

1.4.10

RNA editing

RNA editing by modification of nucleotides in order to change the nucleotide sequence of mRNA transcripts is a distinct mechanism to produce a greater variety of proteins in different cells and to control the expression of certain gene products (Chan, 1993; Landweber & Gilbert, 1993; Sloof & Benne, 1993; Sollner-Webb, 1991 and 1996). The mechanism of RNA editing was found for the first time in mRNA transcripts encoding mitochondrial proteins of trypanosomes (Benne, 1990; Benne *et al.*, 1983).

RNA editing is the insertion (or less frequently deletion) of one or more uracil nucleotides in a transcript. This often results in a frameshift of the open reading frame, thus modifying the message, *i.e.* the encoded protein, of the transcript. In some cases, RNA editing is very extensive so that more than the half of the transcript consists of uracil (Alberts *et al.*, 1998). The trypanosome RNA editing is a massive processing of mRNA by uracil deletion and insertion, directed by *trans*-acting guide RNAs (gRNAs). These gRNAs are 40 to 80 nucleotides long and are transcribed separately. Guide RNAs contain a 5'-sequence that is complementary to a part of the sequence of the transcript that is to be modified. Located downstream to this 5'-sequence is a section specifying the set of uracil nucleotides that are to be inserted in the mRNA. A section of a single (or multiple) uracil nucleotide(s) that is to be inserted into the transcript is located further downstream.

RNA editing requires two RNA ligases. A complex of seven different polypeptides purified from *Trypanosoma brucei* mitochondria that catalyses accurate RNA editing has been identified (Rusche *et al.*, 2001). The mechanism of RNA editing is surprisingly complex. The uracil nucleotides located at the 3'-end of the gRNAs are transferred directly into the transcript (see figure 1.15).

The processing of mRNAs by RNA editing is not limited to the mitochondria of trypanosomes. RNA editing has been described for mRNAs of mitochondria of several plants in which nearly every mRNA is modified to a certain degree by RNA editing. In the mitochondria and chloroplasts of higher plants, RNA editing is responsible for specific cytosine to uracil conversions and for a few uracil to cytosine conversions leading to RNA sequences different from the corresponding DNA sequences without insertion or deletion of nucleotides (Gray, 1996; Gray & Covello, 1993; Maier *et al.*, 1996).

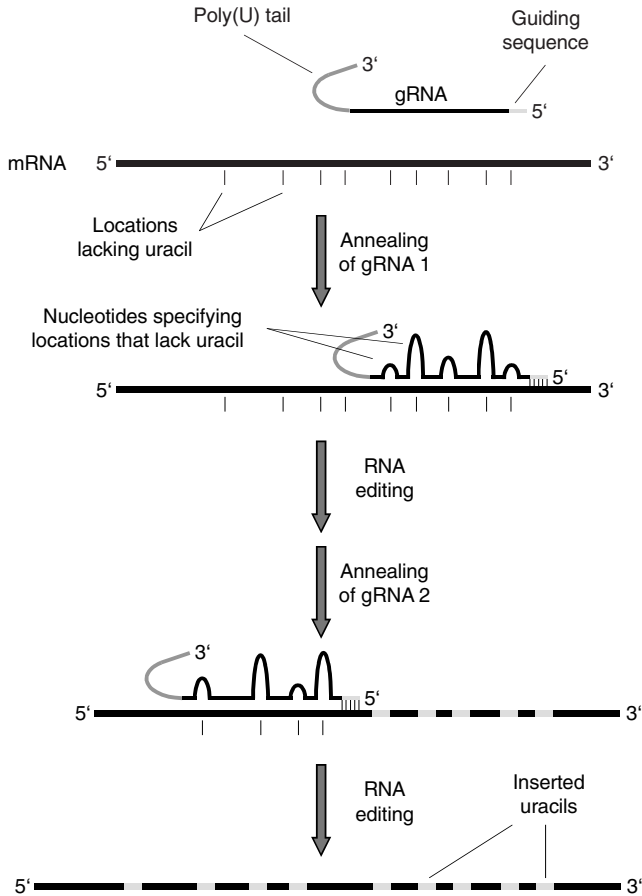


Figure 1.15 Mechanism of RNA editing in mitochondria of trypanosomes. *Trans*-acting guide RNAs (gRNAs) contain a poly(U) tail at the 3'-end. The uracil of the gRNA is inserted into the mRNA transcript at sections of the mRNA that are unable to anneal with the gRNA. During the insertion, the poly(U) tail is shortened by inserting its uracil into the mRNA. Usually, the

processing starts at the 3'-end of the mRNA and is continued in the direction to the 5'-end. This is based on the fact that the guiding sequences of most gRNAs anneal only with the sequences of the mRNA that have not been modified by RNA editing. The figure is adapted from Alberts *et al.* (1998).

It is not clear why RNA editing is so widely used in the mitochondria of plants and trypanosomes. It has been suggested that this mechanism is a relict of a primitive genetic system in mitochondria. There is evidence that RNA editing is essential for correct gene expression and modulates transcriptional activity (Maier *et al.*, 1996). Trypanosomes are primordial single cell eukaryotes. The trypanosome taxon separated very early from the line of life forms leading to the evolution of plants, yeasts and animals (Alberts *et al.*, 1998). To date changes mediated by

RNA editing have been observed in tRNA, rRNA and mRNA molecules of eukaryotes, but not prokaryotes (Brennicke *et al.*, 1999); but the demonstration of RNA editing in prokaryotes may only be a matter of time, considering the range of species in which the various RNA editing processes have been found.

It is noteworthy that proteins translated from edited transcripts differ from the ones deduced from the genes sequences and usually present higher similarity to the corresponding non-plant homologues (Maier *et al.*, 1996). Moreover, RNA editing in kinetoplast-mitochondria of trypanosomes and related protozoa creates translatable mRNAs out of nonsense pre-edited RNAs by correcting encoded defects that vary from simple frameshifts to large 'cryptic' regions. After decades of searching by many groups for evidence of translation-edited mRNA in kinetoplast-mitochondria, Horvath *et al.* (2000) identified the kinetoplast-encoded apocytochrome b whose mRNA is edited in the 5'-region and is translated into a functional protein.

RNA editing has a manifold role of physiological importance including regulation of gene expression, creation of new initiation and stop codons and stabilisation of the secondary structure of introns and tRNAs (Brennicke *et al.*, 1999). RNA editing in mammals is much more simple but also has an important physiological effect. The first case of RNA editing to be described in mammals was the creation of a novel stop codon (cytosine conversion to uracil) in the mRNA of apolipoprotein B that results in the formation of two proteins of different length (apolipoprotein B-48 and apolipoprotein B-100) (Chen *et al.*, 1987; Powell *et al.*, 1987). The formation of apolipoprotein B-48 is tissue-specific and is mediated by an enzyme catalysing the hydrolytic deamination of cytosine into uracil. Although other cases of RNA editing in mammals have been reported (Smith & Sowden, 1996), the mechanism of this process is not known.

1.4.11

Summary and conclusions

Eukaryotic gene expression is a multistep pathway that involves a series of highly regulated events that take place in both nucleus and cytoplasm. The regulation of this process is a complex interplay that allows for sensitive fine-tuning at several stages. In the first step, genes are transcribed in the nucleus into pre-mRNAs after remodelling of chromatin. The initiation of the process of transcription is probably the primary regulator of gene expression. Genes encode not only for the pre-mRNA sequence which in turn encodes for at least one protein, they also contain regulatory elements such as a proximal upstream promoter sequence which binds the basal transcription factors and distant enhancer or repressor sequences, which in turn bind inducible transcription factors. Basal transcription factors are proteins that form the initiation complex with the RNA polymerase at the promoter. Inducible transcription factors bind to short DNA sequences termed response elements. In general, the binding of such factors enhances the formation of the transcription initiation complex. Repression is also possible, but is rare. Several mechanisms for the activation of inducible transcription factors have been de-

scribed. Transcription factors are activated to bind their target response elements by (i) protein kinase-mediated phosphorylation, (ii) binding of activating or removal of inhibitory factors, or (iii) *de novo* protein synthesis.

Pre-mRNAs are then processed in the nucleus and the resulting mature mRNAs are transported to the cytoplasm, where they are translated into protein and degraded at a rate dictated by transcript and cell-type-specific cues.

Whereas the main level of transcriptional control is the initiation of RNA synthesis, in some instances, genes are also regulated by, for example, alternative splicing of the primary transcript, transport of the mature mRNA to the cytoplasm, stability of the transcript and controlled degradation, or control of translation into proteins, *etc.* Large-scale silencing of genes is also mediated by the packing of DNA into highly condensed heterochromatin structures and DNA methylation at cytosines in defined sequences in guanine and cytosine (Kozak, 2001b).

Although an extraordinary amount of regulation at several levels goes into making sure that gene expression is controlled and timed correctly, this is not the whole story. The cell must also ensure that genes are properly transcribed and translated. For example, during translation, the genetic information in mRNA is used to synthesise proteins. tRNAs charged with an amino acid move to the ribosome, where they are paired with the corresponding nucleic acid codon in the mRNA. The nascent polypeptide grows by attaching the amino acid and the ribosome moves on to the next codon. This cycle is repeated to produce a full-length protein. Several proofreading and editing processes are used throughout protein synthesis to ensure the faithful translation of genetic information. The phenomenal accuracy of transcription and translation is ensured at several steps including (i) monitoring of the maturation of tRNAs and mRNAs, and of the identity of amino acids attached to tRNAs, and (ii) selection of aminoacyl-tRNAs on the ribosome and correct base pairing of these aminoacyl-tRNAs with mRNA. Important studies of the molecular mechanisms underpinning the high quality control of these processes have recently been published (Ibba & Soll, 1999).

Proteins emerging from the ribosome must fold into three-dimensional structures and maintain that structure in order to fulfil their tasks. Maintaining quality control over protein structure and function depends on molecular chaperones and proteases. Molecular chaperones promote the proper folding of proteins and prevent aggregation, whereas proteases eliminate irreversibly damaged proteins (Wickner *et al.*, 1999). A detailed description of the cellular quality control of translation and post-translational protein processing goes beyond the scope of this book, however.

The correct spatial and temporal abundance of proteins is regulated by a complex but flexible system. The mechanisms described in this section give a brief introduction to the complexity of the cellular regulation of gene expression that allows a rapid but fine-tuned response of the cell to external and internal influences.

Further mechanisms affecting translation and transcription efficiency have also been reported (Alberts *et al.*, 1998; Johnstone & Lasko, 2001; Kozak, 2001; Preiss & Hentze, 1999) and can only be alluded to here. The interested reader is referred to the specialist literature. Good general reviews are provided by Fry & Peterson (2001

and 2002), Kozak (2001a and 2001b), Maquat & Carmichael (2001), Preiss & Hentze (1999) Schneider *et al.* (2001) and Soutoglou & Talianidis (2002). For details of special regulatory mechanism, the reader is referred to the cited literature. Such special cases of mechanisms not described above include: (i) Inhibition of the binding of the methionylated starting tRNA to the small ribosome subunit by phosphorylation of the initiation factor eIF-2 (Alberts *et al.*, 1998; de Haro *et al.*, 1996; Dever, 1999; Hinnebusch, 1990; Rhoads, 1993; Sarre, 1989). (ii) Negative control of translation by mRNA masking, *e.g.*, inhibition of ferritin translation by binding of a translation repressor protein, aconitase, to regulatory elements near the 5'-end of the ferritin mRNA (Alberts *et al.*, 1998; Melefors & Hentze, 1993; Rouault & Klausner, 1997; Sommerville & Lodomery, 1996). (iii) Conservation of mRNA sequences in the starting regions of protein translation in prokaryotes and eukaryotes, *e.g.*, the Shine-Dalgarno sequence, the Kozak sequence or internal ribosome entry sites (IRES) that influence the efficiency of translation (Alberts *et al.*, 1998; Kozak, 1977 to 1999; Kozak & Shatkin, 1978; Lindahl & Hinnebusch, 1992; Oh & Sarnow, 1993; Shine & Dalgarno, 1975; Vagner *et al.*, 2001). (iv) Circularisation of mRNA (the closed-loop model) in which the 3'-poly(A) tail of mRNAs participates in translation initiation, acting synergistically with the 5'-cap structure of mRNA (Alberts *et al.*, 1998; Preiss & Hentze, 1999). (v) Translation recoding of retroviral mRNAs mediated by a shift of the open reading frame of the mRNA that results in the production of a different protein (Alberts *et al.*, 1999; Hatfield & Oroszlan, 1990; Weiss, 1991).

1.5

Post-translational modification of proteins

1.5.1

Introduction

It is important to recall that the amount of mRNA or protein expressed in a cell is not always the factor determining its function or activity. Nevertheless, many examples exist in which the level of expression of mRNA or protein correlates with a physiological effect. A good example is the high-density lipoprotein-mediated cholesterol efflux from macrophages which is controlled by the amount of ATP-binding cassette transporter A1 (ABCA1) protein present in the cell (Chen *et al.*, 2001; Santamarina-Fojo *et al.*, 2001; Venkateswaran *et al.*, 2000; Wang *et al.*, 2001). It is, however, unclear if the cholesterol is transported directly *via* the ATP-transporter or if the level of ABCA1 mRNA or protein influences the amount of cholesterol taken up by high-density lipoproteins (Attie *et al.*, 2001; Haghpasand *et al.*, 2001). In addition to the amount of protein present, post-translational modifications are now known to play a fundamental role in regulating the activity, intracellular or extracellular location, physiological function and degradation or processing of a wide range of proteins.

Following synthesis on the ribosomes, proteins are modified in many ways. They may be cleaved (thus, eliminating signal sequences, pro-peptides and initiator methionines) and may be decorated with a wide variety of simple chemical groups (e.g., by acetylation, methylation or phosphorylation), or more complex molecules, such as sugars and lipids. Finally, they can be internally or externally cross-linked *via* disulfide bonds or lysinonorleucine. More than a hundred different types of post-translational protein modifications are currently known (Alberts *et al.*, 1998) and it is estimated that 50 to 90 percent of the proteins in the human body are post-translationally modified (Doyle & Mamula, 2001). Thus, the number of different protein molecules expressed by the human genome may be of the order of several hundred thousand, *i. e.* many times the number of genes that the human genome is thought to contain.

Why are proteins processed after translation? The answer to this question is both simple and difficult. On the one hand, in the case of proteins in general, it is obvious that post-translational processing increases the functional flexibility of a cell because these modifications add functionality, affect cellular targeting, regulate activity, modulate mechanical strength, change recognition behaviour, *etc.* of proteins. However, to identify the nature and function of post-translational modification in a particular case is labour and time-intensive. Current proteomic techniques that facilitate protein separation, identification and characterisation of proteins should advance our understanding of these processes.

In the following sections some of the well-known post-translational protein modifications are described in brief, whereas chapter 5 provides an overview of the methods available for the analysis of protein expression.

1.5.2

Proteolytic cleavage of proteins

Several proteins undergo proteolytic cleavage following translation. In its simplest form this is the removal of the start methionine. Many proteins are synthesised as inactive precursors that are activated under physiological conditions by limited proteolysis. Pancreatic enzymes and enzymes involved in clotting are examples of the latter. Inactive precursor proteins that are activated by removal of polypeptides are termed proproteins.

Proteins that are membrane-bound or are destined for excretion are synthesised by ribosomes associated with the membranes of the endoplasmic reticulum. Members of this class of proteins all contain an amino terminal signal sequence or signal peptide. The signal peptide usually consists of 13 to 36 predominantly hydrophobic residues. It is recognised by a multi-protein complex termed the signal recognition particle and is catalytically removed by signal peptidase following passage through the endoplasmic reticulum membrane. Proteins that contain a signal peptide are called preproteins. Some proteins that are destined for secretion also contain pro-sequences and undergo further proteolytic cleavage. These are called preproproteins (Alberts *et al.*, 1998).

A classic example of post-translational processing of a preproprotein is insulin. Since insulin is secreted from the pancreas, it is a prepeptide. Following cleavage of the 24 amino acid signal peptide, the protein folds into proinsulin. Proinsulin is further cleaved yielding active insulin, which is composed of two peptide chains linked together by disulfide bonds (Dodson & Steiner, 1998).

1.5.3

Acylation

Many proteins are post-translational acylated at their amino termini. In most cases, the initiator methionine is replaced by an acetyl group. Acetyl-coenzyme A is the acetyl donor for these reactions. Some proteins have the myristoyl group added to their N-termini which is derived from myristoyl-coenzyme A (Farazi *et al.*, 2001). This latter modification allows association of the modified protein with membranes. Typical examples for acetylated proteins are the catalytic subunit of cyclic AMP (cAMP)-dependent protein kinase (PKA) which is myristoylated (Johnson *et al.*, 1994), some members of the G protein family that can be myristoylated or palmitoylated (Chen & Manning, 2001), p53 tumor suppressor protein (Prives & Manley, 2001) and the four histones of the nucleosome core particle (Turner, 1998).

1.5.4

Prenylation

Prenylation refers to the addition of the farnesyl or geranylgeranyl groups to proteins. Both of these are isoprenoid compounds derived from the cholesterol biosynthetic pathway. The isoprenoid groups are attached to cysteine residues at the carboxy terminus of proteins in a thioether linkage. A common consensus sequence at the carboxy terminus of prenylated proteins has been identified and is composed of CAAX, where C is cysteine, A is any aliphatic amino acid (except alanine) and X is the carboxy terminal amino acid. In order for the prenylation reaction to occur, the three carboxy terminal amino acids (AAX) are first removed and the cysteine is activated by methylation in a reaction utilising S-adenosylmethionine as the methyl donor (Harris & Poulter, 2000).

Important examples of prenylated proteins include the oncogenic GTP-binding and hydrolysing protein Ras and the γ -subunit of the visual protein transducin, both of which are farnesylated. Numerous GTP-binding and hydrolysing proteins (G proteins or GTPases) of signal transduction cascades have γ -subunits modified by geranylgeranylation (Higgins & Casey, 1996; Sinensky, 2000a). In addition, small GTPases of the Rab family regulate membrane transport pathways and must associate with cellular membranes for activity. Their membrane attachment is also mediated by prenyl (geranylgeranyl) post-translational modification (Pereira-Leal *et al.*, 2001).

Post-translational modification with geranylgeranyl or farnesyl substituents was once thought of solely as a mechanism for association of proteins with mem-

branes. In recent years, however, other functional aspects of protein prenylation have come to be appreciated. Although, in almost all instances, such proteins are membrane associated, they are often found also to engage in protein/protein interactions. In some instances, such interactions are critical aspects of prenylated protein trafficking (Sinensky, 2000).

1.5.5

Methylation

Post-translational methylation occurs at lysine and aspartate residues in some proteins such as calmodulin (Murtaugh *et al.*, 1986) and cytochrome c (Paik *et al.*, 1989). The activated methyl donor is S-adenosylmethionine. Stoichiometric methylation of hormones and hormone carrier proteins plays a role in hormone storage and secretion. Substoichiometric methylation at aspartyl residues is involved in a repair mechanism of aged proteins and stoichiometric methylation of calmodulin modulates the sensitivity of calmodulin-dependent processes to calcium (van Waarde, 1987).

1.5.6

Sulphation

Sulphate modification of many secreted and membrane-bound proteins occurs at tyrosine residues such as in fibrinogen (Liu *et al.*, 1985) and in gastrin (Varro & Dockray, 1993). The universal sulphate donor is 3'-phosphoadenosyl-5'-phosphosulphate (PAPS) and the reaction is catalysed by tyrosylprotein sulphotransferase, a membrane-bound enzyme of the *trans*-Golgi network. Sulphate is added permanently and is necessary for the biological activity of the protein and is not used as a regulatory modification like that of tyrosine phosphorylation (Han & Martinage, 1992). However, sulphation is the most abundant post-translational modification of tyrosine residues and occurs in many soluble and membrane proteins passing through the secretory pathway of metazoan cells. Tyrosine sulphation has been shown to be important for protein/protein interactions occurring during the intracellular transport of proteins and upon their secretion (Huttner, 1988; Niehrs *et al.*, 1993), in particular involved in leukocyte adhesion, haemostasis and chemokine signalling (Kehoe & Bertozzi, 2000).

1.5.7

Phosphorylation

Post-translational phosphorylation is one of the most common protein modifications that occurs in animal cells. The vast majority of phosphorylations act as a mechanism to control the biological activity of proteins in many pathways and as such are reversible. The enzymes that phosphorylate proteins are termed kinases and those that remove phosphates are termed phosphatases. In animal cells, serine, threonine and tyrosine are the amino acids subject to phosphorylation. The largest group of kinases are those that phosphorylate either serine or threonine and are

therefore termed serine/threonine kinases. The ratio of phosphorylation of the three different amino acids is approximately 1,000 to 100 to 1 for serine to threonine to tyrosine (Holland & Cooper, 1999; Pallen *et al.*, 1992).

Although the level of tyrosine phosphorylation is minor, the importance of phosphorylation of this amino acid is profound. As an example, the activity of numerous growth factor receptors is controlled by tyrosine phosphorylation. Further physiologically relevant examples are the phosphorylations that occur in glycogen synthase and glycogen phosphorylase in hepatocytes in response to glucagon release from the pancreas (Aggarwal *et al.*, 1995), the phosphorylations of members of the *Bcl2* family of proteins that are important regulators of programmed cell death pathways with individual members that can suppress or promote apoptosis (Ruvolo *et al.*, 2001) and the phosphorylations of the transcription factor p53 (Jimenez *et al.*, 1999).

1.5.8

Ubiquitination

Ubiquitination of cellular proteins plays an important role in a variety of biological processes. Among these are regulation of gene expression, regulation of cell cycle and division, involvement in the cellular stress response, modification of cell surface receptors, DNA repair, import of proteins into mitochondria, uptake of neurotransmitter precursors into synaptosomes, biogenesis of peroxisomes, assembly of ribosomes, targeting for non-lysosomal proteolytic pathway *via* the proteasome, and programmed cell death (apoptosis) (Ben-Neriah, 2002; Ciechanover, 1994; Hasselgren & Fischer, 1997; Hershko & Ciechanover, 1998; Hochstrasser, 1996, 2000; Maniatis, 1998; Plemper & Wolf, 1999; Scheffner *et al.*, 1998; Varshavsky, 1997; Varshavsky *et al.*, 2000; Wilkinson, 1999). The mechanisms that underlie these complex processes and the precise role of ubiquitin are poorly understood.

The best studied modification occurs in the ubiquitin-mediated non-lysosomal proteolytic pathway directing to the proteasome (Ciechanover, 1993 and 1994; Ciechanover & Schwartz, 1994). Recent experimental evidence indicates that the ubiquitin system is involved in the degradation of mitotic cyclins, oncoproteins and tumour suppressors, in the removal of abnormal, incomplete or otherwise defective proteins, and in the processing of antigens restricted to class I major histocompatibility molecules. Degradation of a protein *via* the ubiquitin system involves two discrete steps taking place in the cytoplasm. Initially, multiple copies of the highly conserved ubiquitin molecule are covalently linked in an ATP-dependent fashion to the protein substrate. The targeted protein is then degraded by a specific, energy-dependent and high-molecular mass protease complex into peptides and free amino acids. Ubiquitin is released and can be reutilised (Alberts *et al.*, 1998; Ciechanover, 1993 and 1994; Ciechanover & Schwartz, 1994). In addition, stable mono-ubiquitin adducts are also found in the cell, for example, those involving nucleosomal histones (Ciechanover, 1993 and 1994; Ciechanover & Schwartz, 1994).

Despite the considerable progress that has been made in elucidating the mode of action and roles of the ubiquitin/proteasome system, many problems remain unsolved. For example, little is known of the signals that target proteins for degra-

dation by this pathway. While a few proteins are targeted for degradation following recognition of their amino terminal amino acid residue, the vast majority of cellular proteins are targeted by other signals. The identity of the native cellular substrates of the ubiquitin/proteasome system is another important, yet unresolved problem. Only a few proteins have been recognised so far as substrates of the proteasome system *in vivo*.

1.5.9

Glycosylation

Glycosylation is the covalent modification of proteins with carbohydrate. The resulting glycosylated proteins are termed glycoproteins. The predominant sugars found in glycoproteins are glucose, galactose, mannose, fucose, N-acetyl-D-galactosamine (GalNAc), N-acetyl-D-glucosamine (GlcNAc) and sialic acid (N-acetyl-neuraminic acid; NANA). The distinction between proteoglycans and glycoproteins resides in the level and types of carbohydrate modification. The carbohydrate modifications found in glycoproteins are rarely complex compared to proteoglycans.

Two basic types of glycosylation have been identified. The glycosidic linkage is either to the amide group of asparagine (N-linked glycosylation, N-glycosidic bond) or to the hydroxyl group of serine, threonine or hydroxylysine (O-linked glycosylation, O-glycosidic bond). It is noteworthy that the linkage of carbohydrate to 5-hydroxylysine is generally found only in the protein family of collagens. The only carbohydrates linked to 5-hydroxylysine are either the single sugar galactose or the disaccharide glucosylgalactose, whereas in serine and threonine-type O-linked glycoproteins, the carbohydrate directly attached to the protein is N-acetyl-D-galactosamine and in N-linked glycoproteins it is N-acetyl-D-glucosamine.

The predominant carbohydrate attachment in glycoproteins of mammalian cells is *via* N-glycosidic linkage. The site of carbohydrate attachment to N-linked glycoproteins is found within a consensus sequence of amino acids, N-X-S(T), where X is any amino acid except proline, S is serine and T is threonine. N-linked glycoproteins all contain a common core of carbohydrate attached to the polypeptide. This core consists of three mannose residues and two N-acetyl-D-glucosamine. A variety of other sugars are attached to this core and comprise three major N-linked families. (i) High-mannose type glycoproteins contain all mannose outside the core in varying amounts. (ii) Hybrid type glycosylated proteins have various sugars and amino sugars attached. (iii) Complex type glycoproteins are similar to the hybrid type, but, in addition, contain sialic acids to varying degrees (Alberts *et al.*, 1998). The major distinguishing feature of the complex class is the presence of sialic acid, whereas the hybrid class contains no sialic acid.

Most proteins that are secreted or bound to the plasma membrane are modified by carbohydrate attachment. The part that is modified in plasma membrane-bound proteins, is the extracellular portion. Intracellular proteins are less frequently modified by carbohydrate attachment.

The protein component of all glycoproteins is synthesised by polyribosomes that are bound to the endoplasmic reticulum. Processing of the sugar groups occurs co-

translationally in the lumen of the endoplasmic reticulum and continues in the Golgi apparatus for N-linked glycoproteins. This is in contrast to the attachment of sugars in O-linked glycoproteins, which occurs post-translationally in the Golgi apparatus. The synthesis of O-linked glycoproteins is catalysed by specific glycoprotein glycosyltransferases *via* the stepwise attachment of nucleotide-activated sugars directly onto the polypeptide (Alberts *et al.*, 1998).

In contrast to the step-wise addition of sugar groups to the O-linked class of glycoproteins, N-linked glycoprotein synthesis requires the lipid intermediate dolichol phosphate. Dolichols are polyprenols containing 17 to 21 isoprene units, in which the terminal unit is saturated. The oligosaccharide unit is attached to dolichol phosphate through a pyrophosphate bond in a stepwise reaction. The sugars used for N-linked glycoprotein synthesis are activated by coupling to nucleotides, as in the synthesis of O-linked glycoproteins (Alberts *et al.*, 1998; Carroll *et al.*, 1992).

After the oligosaccharide core is transferred to the protein, additional modifications take place through the action of glycosyltransferases as well as through the removal of certain glycosyl residues. These modifications occur as the protein migrates through the Golgi apparatus to the cell surface. Enzymes that are destined for the lysosomes are directed there by a specific carbohydrate modification.

Some glycoproteins are tethered to the membrane by a lipid linkage. The protein is attached to the carbohydrate through phosphatidylethanolamine linkage, and the carbohydrate is in turn attached to the membrane *via* linkage to phosphatidylinositol, which anchors the structure within the membrane. The linkage is called a glycosylphosphatidylinositol anchor, and proteins that are anchored in this way are referred to as glypiated proteins (Alberts *et al.*, 1998).

1.5.10

Conclusions

This book is devoted to the analysis of gene expression at the level of mRNA and protein. However, this chapter is a cautionary warning against a simplistic reductionist view of biological function based solely on the amount of mRNA or protein present. Moreover, this subchapter can provide only briefest introductions to the topic of post-translational modifications of proteins. For example, a wide variety of further modifications have been described such as hydroxylation at proline and lysine residues of collagens, carboxy terminal amidation of peptide hormones (*e.g.*, oxytocin and vasopressin), carboxylation of glutamine residues within several proteins of the blood clotting cascade and also deformylation, acetylation, attachment of prosthetic groups, addition of metal ions and co-factors, subunit multimerisation, *etc.* However, a detailed description of all of these would exceed the scope of this book and the interested reader is referred to the appropriate specialist literature (*e.g.*, Aletta *et al.*, 1998; Brodbeck & Bordier, 1988; Chapman-Smith & Cronan, 1999; Comer & Hart, 1999; Freedman & Hawkins, 1980; Grant & Berger, 1999; Graves *et al.*, 1994; Helenius & Aebi, 2001; Nalivaeva & Turner, 2001; Parekh & Rohlff, 1997; Stallcup, 2001; Tuboi *et al.*, 1992; Vosseller *et al.*, 2001).

1.6

Correlation of mRNA and protein expression

1.6.1

Introduction

The classical biochemical approach to the study of biological processes has up to now been based on the purification to homogeneity of specific proteins by following their biological activity through extensive steps of fractionation. Each purified protein was then subjected to detailed structural, functional and regulatory analyses and the final step was the attempt to reconstitute the biological process *in vitro* from its isolated components. This type of research was conducted in a hypothesis-driven manner in which questions such as which protein causes a given biological activity or effect was frequently central to a research project. In this environment, individual laboratories championed individual proteins or biochemical pathways and studied them with great success as isolated modules. While this approach has been very successful in some instances, particularly in the study of core metabolic pathways, it is fundamentally quite limited in detecting and analysing the complex networks of functional and regulatory interactions that are common to cellular function. The availability of complete genomic sequences of numerous species and therefore, in principle, the knowledge of every gene and protein that constitutes a biological system, has catalysed an idea that biological processes could also be analysed systematically by concurrently analysing all the components that constitute the process. This has been referred to as discovery science (Aebersold *et al.*, 2001). In addition to complete genomic databases, discovery science also requires high-throughput technologies for the (quantitative) analysis of large numbers of genes and proteins. Such technology has been emerging over the past few years, first, for the systematic analysis of gene expression at the mRNA level and more recently for the analysis of proteins.

Studies measuring gene expression at the mRNA level and changes thereof have been most commonly carried out using any one of a variety of nucleotide array technologies (DeRisi *et al.*, 1997; Lashkari *et al.*, 1997; Lockhart & Winzeler, 2000; Schena *et al.*, 1995), or alternative methods for mRNA analysis such as serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995), differential display RT-PCR (DDRT-PCR) (Liang & Pardee, 1992), cDNA fingerprinting (Ivanova & Belyavsky, 1995) and several others more (see chapters 3 and 4). These technologies have been used successfully to diagnostically distinguish between cell types (Alizadeh *et al.*, 2000; Golub *et al.*, 1999; Lockhart & Winzeler, 2000; Ross *et al.*, 2000; Scherf *et al.*, 2000; Zhang *et al.*, 1997) and to differentiate between states (metabolic, activation, pathological) of a particular cell type (Lashkari *et al.*, 1997; Marton *et al.*, 1998), as well as the comprehensive analysis of cellular pathways and processes by targeted perturbations of cells (Holstege *et al.*, 1998; Ideker *et al.*, 2001; Marton *et al.*, 1998; Roberts *et al.*, 2000). In comparison to systematic gene expression measurements at the mRNA level, the measurement of protein expression profiles has been slower, more difficult and more expensive. It might

therefore be tempting to try to infer proteomic properties such as protein quantity, modification, activity, linkage with other proteins and structure from genomic data. In this part of the book, a number of studies are summarised, mostly from yeast biology, to examine the question to what extent absolute protein abundance and perturbation-induced changes of protein abundance can be predicted from the corresponding measurements at the mRNA level. This is a critical question for the new systematic biology because it asks whether one type of global measurement by itself is likely to be sufficient for the study of the structure and function of biological processes or whether the concurrent measurement of diverse properties of the system will be required.

1.6.2

Levels of mRNA and protein expression: correlations and discrepancies

First, the correlation of the absolute number of mRNA transcripts and proteins has been examined. A number of studies on this topic have been published over the last few years. The research group around Ruedi Aebersold has compared the absolute mRNA transcript abundances measured by SAGE analysis (Velculescu *et al.* 1995) with the abundance of the corresponding proteins in exponentially growing cells of *Saccharomyces cerevisiae* (Gygi *et al.*, 1999). While a loose correlation between the two measurements was observed, the study conclusively showed that neither the protein abundance nor the mRNA level could be inferred from the data obtained from the alternate type of molecule. Similar conclusions were reached in other studies in yeast (Futcher, 1999) and other species and in different cell types (Anderson & Seilhamer, 1999; Coghlan & Wolfe, 2000). This is not surprising since the protein abundance is also controlled translationally and post-translationally (McCarthy, 1998).

It is conceivable that perturbation-induced responses induced at the mRNA and protein levels could be tightly correlated even if the steady state levels of the respective molecules are disparate. For the study of regulated biological systems, the question whether induced changes in protein abundance and ultimately in protein function and activity can be inferred from mRNA profile measurements therefore appears more relevant. There are numerous anecdotal cases where the magnitude of induction of a specific gene at the protein and mRNA levels differ significantly. Recently, as part of a broader study and building on the availability of new mass spectrometric methods (Aebersold & Goodlett, 2001), a comparison of changes in protein and mRNA expression profiles induced in response to activating the galactose utilisation pathway in yeast was published (Ideker *et al.*, 2001). The study used global techniques including cDNA microarrays (see chapter 4) and the new isotope-coded affinity tag (ICAT™) methodology (see chapter 5) to determine expression of mRNA and protein separately and then compared the two for correlation (see figure 1.16). As defined by classical hypothesis-driven biochemistry, galactose utilisation is represented as a pathway to convert galactose into glucose-6-phosphate and a regulatory mechanism for on/off control. The process involves at least three types of proteins: (i) a transporter gene (*GAL2*) encoding a

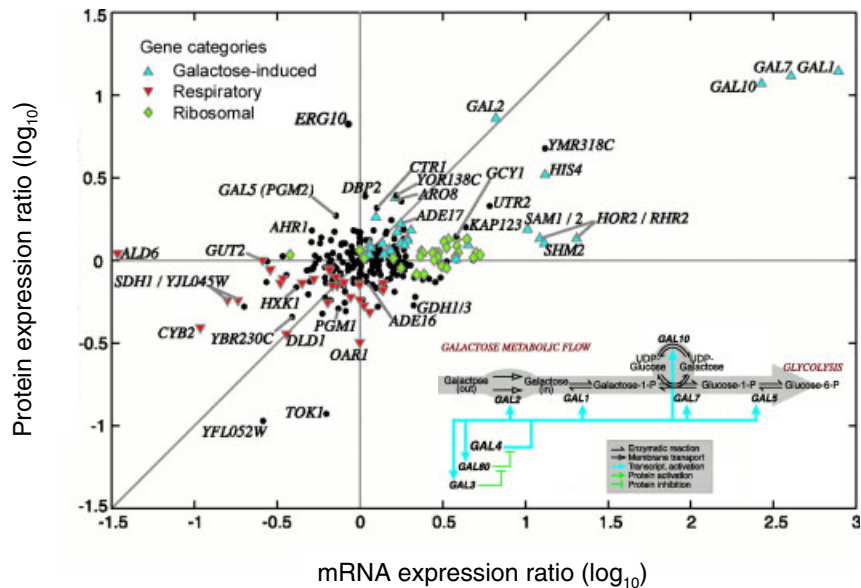


Figure 1.16 Scatter plot of protein versus mRNA expression ratios. Ratios of wild-type_{+gal} to wild-type_{-gal} protein expression (using isotope-coded affinity tag, ICAT™, technology) plotted against corresponding mRNA expression ratios (derived from cDNA microarray experiments). Note that many genes with elevated mRNA or protein expression in wild-type_{+gal} were metabolic or ribosomal, but respiratory genes were decreased. Model of

galactose utilisation shown as an inset. Yeast metabolise galactose via a series of steps involving GAL2 transporter and enzymes produced by GAL1, GAL7, GAL10 and GAL5 while GAL4, GAL80 and GAL3 transcriptionally regulate the enzyme genes. The figure is kindly provided by Dr. Trey Ideker (The Institute for Systems Biology, Seattle, Washington, USA) and reprinted with permission from *Science* (Ideker *et al.*, 2001).

permease that transports galactose into the cell, (ii) a group of genes encoding enzymes required for conversion of intracellular galactose, including galactokinase (GAL1), uridylyltransferase (GAL7), epimerase (GAL10), and phosphoglucomutase (GAL5/PGM2), and (iii) three genes, GAL3 (encoding the protein GAL3p which associates with GAL80p when galactose is in the cell and thus releases the repression of GAL4p imposed by GAL80p binding to it; this results in transporter and enzymes being expressed at high levels), GAL4 (encoding the protein GAL4p, a DNA-binding factor), and GAL80 (encoding the protein GAL80p which binds GAL4p and inhibits its activity), exerting tight transcriptional control over the transporter, the enzymes, and to a certain extent, each other. In the absence of galactose, GAL80p binds GAL4p, a DNA-binding factor that can strongly activate transcription, and inhibits its activity. However, intracellular galactose causes GAL3p to associate with GAL80p and this association in turn causes GAL80p to release its repression of GAL4p resulting in the transporter and the enzymes being expressed at a high level.

In the study, wild-type and nine genetically-altered yeast strains were used, each with a complete deletion of one of the nine *GAL* genes grown in the presence (+*gal*) or absence (–*gal*) of two percent galactose and two percent raffinose in both states. The question of correlation of protein and mRNA expression was addressed by examination of the global changes in mRNA resulting from the knock-out perturbations with DNA microarrays of approximately 6,200 nuclear yeast genes and global changes in protein abundance between wild-type_{+gal} and wild-type_{–gal} conditions using ICAT™ methodology (Han *et al.*, 2001; Goodlett & Yi, 2002). Using a statistical method based on maximum-likelihood estimation, 997 genes were identified whose mRNA levels differed significantly from reference under one or more perturbations. Correlative database searching of the yeast database with SEQUEST (Eng *et al.*, 1994) was used to identify the proteins from which specific peptides originated and a separate algorithm called Xpress™ (Han *et al.*, 2001) used to indicate relative abundances of the parent proteins from ICAT™-labelled peptides identified by SEQUEST. Comparison of protein expression was limited to wild-type_{+gal} and wild-type_{–gal} cultures because at that time protein expression experiments took considerably longer to conduct and interpret than the corresponding mRNA expression experiments. This last point has changed little in the intervening time since the publication of the above results, but technological advances will eventually allow the ICAT™-type experiments to be done at a rate comparable to the mRNA expression analysis (*i.e.*, differential isotopic tagging of peptides for quantification by mass spectrometry; Aebersold & Goodlett 2001; Goodlett *et al.*, 2001).

The study identified and quantified a total of 289 regulated yeast proteins. Given the complexity of the lysate, this number of protein identifications is at first glance notably low. However, it is a reasonable number of identifications considering the low amount of starting material (100 micrograms protein from wild-type_{+gal} and wild-type_{–gal} cultures) and the use of top-down data-dependent methods for ion selection (Goodlett & Yi, 2002). The data showed that as a whole, the protein abundance ratios were moderately correlated with their mRNA counterparts. While approximately 30 proteins displayed clear changes in abundance between the wild-type_{+gal} and wild-type_{–gal} conditions, mRNA levels for 15 of these did not change significantly in response to any perturbation, suggesting that these proteins may be regulated post-transcriptionally. In addition, many ribosomal protein genes increased three to fivefold at the level of mRNA but not in terms of protein abundance in response to galactose addition. This imbalance is perhaps related to the high energetic cost of ribosomal assembly, rapid degradation of the ribosomal subunits, or an extremely long time interval between ribosomal-subunit mRNA and protein synthesis (more than twelve to sixteen hours of overnight growth prior to harvest).

An additional limitation of this study was that the identified proteins were almost exclusively ‘blue collar’ (*i.e.* working class) *GAL* enzymes and the transporter protein, but none of the *GAL* regulatory proteins. However, if one assumes that the reasons for failure to observe these ‘white collar’ proteins are due to their presence at low copy numbers per cell, then again the common top-down data-dependent

protocols used for ion selection would seem to favour the high-abundance peptide signals. A number of solutions are available to solve this issue. The most obvious is fractionation of the peptides from the high-performance liquid chromatography (HPLC) separation onto a matrix-assisted laser desorption/ionisation (MALDI) plate rather than directly into the mass spectrometer. Once on the MALDI plate, the ICATTM-labelled peptides could be interrogated off-line in a mode where the survey scan for each MALDI spot reveals a set of ions available for tandem mass spectrometry (Griffin *et al.*, 2001; Medzihradszky *et al.*, 2000) and where only those changing in abundance need be subjected to tandem mass spectrometry. As a final note for the use of mass spectrometry and stable isotope labelling to obtain protein expression data, several caveats are worth mentioning. Just as cross-hybridisation can lead to misidentification of mRNA species on a DNA microarray, a similar situation may lead to ambiguous or misleading results where peptides are used to identify parent proteins because: (i) peptides with sequence homology in multiple parent proteins or to proteins that may have existed in biologically active but truncated forms will lead to ambiguous parent protein identification, and (ii) the parent protein, identified by a database search of a tandem mass spectrum of a peptide without a post-translational modification, might have existed in a post-translationally modified form of biological importance that was not selected for collision-induced dissociation because the peptide with the post-translational modification was present at sub-stoichiometric amounts (Katze *et al.*, 2000).

1.6.3

Conclusions

In summary, a key question in the quest to assign function to gene sequences and to integrate the functions of expressed genes and proteins into regulated systems and networks was examined. The question is whether key properties of biological systems can be inferred from the mature gene expression profiling technologies such as DNA microarrays or SAGE. It was found that neither the steady state levels in mRNA and protein abundance nor their perturbation-induced changes correlate strongly enough to allow prediction of the one value from the other. Given the presence of other mechanisms of regulation of protein properties, such as protein modifications and intermolecular protein interactions, it is likely that these properties and their dynamic change will also be difficult to predict from genomic information alone. Therefore, this argues for a need for the development of more comprehensive methodologies that characterise these modifications and interactions, and recent reports document significant progress towards these goals (Gavin *et al.*, 2002; Ho *et al.*, 2002; Oda *et al.*, 2001; Weckwerth *et al.*, 2000; Zhou *et al.*, 2001; Zhu & Snyder, 2001). In addition, it is evident that new approaches need to be developed in which traditional hypothesis-driven research is being combined with the practice of discovery science. This emerging combination of traditional and discovery science tools is becoming known as the 'systems approach to biology'. Recently, a proof-of-principle study in systems biology was carried out using galactose induction in yeast as a model system. The study provided new insights into the well-

studied galactose induction pathway and its connections to other cellular pathways by combining systematic and quantitative mRNA and protein profiling technologies, systematic perturbation of all known components of the core pathway and mathematical integration and modelling of the data. It is clear that advances in the comprehensive analysis of protein products, in conjunction with the already mature methods in global measurement of mRNA expression, will one day enable the characterisation of cellular circuitry on a system-wide scale.

1.7

Housekeeping genes, internal and external standards

1.7.1

What are housekeeping genes?

Many cellular processes concerning survival, growth and differentiation are reflected in altered patterns of gene expression. The ability to quantitatively measure the level of transcription of specific genes is central to research into gene function. Such research is a crucial element of understanding the causes and mechanisms of diseases and thus makes an important contribution to the discovery of potential therapeutic and diagnostic targets. For example, one important application is the quantification of the regulation and expression of drug resistance markers in tumour cells (Ramachandran & Melnick, 1999).

Quantitative assays for the analysis of mRNA expression widely use so-called 'housekeeping genes', genes whose expression is presumed to remain constant under changing cellular conditions. Therefore, a housekeeping gene acts as an internal standard which allows normalisation and thus relative quantification of the expression of other genes.

Several years ago, housekeeping genes were simply defined as those genes that are always expressed (Watson *et al.*, 1965). Today, the housekeeping genes are defined as those genes which are expressed in all tissues, are highly conserved throughout evolution, and usually serve basic metabolic functions in the cell so that they are responsible for the successful completion of the cell cycle (Foss *et al.*, 1994). The latter point explains why the synthesis of these molecules is thought to fluctuate much less than that of other cell constituents.

Housekeeping genes are commonly used in semi-quantitative and quantitative RNA analyses such as Northern blot analysis, *in situ* hybridisation, ribonuclease protection assay, reverse transcription-based PCR (RT-PCR), differential display RT-PCR, and DNA microarrays. The appropriate choice of an internal standard is a very important element of quantitative RNA analyses.

Ideally, internal RNA controls should show constant expression across a wide range of physiological and experimental conditions. This is a prerequisite for valid quantitative comparisons of gene expressions among different tissue types, varying developmental stages, and experimentally treated cells (compare figures 1.17 and 1.18.A). However, in practise, the commonly used housekeeping

genes do not always manifest stable expression levels under all experimental conditions (Thellin *et al.*, 1999) (compare figures 1.17 and 1.18.B).

In fact, the expression of several widely used internal standards often varies in response to various factors so that they may not be useful for many routine applications (Zhong & Simons, 1999). This may partly be explained by the fact that housekeeping proteins are not only implicated in basal cell metabolism but also participate in other functions (Petersen *et al.*, 1990; Singh & Green, 1993; Ishitani *et al.*, 1996). In addition, a consequence of the conservation of the housekeeping genes is the frequent presence in mammalian genomes of corresponding intron-less pseudogenes that mimic cDNAs in size and sequence (homology more than 90 percent) and may cause quantification errors (Hanauer & Mandel, 1984). Such pseudogenes exist for GAPDH (Foss *et al.*, 1994) and β -actin (Finke *et al.*, 1993), for example. To circumvent this problem, the PCR-primers used to amplify the internal standard must be designed in such a way that amplification of pseudogenes can be excluded.

Another problem is contamination of RNA samples with genomic DNA, which can produce false-positive or falsely raised levels of gene expression (Kreuzer *et al.*, 1999). This effect can be reduced by prior DNase digestion, although this may lead to loss of material.

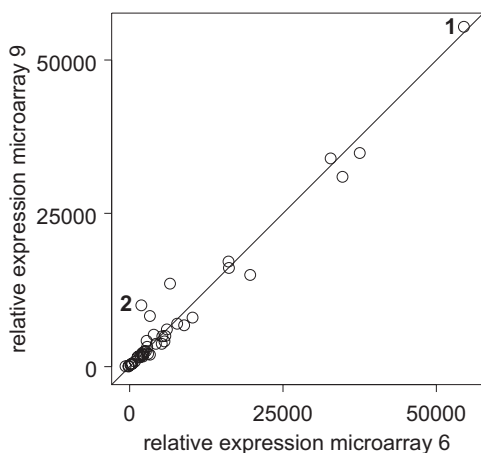


Figure 1.17 Comparison of the relative expression levels of 47 supposed housekeeping genes between two DNA microarrays (see in addition figure 1.18: DNA microarrays numbers 6 and 9). Gene expression was measured in the human cell line THP-1 under nine different experimental conditions (treatment with different drugs and with different incubation times). If the genes are not differentially regulated under the experimental conditions the relative expression levels should lie on the straight line

showing a nearly identical signal intensity in both experiments. This is the case for gene 1 representing an appropriate housekeeping gene (compare figure 1.18.A). In contrast, gene number 2 is up-regulated on DNA microarray number 9 in comparison to DNA microarray number 6 and is therefore not useful as an internal control (compare figure 1.18.B).

Figure and data were kindly provided by Dr. Reinhard Voss (Institut für Arterioskleroseforschung, University of Münster, Germany).

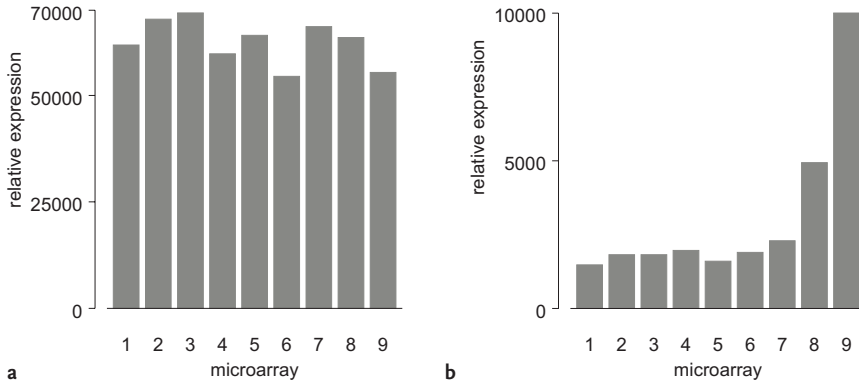


Figure 1.18 Measurement of the relative expressions by nine different DNA microarrays of **A)** a housekeeping gene that is not regulated under changing experimental conditions (gene number 1 in figure 1.17; mRNA for elongation factor EF-1 α) and **B)** a differentially regulated supposed housekeeping gene (gene number 2 in figure 1.17; mRNA for interferon-inducible γ 2 protein) which is not suitable as an internal control because a large difference in its mRNA expression level was seen between the different experimental treatments. Figures and data were kindly provided by Dr. Reinhard Voss (Institut für Arterioskleroseforschung, University of Münster, Germany).

1.7.2

Survey of the most important housekeeping genes

1.7.2.1 Glyceraldehyde-3-phosphate dehydrogenase

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH or G3PDH) is a key enzyme in glycolysis that plays an important role in energy metabolism. GAPDH catalyses the conversion of glyceraldehyde-3-phosphate to 1,3-bisphosphoglycerate with generation of the high-energy electron carrier NADH from NAD⁺ (nicotinamide adenine dinucleotide) (Tatton *et al.*, 2000).

The RNA encoding GAPDH is a ubiquitously expressed, moderately abundant message. It is frequently used as an endogenous control for quantitative RNA analyses based on the assumption that its expression remains constant under changing cellular conditions (Edwards & Denhardt, 1985).

However, in recent years, it has become clear that use of GAPDH as an internal standard may be inappropriate because GAPDH is now known to contribute to diverse cellular functions unrelated to glycolysis such as nuclear RNA export, DNA replication, DNA repair, exocytotic membrane fusion, cytoskeletal organisation and phosphotransferase activity (Sirover, 1999).

Furthermore, the above mentioned occurrence of pseudogenes causes problems when GAPDH is used as an internal standard for measuring mRNA expression (Foss *et al.*, 1994).

There are other cases in which GAPDH is an inappropriate standard. For example, GAPDH concentrations vary significantly between different individuals (Bus-

tin *et al.*, 1999), during pregnancy (Cale *et al.*, 1997), during different developmental stages (Puissant *et al.*, 1994), and during the cell cycle (Mansur *et al.*, 1993). Expression levels of GAPDH varied by eight-fold among seven different human adult tissues, with highest expression in brain, heart, and kidney and lowest expression in the pancreas (Warrington *et al.*, 2000). Insulin (Rolland *et al.*, 1995), vitamin D (Desprez *et al.*, 1992), oxidative stress (Ito *et al.*, 1996), and hypoxia (Zhong & Simons, 1999) stimulate GAPDH transcription. Furthermore, GAPDH is up-regulated in cancer, including human prostate carcinoma (Ripple & Wilding, 1995). Finally, GAPDH has been implicated in apoptosis, neurodegenerative disease and viral pathogenesis (Tatton *et al.*, 2000).

In light of this evidence, it is necessary to test the suitability of GAPDH as an internal control under each experimental condition prior to its use as an internal standard for any quantitative analysis.

1.7.2.2 β -Actin

β -Actin is a protein that is essential for the structure and kinetics of the cytoskeleton. Its mRNA is expressed at moderately abundant levels in most cell types (Bustin, 2000). β -Actin was one of the first RNAs to be used as an internal standard and it still serves as a quantitative reference.

However, as with GAPDH, the expression levels of β -actin have been reported to vary in some tissues and cell lines and in response to experimental manipulation, for example, under hypoxia (Zhong & Simons, 1999) and in blastomeres (Krussel *et al.*, 1998). In the above mentioned study by Warrington *et al.* (2000), β -actin was found to vary by seven-fold among seven different human adult tissues, with highest expression in brain and uterus and lowest expression in heart. In addition, the presence of pseudogenes interferes with the interpretation of results (Finke *et al.*, 1993).

All in all, the use of β -actin as an internal standard for quantification of gene expression is as problematic as that of GAPDH. The same problems occur when γ -actins are used in place of β -actin.

1.7.3

Other commonly used housekeeping genes

Besides GAPDH and β - and γ -actins, other mRNAs have been used occasionally as internal standards – for example, albumin which serves as an intravascular osmotic pressure contributor, hypoxanthine phosphoribosyltransferase, α -tubulin, β -tubulin which are essential for the structure and kinetics of the cytoskeleton, cyclophilin which is involved in cellular protein folding and protein interactions, and L32 which represents a ribosomal protein (Thellin *et al.*, 1999).

Unfortunately, quantification of RNA using these genes may also be problematic. To name just two critical applications, cyclophilin displays hypoxia-induced changes in expression levels (Zhong & Simons, 1999) while the level of L32 transcripts changes in the presence of mitogens (Thellin *et al.*, 1999).

On the other hand, there are special experimental conditions where expression levels are indeed constant. One example is hypoxanthine phosphoribosyltransferase which represents the control system of choice for the evaluation of RNA in paraffin-embedded tissues (Foss *et al.*, 1994) and which does not present pseudogenes.

Therefore, it is necessary to characterise the suitability of various housekeeping genes to serve as internal RNA controls under the particular conditions of each experiment.

1.7.3.1 Ribosomal RNA (rRNA)

Ribosomal RNAs (rRNA), which constitute 85 to 90 percent of total cellular RNA, have also been used as internal standards (Bustin, 2000). The levels of rRNA are thought less likely to vary under conditions that affect the expression of mRNAs (Barbu & Dautry, 1989) since rRNAs are transcribed by a distinct RNA polymerase (Paule & White, 2000).

The stable expression levels of rRNA relative to other housekeeping genes under several experimental conditions has been shown in some breast cell lines (Spanakis, 1993), in rat liver (de Leeuw *et al.*, 1989), in human fibroblasts (Mansur *et al.*, 1993) and in human (Zhong & Simons, 1999) and mouse (Bhatia *et al.*, 1994) malignant cell lines.

There are, however, two drawbacks to the use of rRNA: firstly, it cannot be used for normalisation when quantifying targets that have been enriched for mRNA because it is lost during mRNA purification as it has no poly(A) tail, and secondly, rRNA is expressed at much greater levels than any target mRNA.

Nevertheless, if the other commonly used housekeeping genes fail in certain experimental conditions the best way is to refer to 28S and 18S rRNA as internal standards for mRNA quantification (Thellin *et al.*, 1999).

1.7.4

New identified 'maintenance genes'

The development of massively parallel analyses of gene expression using either DNA microarrays or methods such as serial analysis of gene expression (SAGE) means that it is now possible to simultaneously measure the expression of thousands of genes that are expressed at low, moderate, and high abundance levels. Using DNA microarrays, measurement of the expression levels of about 7,000 genes in eleven different human adult and fetal tissues identified 535 genes that are expressed in each of the eleven tissues (Warrington *et al.*, 2000). These genes are turned on early in fetal development and stay on throughout adulthood and are therefore good potential candidates for being cellular 'maintenance genes'. Moreover, 47 of the transcripts were detected at similar levels in each of these tissues. For example, the transcript for elongation factor EF-1 was detected at high abundance and the transcript for E2 ubiquitin was detected at low abundance in all tissues surveyed (table 1.2). It is of note that the classical housekeeping genes, as for instance GAPDH and β -actin, do not occur in this list.

Table 1.2 47 transcripts expressed at the same level in eleven different human adult and fetal tissues (according to Warrington *et al.*, 2000).

Accession no.	Description	Bin[*]
AJ000099	Lysosomal hyaluronidase 2 (hyaluronoglucosaminidase)	M
D14662	Anti-oxidant protein 2 (non-selenium glutathione peroxidase; acidic calcium-independent phospholipase A2)	LM
D21261	Transgelin 2	M
D23673	KIAA0864	M
D29012	Proteasome, subunit Y (macropain, subunit δ)	M
D32129	Major histocompatibility complex, class I, A (HLA-A26) heavy chain	M
D50063	Proteasome, subunit p40 (Mov34 protein; macropain 26S subunit)	M
D63476	PAK-interacting exchange factor β	LM
J04617	Elongation factor EF-1 α	H
L06499	Ribosomal protein L37a	H
L37368	RNA-binding protein S1	M
L40395	Eukaryotic translation initiation factor 2B, subunit 2 β	LM
M15990	v-src-1 (Yamaguchi sarcoma viral oncogene homolog 1)	L
M23254	Ca ²⁺ -activated neutral protease, large subunit (calpain 2)	M
U06155	60S ribosomal protein L23A	H
U07857	Signal recognition particle 14kD (Alu RNA-binding protein)	M
U10323	Interleukin enhancer-binding factor 2 (nuclear factor NF45)	M
U20285	G protein pathway suppressor 1	M
U33821	Tax1-binding protein 1	M
U39317	Ubiquitin conjugating enzyme E2D 2	L
U46570	Tetratricopeptide repeat protein	M
U50079	Histone deacetylase 1	LM
U56417	1-acylglycerol-3-phosphate O-acyltransferase 1 (lysophosphatidic acid acyltransferase α)	M
U57341	Neurofilament triplet L protein	MH
U57342	Myelodysplasia/myeloid leukemia factor 2	M
U62317	Mitogen-activated protein kinase 8 interacting protein 2	M
U62739	Mitochondrial branched-chain amino acid aminotransferase 2	M
U79528	Sigma receptor (SR31747-binding protein 1)	M
U86070	Phosphomannomutase	M
X12432	Major histocompatibility complex, class I, A (HLA-A)	M
X16135	Heterogeneous nuclear ribonucleoprotein L	M
X51521	Villin 2 (ezrin)	M
X59892	Tryptophanyl-tRNA synthetase (IFNy-inducible γ 2 protein)	M
X61970	Proteasome subunit, α -type, 5 (prosome; macropain, subunit ζ)	LM
X72964	Centrin 2 (EF-hand protein; caltractin)	M
X74008	Protein phosphatase 1, catalytic subunit, isoform γ	LM
X76013	Glutaminyl-tRNA synthetase	M
X80199	MLN51 protein	M
X80763	5-hydroxytryptamine (serotonin) receptor 2C	LM
X83218	ATP synthase, H ⁺ transporting, mitochondrial F1 complex, subunit O	M
X91247	Thioredoxin reductase 1	LM
X96484	DiGeorge syndrome critical region gene 6	LM
Z11584	Nuclear mitotic apparatus protein 1, alternative splice form 2	M
Z12962	Ribosomal protein L41	H
Z27113	RNA polymerase II (DNA-directed), subunit F	M
Z47727	RNA polymerase II (DNA-directed), subunit K	LM
Z70759	Mitochondrial 16S ribosomal RNA	H

* The abundance levels in copies per cell: L ≤ 5 , LM $>5 \leq 10$, M $>10 \leq 50$, MH $>50 \leq 100$, H >100 .

The identification of these ‘maintenance genes’ is of great importance because they may well serve as useful quantitative internal controls in studies of normal adult and fetal gene expression in the future.

1.7.5

Methods of quantification

Errors in the quantification of mRNA transcripts are easily compounded by any variation in the amount of starting material between samples. This is especially relevant when the samples have been obtained from different individuals, and will result in the misinterpretation of the expression profiles of the target genes. Therefore, it is necessary to refer to one of the following evaluation methods.

1.7.5.1 Normalisation (*via* internal standards)

One accepted method for minimising errors in the quantification of mRNA transcripts and correction for sample-to-sample variation is to amplify, simultaneously with the target, a cellular RNA that serves as an internal reference against which other RNA values can be normalised (Karge *et al.*, 1998). This can be conducted either in one tube or in two tubes as a duplex-PCR. The housekeeping genes which are described in this chapter are most commonly used to normalise patterns of gene expression so far.

As has been already mentioned, the ideal internal standard or housekeeping gene should be expressed at a constant level among different tissues, at all stages of development, and in particular should be unaffected by the experimental treatment. In addition, an endogenous control should also be expressed at roughly the same level as the RNA under study. Therefore, the efficiency of PCR amplification of both genes has to be checked.

1.7.5.2 Standard curves (*via* external standards)

Today, quantification of mRNA is based on RT-PCR, hybridisations and sequence-based techniques mostly using fluorescence signals.

In real-time RT-PCR (compare chapter 3), fluorescence values are recorded during every PCR cycle and represent the amount of product amplified up to that point in the amplification reaction. The more template present at the beginning of the reaction, the fewer number of cycles it takes to reach a point in which the fluorescent signal is first recorded as being statistically significant (*i.e.* usually ten times the standard deviation of the baseline) above background (Gibson *et al.*, 1996). This point is defined as the threshold cycle (C_t), and will always occur during the exponential phase of amplification. For each sample a C_t value is measured which can be translated into a quantitative result by constructing a standard curve.

Using these techniques, it is possible to precisely determine the gene copy number per cell, per total RNA concentration, or per unit mass of tissue. This requires

the construction of an absolute standard curve for each individual amplicon to ensure accurate reverse transcription and PCR amplification profiles (Bustin, 2000).

A standard curve is prepared by measuring the C_t values of serial dilutions of known amounts of an individual amplicon and by plotting the C_t values against the logarithm of the initial copy numbers because the C_t values are inversely proportional to the log of the initial copy number (Higuchi *et al.*, 1993). The copy numbers of experimental RNAs can be calculated after real-time amplification from the linear regression of the standard curve. Normally, copy numbers are related to total RNA concentrations.

A standard curve can be constructed in two different ways (Bustin, 2000). One method is to subclone the amplicon behind a T7 or SP6 polymerase promoter into a plasmid vector and to generate an *in vitro*-transcribed sense RNA transcript which must be quantified and diluted. Afterwards, the C_t values are measured and a standard curve can be constructed. The second method makes use of single-stranded sense-strand oligodeoxynucleotides. This way of obtaining a standard curve is simpler but more cost-intensive and is an acceptable alternative to the T7 RNA polymerase-generated standard amplicon.

1.7.5.3 Alternative methods (*via* internal standards)

One commonly used but time-consuming alternative for quantification of mRNA is the technique of competitive RT-PCR (see chapter 3). This makes use of an internal standard which shares the same primer recognition and internal sequences with the primary target, leading to competition for reagents and amplification with more or less the same efficiency. A series of PCR tubes containing the target are spiked with serial dilutions of known copy numbers of the internal standard. The greater the concentration of the internal standard, the more likely it is that the primers will bind and amplify it, rather than the target (Raeymaekers, 1999). The concentration of the target can be determined by comparing the amplification ratios. If the ratio is 1:1, the concentration of the target is identical to that of the internal standard.

A further alternative method is real-time quantitative RT-PCR using quantitative spiking. Duplicate sets of tubes containing fixed concentrations of target mRNA are co-amplified with serial dilutions of an internal control RNA. The target hybridisation probe is added to the first, the internal control hybridisation probe to the second set of tubes. C_t values are measured and plotted against the relative copy number of internal control in each reaction tube. The number of unknown target molecules can be determined from the intersection of the two lines (figure 1.19), where C_t of target equals C_t of internal control, after multiplying by the respective target dilution (Gibson *et al.*, 1996).

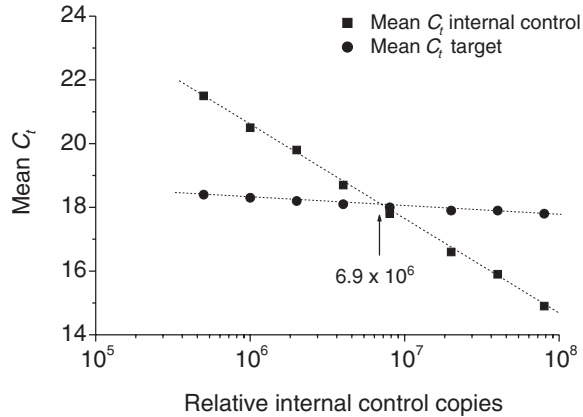


Figure 1.19 Real-time quantitative RT-PCR. Two sets of tubes containing eight serial dilutions of a known amount of internal control RNA are prepared. After adding the target hybridisation probe to the first and the internal control hybridisation probe to the second set of tubes, the C_t values are measured. The mean C_t values of internal control and target are plotted against the known internal control copy number. The number of unknown target molecules can be determined from the intersection of the two lines, i.e. the point where C_t of target equals C_t of internal control, and multiplication by the respective target dilution.

1.7.6

Summary

The use of housekeeping genes as internal controls should be examined carefully in relation to the cell types and the cell metabolism, otherwise false quantification results may be obtained. Good knowledge of the modulations of the internal standard is important prior to beginning any data interpretation (Thellin *et al.*, 1999). Dent *et al.* (1997) have proposed using at least two types of housekeeping gene transcripts as internal standards at the same time. If the ratios of both housekeeping genes do not significantly vary between different experimental conditions they may be used as internal standards without any problem. The use of only one internal standard is nevertheless acceptable providing it has been shown that the expression of the housekeeping gene is not influenced by the experimental conditions under investigation (Thellin *et al.*, 1999).

In summary, the appropriate choice of an internal or external standard is critical for absolute quantification of target nucleic acids. Therefore, it is very urgent to search for a relevant and universal means of comparing and normalising between samples.

1.8

Classification of differential gene expression technologies

1.8.1

Introduction

In the past 30 years, since the development of the classic approaches Northern blotting and nuclear run-on assay for RNA expression analysis (Alwine *et al.*, 1977; Eikhom *et al.*, 1975; Marzluff *et al.*, 1973; Marzluff, 1978) and electroblotting and Western blotting for the analysis of protein expression (Burnette, 1981; Towbin *et al.*, 1979), the analysis of gene expression has become one of the most widely used set of techniques in biomedical research. The significance of these techniques is reflected in the thousands of scientific papers that cite the original descriptions.

1.8.2

From single genes to transcriptomes

Since the early 1990s, differential gene expression technologies have been applied to many biological problems. However, the scientist's view of analysing gene expression has changed dramatically. Where once the analysis of differential gene expression was viewed in terms of a single gene on Northern or Western blots, it is now possible to assess the expression of nearly all genes of a genome simultaneously. This genome-wide expression analysis was facilitated by the development of several methods that are described in the following chapters of this book.

The genome-wide landscape of gene expression is generally termed the 'transcript profile' or the 'expression profile' of an organism (Green *et al.*, 2001). This profile is a complex, context-dependent pattern of expressed genes, that is as typical for a certain cell or tissue in a defined biological state (*e.g.*, developmental stage, differentiation stage, healthy or cancer cells) as it is a fingerprint of an individual finger tip. Therefore, these profiles are often referred as the RNA or protein fingerprints of a cell or tissue.

Expression analysis at the level of a single gene at a single time point is a relatively simple task. The message of a gene in the experimental situation is either up-regulated, down-regulated, or unchanged compared to its expression in the control situation. By contrast, the analysis of the expression of a multitude of genes or, ideally, the totality of genes in response to a stimulus is an immensely complex and difficult endeavour (see figure 1.20).

Technologies providing comprehensive expression profiles of a given genome are at the core of a new scientific field called 'functional genomics'. The aim of this field is the precise understanding of global changes in gene expression patterns in well-defined experimental contexts (Brent, 2000; Green *et al.*, 2001).

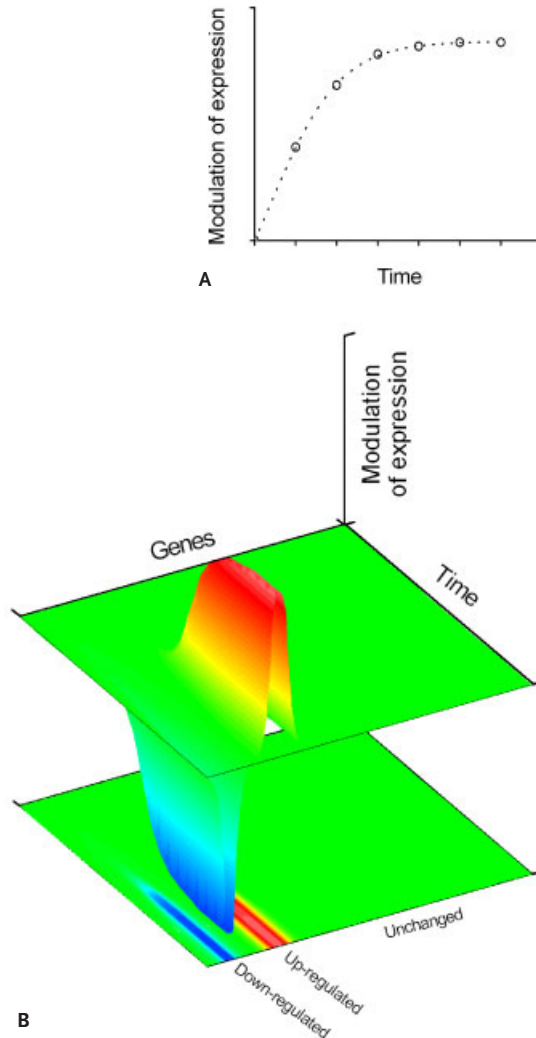


Figure 1.20 Graphical representation of the relative complexity of single gene and genome-wide assessment of expression analysis. The increase in relative scale and complexity of data management becomes quite obvious by comparing the hypothetical schematics of the types of data generated by **A**) Northern blot analysis or quantitative RT-PCR, with a handful of data points, to the data output from **B**) a genome-wide expression analysis, *e.g.*, generated by serial analysis of gene expression (SAGE) or DNA microarrays, with several thousands of data points. Assuming that there are 30,000 to 40,000 genes represented in the transcriptome of any particular cell, as it is estimated for a typical human cell, and, for example, a reasonable time course with at least five time points, one might expect the problems to handle the

order of more than 150,000 data points. In this figure, the coloured surface represents the relative change in the expression of a certain gene at a given time point compared to the initial state (time point zero). The expression of genes that are up-regulated (up to a few hundred) are coloured in yellow to red (with red as the highest change in expression) whereas down-regulated genes (also up to a few hundred) are coloured in turquoise to blue (with dark blue as the greatest change in expression). The plane coloured in green shows cohorts of genes that are not modulated in their expression in the given experiment. These genes represent the majority of all measured genes (about several hundreds to several thousands). The figure is adapted from Green *et al.* (2001).

1.8.3

Methods of classification

State-of-the-art methods in functional analysis of gene expression are broadly divided into the methods to analyse RNA expression and methods to analyse protein expression (figure 1.21). That can be divided further into two sub-families based on the following background. Depending on their abilities, methods for the analysis of gene expression are termed ‘closed architecture’ or ‘open architecture’ systems. While closed systems, such as DNA arrays, only allow analysis of pre-defined known genes, open systems, such as PCR-based differential display methods, allow analysis of the expression of genes that were previously unknown. The quality of closed systems therefore depends on the data known before. A complete analysis of the transcriptome of a genome using closed systems is strictly dependent upon the completeness of knowledge of that genome, thus severely limiting comprehensive applicability in all but the most well-characterised and often-used species for the near future. DNA microarrays (Marshall & Hodgson, 1998) and quantitative RT-PCR methods, such as competitive RT-PCR (Freeman *et al.*, 1999) and

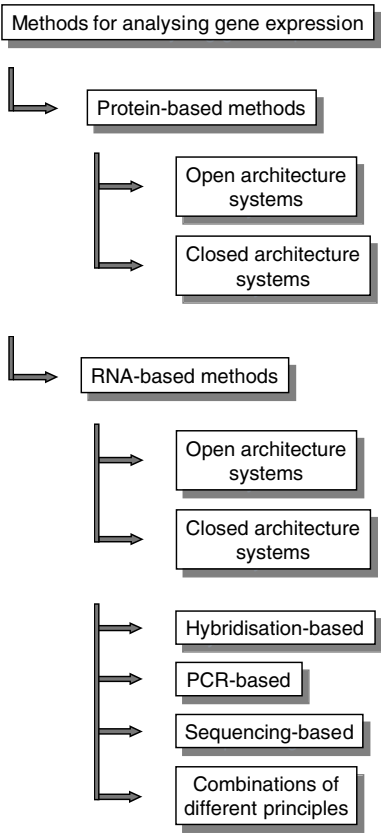


Figure 1.21 Classifications of technologies used for the analysis of gene expression. See text for details.

real-time RT-PCR (Heid *et al.*, 1996) are some of the most commonly used closed system methods for the analysis of gene expression.

Nevertheless, closed systems work well for certain types of analyses, as it is demonstrated by several studies using conventional techniques such as Northern blotting or quantitative RT-PCR. The use of closed systems is restricted in functional genomics because the topology of genes in this application is not well defined and novelty is at a premium. Even in species where the whole genome is known, the transcriptome is thought to be far more complex than the number of genes would at first suggest because of alternatively splicing, gene polymorphisms, RNA editing, *etc.* Therefore, one must keep in mind that even closed systems with huge capacity, such as DNA microarrays, are limited to our knowledge of the transcriptome. On the other hand, open systems may also be unable to exhaustively cover the whole transcriptome (see comments below).

In contrast to closed architecture systems, open systems are defined by the fact that no *a priori* knowledge of the transcriptome is necessary. Therefore, such systems can identify transcripts that are novel and no prior information is required on the modulation of expression of such a new transcript. At least in theory, it is possible using open architecture systems to identify all alternatively spliced, edited or otherwise modified RNAs, so that the complexity of a given transcriptome does not affect the quality of expression analysis.

Although open systems are not dependent upon existing information about the investigated transcriptome, they normally require the support of available data. For example, in the simplest case, a comparison of a transcriptome investigated with an open architecture approach using pre-existing data on the transcriptome reveals novel and so far unknown transcripts or edited transcripts by filtering out known genes. Comparison of the expression modulation of a novel transcript with data of known genes may also facilitate the initial characterisation and may give a first idea of the function of the novel transcript. The fact that the modulation of expression of such a gene shows a behaviour similar to others may be very informative if a known biological pathway or phenomenon is implicated. Nevertheless, we cannot absolutely ascertain the function of a novel transcript by virtue of simple association but this can be useful for the generation of an initial testable hypothesis to address the function of the gene.

In addition, the capacity of open systems is also of a limited nature. If knowledge of a transcriptome is comprehensive enough, then open systems cannot really compete with closed architecture systems. Even a method that theoretically permits analysis of a whole genome is limited by the fact, which is valid for both closed and open methods currently available, that the analysis of low abundance transcripts, rarely expressed splicing variants or other modified/edited RNAs is difficult and restricted by limited detection sensitivities. Most of these arguments are also valid for the analysis of protein expression.

In the field of protein expression analysis, open systems are of main interest because proteins are often highly post-translationally modified (*e.g.*, glycosylation, phosphorylation, ubiquitination, removal of leader sequences) and little knowledge exists on this with respect to the whole proteome. Therefore, open architecture sys-

tems are highly recommended in the field of protein expression analyses, because they allow the identification of nearly all gene products of interest including all their modifications.

Open systems have a voracious appetite for bioinformatic resources. Huge and complex computing capacity is required for data management and mining, statistical analysis, and for normalisation and calibration.

Contrary to opinion by some of their respective proponents, closed and open architecture systems are not competitive but complementary technologies. Once novelty has been identified, interesting transcripts should be investigated in detail in direct ways using closed architecture systems (Green *et al.*, 2001). Furthermore, these two systems have in common that they create lists of genes as a result, which requires a well-planned strategy for annotation and classification by functional role hierarchies, as it has been applied already to completed genomes (Rubin *et al.*, 2000).

In addition to the classification mentioned above, it is also possible to classify RNA-based methods for the analysis of differential gene expression depending on the procedure upon which the method is based. RNA-based methods may be divided into the following three groups: (i) hybridisation-based methods, (ii) PCR-based techniques, and (iii) sequencing-based approaches. Typical examples for hybridisation-based technologies are DNA microarrays, for PCR-based methods are reverse transcription-based PCR (RT-PCR) methods and for sequencing-based methods is serial analysis of gene expression (SAGE).

This latter classification is not entirely satisfactory, however. Many state-of-the-art systems combine different kinds of molecular biological methods. A PCR amplification step to increase signal intensity may be combined with a hybridisation procedure to simplify the detection of differentially expressed genes or to increase the number of genes of which differential expression can be detected simultaneously. Amplified differential gene expression (ADGE), for example, combines a PCR-based technique with DNA microarray-based hybridisation for detection of differentially expressed transcripts.

Open system transcript analysis provides the possibility to associate a novel modulated gene with the expression of several other genes that are regulated in similar fashion. This facilitates the generation of a testable hypothesis to investigate the functional role of this differentially expressed gene in the experimental system.

1.8.4

Conclusions

Irrespective of the kind of classification, the primary considerations for a scientist in selecting a technology for gene expression analysis are (i) the amount of starting material required, (ii) the sensitivity of the technique, (iii) coverage of the method, and (iv) additional follow-up costs which are often underestimated. The amount of biological starting material is a primary concern because it can limit the applicability of the method. In addition, it is critical to assess the sensitivity of the method. It

is estimated that 90 to 95 percent of all eukaryotic mRNA species are present at five or less copies per cell (Alberts *et al.*, 1998; Bonaldo *et al.*, 1996; Zhang *et al.*, 1997). However, these rare transcripts make up only 35 to 50 percent of the total cellular mRNA mass which is estimated to be only one to three percent of the amount of total RNA. The percentage of all possible mRNAs assayed by the selected technology (the coverage of the method) is also important for success. Closed systems, by definition, do not enable coverage of the entire transcriptome. At the present time, however, even open systems cannot guarantee comprehensive transcriptome coverage. Therefore, the choice of the system depends on the task to be undertaken. An aspect that is often insufficiently realised are the follow-up costs of a method. These may be substantial. Genome-wide expression analyses are in general time, labour and cost-intensive. Follow-up costs are high. In addition, the evaluation of an expression experiment requires many time-intensive investigations using additional technologies such as statistical analysis, management of huge amounts of data, and specific confirmatory follow-up investigations.

Classification of methods is therefore not an exhaustive or exact science. It is simply a tool to enable the researcher to assess rapidly the potential advantages and disadvantages of a method and to enable him or her to select the right approach for a particular expression experiment.

1.9

Summary

Gene expression is the result of a closely regulated and highly ordered process which involves the sequential assembly and disassembly of molecular nanomachines within the cell. While much has been learned about this process, the complete details are far from understood. For example, several recent reports point to the importance of small, non-coding RNA molecules (small temporally expressed RNAs, stRNAs, or microRNAs, miRNAs) being involved in the post-transcriptional fine tuning of gene expression (Banerjee & Slack, 2002; Eddy, 2001; Grosshans & Slack, 2002; Ruvkun, 2001). Nevertheless, sufficient information exists for the useful application of the powerful technologies that will be described in subsequent chapters.

1.10

References

Recommended readings

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D., *Molecular biology of the cell*. Garland Publishing, New York, 1998.

Lehninger, A. L., Nelson, D. L., and Cox, M. M., *Principles of biochemistry*, 3rd edition, Worth Publishing, New York, 2000.

Stryer, L., *Biochemistry*, 4th edition, W. H. Freeman and Company, New York, 1995.

Voet, D., and Voet, J., *Biochemistry*, 2nd edition, John Wiley & Sons, New York, 1995.

White, R., and White, R. J., *Gene transcription: mechanisms and control*, Blackwell Science, Williston, 2001.

Internet resources

Eukaryotic Promoter Database (EPD) of the Swiss Institute for Experimental Cancer Research (ISREC); Epalinges sur Lausanne, Switzerland; accessible at <http://www.epd.isb-sib.ch/>.

Alternative Splicing Database of the National Energy Research Scientific Computing Center (Lawrence Berkeley National Laboratory, Berkeley, California, USA); accessible at <http://cbcg.nersc.gov/asdb>.

Cited publications

Aebersold, R., Hood, L. E., Watts, J. D. Equipping scientists for the new biology. *Nature Biotechnology* 2000; 18(4):359.

Aebersold, R., and Goodlett, D. R. Mass spectrometry in proteomics. *Chemical Reviews* 2001; 101(2):269-295.

Agabian, N. Trans splicing of nuclear pre-mRNAs. *Cell* 1990; 61(7):1157-1160.

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., and Thanos, D. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell* 2000; 103(4):667-678.

Aggarwal, S. R., Lindros, K. O., and Palmer, T. N. Glucagon stimulates phosphorylation of different peptides in isolated periportal and perivenous hepatocytes. *FEBS Letters* 1995; 377(3):439-443.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. *Molecular biology of the cell*. Garland Publishing, New York, New York, USA 1998:528.

Aletta, J. M., Cimato, T. R., and Ettinger, M. J. Protein methylation: a signal event in post-translational modification. *Trends in Biochemical Sciences* 1998; 23(3):89-91.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; 403(6769):503-511.

Alwine, J. C., Kemp, D. J., and Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the USA* 1977; 74(12):5350-5354.

Ambros, V. MicroRNAs: tiny regulators with great potential. *Cell* 2001; 107(7):823-826.

Anderson, L., and Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 1997; 18(3-4):533-537.

Attie, A. D., Kastelein, J. P., and Hayden, M. R. Pivotal role of ABCA1 in reverse cholesterol transport influencing HDL levels and susceptibility to atherosclerosis. *Journal of Lipid Research* 2001; 42(11):1717-1726.

Banerjee, D., and Slack, F. Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays* 2002; 24(2):119-129.

Barbu, V., and Dautry, F. Northern blot normalization with a 28S rRNA oligonucleotide probe. *Nucleic Acids Research* 1989; 17(17):7115.

Bartel, P. L., Roedel, J. A., SenGupta, D., and Fields, S. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genetics* 1996; 12(1):72-77.

Bashirullah, A., Halsell, S. R., Cooperstock, R. L., Kloc, M., Karauskakis, A., Fisher, W. W., Fu, W., Hamilton, J. K., Etkin, L. D., and Lipshitz, H. D. Joint action of two RNA degradation pathways controls the timing of ma-

- ternal transcript elimination at the midblastula transition in *Drosophila melanogaster*. *EMBO Journal* **1999**; 18(9):2610-2620.
- Bastin, P., Galvani, A., and Sperling, L. Genetic interference in protozoa. *Research in Microbiology* **2001**; 152(2):123-129.
- Bell, A. C., and Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **2000**; 405(6785):482-485.
- Ben-Neriah, Y. Regulatory functions of ubiquitination in the immune system. *Nature Immunology* **2002**; 3(1):20-26.
- Bender, M. A., Reik, A., Close, J., Telling, A., Epner, E., Fiering, S., Hardison, R., and Groudine, M. Description and targeted deletion of 5' hypersensitive site 5 and 6 of the mouse beta-globin locus control region. *Blood* **1998**; 92(11):4394-4403.
- Benne, R. RNA editing in trypanosomes: is there a message? *Trends in Genetics* **1990**; 6(6):177-181.
- Benne, R., van den Burg, J., Brakenhoff, J. P., Sloof, P., van Boom, J. H., Tromp, M. C. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **1986**; 46(6):819-826.
- Bentley, D. L. Regulation of transcriptional elongation by RNA polymerase II. *Current Opinion in Genetics and Development* **1995**; 5(2):210-216.
- Bhatia, P., Taylor, W. R., Greenberg, A. H., and Wright, J. A. Comparison of glyceraldehyde-3-phosphate dehydrogenase and 28S-ribosomal RNA gene expression as RNA loading controls for northern blot analysis of cell lines of varying malignant potential. *Analytical Biochemistry* **1994**; 216(1):223-226.
- Bird, A. DNA methylation patterns and epigenetic memory. *Genes & Development* **2002**; 16(1):6-21.
- Blaxter, M., and Liu, L. Nematode spliced leaders-ubiquity, evolution and utility. *International Journal of Parasitology* **1996**; 26(10):1025-1033.
- Bonaldo, M. F., Lennon, G., and Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research* **1996**; 6(9):791-806.
- Bonen, L. Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB Journal* **1993**; 7(1):40-46.
- Bosher, J. M., and Labouesse, M. RNA interference: genetic wand and genetic watchdog. *Nature Cell Biology* **2000**; 2(2):E31-E36.
- Bouget, F. Y., Gerttula, S., Shaw, S. L., and Quatrano, R. S. Localization of actin mRNA during the establishment of cell polarity and early cell divisions in *Fucus* embryos. *Plant Cell* **1996**; 8(2):189-201.
- Brennicke, A., Marchfelder, A., and Binder, S. RNA editing. *FEMS Microbiology Reviews* **1999**; 23(3):297-316.
- Brent, R. Genomic biology. *Cell* **2000**; 100(1):169-183.
- Brivanlou, A. H., and Darnell, J. E., Jr. Signal transduction and the control of gene expression. *Science* **2002**; 295(5556):813-818.
- Brodbeck, U., and Bordier, C. Post-translational modification of proteins by lipids. A laboratory manual. Brodbeck, U., and Bordier, C., editors. Springer-Verlag, Berlin, Germany **1988**.
- Bulger, M., and Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes and Development* **1999**; 13(19):2465-2477.
- Burnette, W. N. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Analytical Biochemistry* **1981**; 112(2):195-203.
- Bustin, S. A., Gyselman, V. G., Williams, N. S., and Dorudi, S. Detection of cytokeratins 19/20 and guanylyl cyclase C in peripheral blood of colorectal cancer patients. *British Journal of Cancer* **1999**; 79(11-12):1813-1820.
- Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology* **2000**; 25(2):169-193.
- Cale, J. M., Millican, D. S., Itoh, H., Magness, R. R., and Bird, I. M. Pregnancy induces an increase in the expression of glyceraldehyde-3-phosphate dehydrogenase in uterine artery endothelial cells. *Journal of the Society for Gynecologic Investigation* **1997**; 4(6):284-292.
- Caponigro, G., and Parker, R. Mechanisms and control of mRNA turnover in *Saccharomyces cerevisiae*. *Microbiological Reviews* **1996**; 60(1):233-249.
- Carey, M. F. Transcriptional activation – A holistic view of the complex. *Current Biology* **1995**; 5(9):1003-1005.

- Carey, M. F., and Smale, S. T. Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, New York, USA 2000.
- Carles, C., and Riva, M. Yeast RNA polymerase I subunits and genes. In: Transcription of ribosomal RNA genes by eukaryotic RNA polymerase I. Paule, M. R., editor. Springer-Verlag, Berlin, Germany 1998:9-38.
- Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., and Versteeg, R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001; 291(5507):1289-1292.
- Carroll, K. K., Guthrie, N., and Ravi, K. Dolichol: function, metabolism, and accumulation in human tissues. *Biochemistry and Cell Biology* 1992; 70(6):382-384.
- Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* 2001; 12(2):323-37.
- Chan, L. RNA editing: exploring one mode with apolipoprotein B mRNA. *Bioessays* 1993; 15(1):33-41.
- Chapman-Smith, A., and Cronan, J. E., Jr. The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity. *Trends in Biochemical Sciences* 1999; 24(9):359-363.
- Chartrand, P., Singer, R. H., and Long, R. M. RNP localization and transport in yeast. *Annual Review of Cell and Developmental Biology* 2001; 17:297-310.
- Chédin, S., Ferri, L., Peyroche, G., Andrau, J. C., Jourdain, S., Lefevre, O., Werner, M., Carles, C., and Sentenac, A. The yeast RNA polymerase III transcription machinery. A paradigm for eukaryotic gene activation. *Cold Spring Harbor Symposia on Quantitative Biology* 1998; 63:381-389.
- Chee, M., Yang, R., Hubell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. Accessing genetic information with high-density DNA arrays. *Science* 1996; 274(5287):610-614.
- Chen, S. H., Habib, G., Yang, C. Y., Gu, Z. W., Lee, B. R., Weng, S. A., Silberman, S. R., Cai, S. J., Deslypere, J. P., Rosseneu, M., Gotto, A. M., Li, W. H., and Chan, L. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 1987; 238(4825):363-366.
- Chen, C. Y., and Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences* 1995; 20(11):465-470.
- Chen, H., Lin, R. J., Xie, W., Wilpitz, D., and Evans, R. M. Regulation of hormone-induced histone hyperacetylation and gene activation via acetylation of an acetylase. *Cell* 1999; 98(5):675-686.
- Chen, C. A., and Manning, D. R. Regulation of G proteins by covalent modification. *Oncogene* 2001; 20(13):1643-1652.
- Chen, W., Sun, Y., Welch, C., Gorelik, A., Leventhal, A. R., Tabas, I., and Tall, A. R. Preferential ATP-binding cassette transporter A1-mediated cholesterol efflux from late endosomes/lysosomes. *Journal of Biological Chemistry* 2001; 276(47):43564-43569.
- Cho, R. J., Campbell, M. J., Winzler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 1998; 2(1):65-73.
- Ciechanover, A. The ubiquitin-mediated proteolytic pathway. *Brain Pathology* 1993; 3(1):67-75.
- Ciechanover, A. The ubiquitin-mediated proteolytic pathway: mechanisms of action and cellular physiology. *Biological Chemistry Hoppe-Seyler* 1994; 375(9):565-581.
- Ciechanover A, Schwartz AL. The ubiquitin-mediated proteolytic pathway: mechanisms of recognition of the proteolytic substrate and involvement in the degradation of native cellular proteins. *FASEB Journal* 1994; 8(2):182-91.
- Coghlan, A., and Wolfe, K. H. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* 2000; 16(12):1131-1145.
- Cogoni, C., and Macino, G. Post-transcriptional gene silencing across kingdoms. *Current Opinion in Genetics and Development* 2000; 10(6):638-643.
- Colgan, D. F., and Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes and Development* 1997; 11(21):2755-2766.

- Comer, F. I., and Hart, G. W. O-GlcNAc and the control of gene expression. *Biochimica et Biophysica Acta* **1999**; 1473(1):161-171.
- Conaway, J. W., Shilatifard, A., Dvir, A., and Conaway, R. C. Control of elongation by RNA polymerase II. *Trends in Biochemical Sciences* **2000**; 25(8):375-380.
- Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R., and Kornberg, R. D. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **2000**; 288(5466):640-649.
- Cramer, P., Bushnell, D. A., and Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science* **2001**; 292(5523):1863-1876.
- Cullen, B. R. Mechanism of action of regulatory proteins encoded by complex retroviruses. *Microbiological Reviews* **1992**; 56(3):375-394.
- Dantanel, J. C., Wurtz, J. M., Poch, O., Moras, D., and Tora L. The TBP-like factor: an alternative transcription factor in metazoa? *Trends in Biochemical Sciences* **1999**; 24(9):335-339.
- De Haro, C., Mendez, R., and Santoyo, J. The eIF-2 α kinases and the control of protein synthesis. *FASEB Journal* **1996**; 10(12):1378-1387.
- Dever, T. E. Translation initiation: adept at adapting. *Trends in Biochemical Sciences* **1999**; 24(10):398-403.
- De Laat, W. L., Jaspers, N. G., and Hoeijmakers, J. H. Molecular mechanism of nucleotide excision repair. *Genes and Development* **1999**; 13(7):768-785.
- De Leeuw, W. J., Slagboom, P. E., and Vijg, J. Quantitative comparison of mRNA levels in mammalian tissues: 28S ribosomal RNA level as an accurate internal control. *Nucleic Acids Research* **1989**; 17(23):10137-10138.
- Dent, A. L., Shaffer, A. L., Yu, X., Allman, D., and Staudt, L. M. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* **1997**; 276(5312):589-592.
- DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**; 278(5338):680-686.
- Desprez, P. Y., Poujol, D., and Saez, S. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH, E. C. 1.2.1.12.) gene expression in two malignant human mammary epithelial cell lines: BT-20 and MCF-7. Regulation of gene expression by 1,25-dihydroxyvitamin D3 (1,25-(OH) $_2$ D3). *Cancer Letters* **1992**; 64(3):219-224.
- Dieci, G., and Sentenac, A. Facilitated recycling pathway for RNA polymerase III. *Cell* **1996**; 84(2):245-252.
- Ding, D., Parkhurst, S. M., Halsell, S. R., and Lipshitz, H. D. Dynamic Hsp83 RNA localization during *Drosophila* oogenesis and embryogenesis. *Molecular and Cellular Biology* **1993**; 13(6):3773-3781.
- Dodson, G., and Steiner, D. The role of assembly in insulin's biosynthesis. *Current Opinion in Structural Biology* **1998**; 8(2):189-194.
- Doyle, H. A., and Mamula, M. J. Post-translational protein modifications in antigen recognition and autoimmunity. *Trends in Immunology* **2001**; 22(8):443-449.
- Dralyuk, I., Brudno, M., Gelfand, M. S., Zorn, M., and Dubchak, I. ASDB: database of alternatively spliced genes. *Nucleic Acids Research* **2000**; 28(1):296-297.
- Dubowy, J., and Macdonald, P. M. Localization of mRNAs to the oocyte is common in *Drosophila* ovaries. *Mechanisms of Development* **1998**; 70(1-2):193-195.
- Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2001**; 2(12):919-929.
- Edwards, D. R., and Denhardt, D. T. A study of mitochondrial and nuclear transcription with cloned cDNA probes. Changes in the relative abundance of mitochondrial transcripts after stimulation of quiescent mouse fibroblasts. *Experimental Cell Research* **1985**; 157(1):127-143.
- Eikhom, T. S., Abraham, K. A., and Dowben, R. M. Ribosomal RNA metabolism in synchronized plasmacytoma cells. *Experimental Cell Research* **1975**; 91(2):301-309.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(25):14863-14868.
- Elefant, F., Su, Y., Liebhaber, S. A., and Cooke, N. E. Patterns of histone acetylation suggest dual pathways for gene activation by a bifunctional locus control region. *EMBO Journal* **2000**; 19(24):6814-6822.

- Eng, J. K., McCormack, A. L., and Yates, J. R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of American Society for Mass Spectrometry* **1994**; *5*(11):976-989.
- Engel, J. D., and Tanimoto, K. Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. *Cell* **2000**; *100*(5):499-502.
- Ernst, P. and Smale, S. T. Combinatorial regulation of transcription II: The immunoglobulin μ heavy chain gene. *Immunity* **1995**; *2*(5):427-438.
- Farazi, T. A., Waksman, G., and Gordon, J. I. The biology and enzymology of protein N-myristoylation. *Journal of Biological Chemistry* **2001**; *276*(43):39501-39504.
- Fields, S., and Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **1989**; *340*(6230):245-246.
- Fiering, S., Epner, E., Robinson, K., Zhuang, Y., Telling, A., Hu, M., Martin, D. I., Enver, T., Ley, T. J., and Groudine, M. Targeted deletion of 5'HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus. *Genes and Development* **1995**; *9*(18):2203-2213.
- Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenkov, V. V. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology* **1996**; *16*(6):2802-2813.
- Finke, J., Fritzen, R., Ternes, P., Lange, W., and Dölken, G. An improved strategy and a useful housekeeping gene for RNA analysis from formalin-fixed, paraffin-embedded tissues by PCR. *Biotechnology* **1993**; *14*(3):448-453.
- Flaherty, K. T., Stevenson, J. P., O'Dwyer, P. J. Antisense therapeutics: lessons from early clinical trials. *Current Opinion in Oncology* **2001**; *13*(6):499-505.
- Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C., and Legrain, P. A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* **2000**; *242*(1-2):369-379.
- Flores, A., Briand, J.-F., Gadal, O., Andrau, J. C., Rubbi, L., Van Mullem, V., Boschiero, C., Goussot, M., Marck, C., Carles, C., Thuriaux, P., Sentenac, A., and Werner, M. A protein-protein interaction map of yeast RNA polymerase III. *Proceedings of the National Academy of Sciences of the USA* **1999**; *96*(14):7815-7820.
- Fornerod, M., Ohno, M., Yoshida, M., and Mattaj, I. W. CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell* **1997**; *90*(6):1051-1060.
- Foss, R. D., Guha-Thakurta, N., Conran, R. M., and Gutman, P. Effects of fixative and fixation time on the extraction and polymerase chain reaction amplification of RNA from paraffin-embedded tissue. *Diagnostic Molecular Pathology* **1994**; *3*(3):148-155.
- Freedman, R. B., and Hawkins, H. C. The enzymology of post-translational modification of proteins. Freedman, R. B., and Hawkins, H. C., editors. Academic Press, London, United Kingdom **1980**.
- Freeman, W. M., Walker, S. J., and Vrana, K. E. Quantitative RT-PCR: pitfalls and potential. *Biotechniques* **1999**; *26*(1):112-125.
- Fry, C. J., and Peterson, C. L. Chromatin remodeling enzymes: who's on first? *Current Biology* **2001**; *11*(5):R185-197.
- Fry, C. J., and Peterson, C. L. Transcription: Unlocking the gates to gene expression. *Science* **2002**; *295*(5561):1847-1848.
- Futcher, B. Cell cycle synchronization. *Methods Cell Science* **1999**; *21*(2-3):79-86.
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **2000**; *11*(12):4241-4257.
- Gasser, S. M., and Laemmli, U. K. Improved methods for the isolation of individual and clustered mitotic chromosomes. *Experimental Cell Research* **1987**; *173*(1):85-98.
- Gavin, A.-C., Bösch, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. Functional organization of the yeast pro-

- teome by systematic analysis of protein complexes. *Nature* **2002**; 415(6868):141-147.
- Ge, H., Liu, Z., Church, G. M., and Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics* **2001**; 29(4):482-486.
- Geiduschek, E. P., and Kassavetis, G. A. The RNA polymerase III transcription apparatus. *Journal of Molecular Biology* **2001**; 310(1):1-26.
- Gelfand, M. S., Dubchak, I., Dralyuk, I., and Zorn, M. ASDB: database of alternatively spliced genes. *Nucleic Acids Research* **1999**; 27(1):301-302.
- Gibson, U. E. M., Heid, C. A., and Williams, P. M. A novel method for real time quantitative RT-PCR. *Genome Research* **1996**; 6(10):995-1001.
- Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., and Kornberg, R. D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* **2001**; 292(5523):1876-1882.
- Golden, T., Dean, N. M., and Honkanen, R. E. Use of antisense oligonucleotides: advantages, controls, and cardiovascular tissue. *Microcirculation* **2002**; 9(1):51-64.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**; 286(5439):531-537.
- Goodlett, D. R., Keller, A., Watts, J. D., Newitt, R., Yi, E. C., Purvine, S., Eng, J. K., von Haller, P., Aebersold, R., and Kolker, E. Differential stable isotope labeling of peptides for quantitation and *de novo* sequence derivation. *Rapid Communications in Mass Spectrometry* **2001**; 15(14):1241-1221.
- Goodlett, D. R., and Yi, E. C. Proteomics without polyacrylamide: qualitative and quantitative uses of tandem mass spectrometry in proteome analysis. *Functional and Integrative Genomics* **2002**; in press.
- Gorlich, D., and Mattaj, I. W. Nucleocytoplasmic transport. *Science* **1996**; 271(5255):1513-1518.
- Grant, S. R. Dissecting the mechanisms of posttranscriptional gene silencing: divide and conquer. *Cell* **1999**; 96(3):303-306.
- Grant, P. A., and Berger, S. L. Histone acetyltransferase complexes. *Seminars in Cell and Developmental Biology* **1999**; 10(2):169-177.
- Graves, D. J., Martin, B. L., and Wang, J. H. Co- and post-translational modification of proteins: Chemical principles and biological effects. Oxford University Press, New York, New York, USA **1994**.
- Gray, M. W. RNA editing in plant organelles: a fertile field. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(16):8157-8259.
- Gray, M. W., and Covello, P. S. RNA editing in plant mitochondria and chloroplasts. *FASEB Journal* **1993**; 7(1):64-71.
- Green, C. D., Simons, J. F., Taillon, B. E., and Lewin, D. A. Open systems: panoramic views of gene expression. *Journal of Immunological Methods* **2001**; 250(1-2):67-79.
- Greenblatt, J., Nodwell, J. R., and Mason, S. W. Transcriptional antitermination. *Nature* **1993**; 364(6436):401-406.
- Griffin, T. J., Gygi, S. P., Rist, B., Aebersold, R., Loboda, A., Jilkine, A., Ens, W., and Standing, K. G. Quantitative proteomic analysis using a MALDI quadrupole time-of-flight mass spectrometer. *Analytical Chemistry* **2001**; 73(5):978-986.
- Grosschedl, R., Giese, K., and Pagel, J. HMG domain proteins: Architectural elements in the assembly of nucleoprotein structures. *Trends in Genetics* **1994**; 10(3):94-100.
- Grosshans, H., and Slack, F. J. Micro-RNAs: small is plentiful. *Journal of Cell Biology* **2002**; 156(1):17-21.
- Grosveld, F. Activation by locus control regions? *Current Opinion in Genetics and Development* **1999**; 9(2):152-157.
- Gygi, S. P., Rochon, Y., Fianza, B. R., and Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular Cellular Biology* **1999**; 19(3):1720-1730.
- Haghpasand, M., Bourassa, P. A., Franccone, O. L., and Aiello, R. J. Monocyte/macrophage expression of ABCA1 has minimal contribution to plasma HDL levels. *Journal of Clinical Investigation* **2001**; 108(9):1315-1320.
- Hammond, S. M., Caudy, A. A., and Hannon, G. J. Post-transcriptional gene silencing by double-stranded RNA. *Nature Reviews Genetics* **2001**; 2(2):110-119.
- Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional ma-

- chinery. *Microbiology and Molecular Biology Reviews* **1998**; 62(2):465-503.
- Han, K. K., and Martinage, A. Post-translational chemical modification(s) of proteins. *International Journal of Biochemistry* **1992**; 24(1):19-28.
- Han, D. K., Eng, J., Zhou, H., and Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology* **2001**; 19(10):946-951.
- Hanauer, A., and Mandel, J. L. The glyceraldehyde 3 phosphate dehydrogenase gene family: structure of a human cDNA and of an X chromosome linked pseudogene: amazing complexity of the gene family in mouse. *The EMBO Journal* **1984**; 3(11):2627-2633.
- Harris, C. M., and Poulter, C. D. Recent studies of the mechanism of protein prenylation. *Natural Product Reports* **2000**; 17(2):137-144.
- Hart, C. M., and Laemmli, U. K. Facilitation of chromatin dynamics by SARs. *Current Opinion in Genetics and Development* **1998**; (8)5:519-525.
- Hasselgren, P.-O., and Fischer, J. E. The ubiquitin-proteasome pathway. *Annals of Surgery* **1997**; 225(3):307-316.
- Hatfield, D., and Oroszlan, S. The where, what and how of ribosomal frameshifting in retroviral protein synthesis. *Trends in Biochemical Sciences* **1990**; 15(5):186-190.
- Heid, C. A., Stevens, J., Livak, K. J., and Williams, P. M. Real time quantitative PCR. *Genome Research* **1996**; 6(10):986-994.
- Helenius, A., and Aebi, M. Intracellular functions of N-linked glycans. *Science* **2001**; 291(5512):2364-2369.
- Herbert, A., and Rich, A. Left-handed Z-DNA: structure and function. *Genetica* **1999**; 106(1-2):37-47.
- Hershko, A., and Ciechanover, A. The ubiquitin system. *Annual Review of Biochemistry* **1998**; 67:425-479.
- Higgins, J. B., and Casey, P. J. The role of prenylation in G-protein assembly and function. *Cellular Signalling* **1996**; 8(6):433-437.
- Higuchi, R., Fockler, C., Dollinger, G., and Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology* **1993**; 11(9):1026-1030.
- Hill, M. A., and Gunning, P. β and γ actin mRNAs are differentially located within myoblasts. *Journal of Cell Biology* **1993**; 122(4):825-832.
- Hinnebusch, A. G. Involvement of an initiation factor and protein phosphorylation in translational control of GCN4 mRNA. *Trends in Biochemical Sciences* **1990**; 15(4):148-152.
- Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreau, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Mathiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figge, D., and Tyers, M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **2002**; 415(6868):180-183.
- Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annual Review of Genetics* **1996**; 30:405-439.
- Hochstrasser, M. Biochemistry. All in the ubiquitin family. *Science* **2000a**; 289(5479):563-564.
- Hochstrasser, M. Evolution and function of ubiquitin-like protein-conjugation systems. *Nature Cell Biology* **2000**; 2(8):E153-E157.
- Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., and Cooke, M. P. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **2001**; 106(4):413-415.
- Holland, P. M., and Cooper, J. A. Protein modification: docking sites for kinases. *Current Biology* **1999**; 9(9):R329-R331.
- Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **1998**; 95(5):717-728.
- Horvath, A., Berry, E. A., and Maslov, D. A. Translation of the edited mRNA for cytochrome b in trypanosome mitochondria. *Science* **2000**; 287(5458):1639-1640.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker,

- D. D., Gachotte, D., Chakraborty, K., Simon, J., Bard, M., and Friend, S. H. Functional discovery via a compendium of expression profiles. *Cell* **2000**; 102(1):109-126.
- Huttner, W. B. Tyrosine sulfation and the secretory pathway. *Annual Review of Physiology* **1988**; 50:363-376.
- Ilbba, M., and Soll, D. Quality control mechanisms during translation. *Science* **1999**; 286(5446):1893-1897.
- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **2001**; 292(5518):929-934.
- Ishitani, R., Sunaga, K., Hirano, A., Saunders, P., Katsube, N., and Chang, D. M. Evidence that glyceraldehyde-3-phosphate dehydrogenase is involved in age-induced apoptosis in mature cerebellar neurons in culture. *Journal of Neurochemistry* **1996**; 66(3):928-935.
- Ito, Y., Pagano, P. J., Tornheim, K., Brecher, P., and Cohen, R. A. Oxidative stress increases glyceraldehyde-3-phosphate dehydrogenase mRNA levels in isolated rabbit aorta. *American Journal of Physiology* **1996**; 270(1 part 2):H81-H87.
- Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(34):1143-1147.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(3):4569-4574.
- Ivanova, N. B., and Belyavsky, A. V. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research* **1995**; 23(15):2954-2958.
- Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **2001**; 409(6819):533-538.
- Jacob, F., Perrin, D., Sanchez, C., and Monod, J. L'opéron groupes de gènes à expression coordonnée par un opérateur. *Comptes Rendus des Seances de L'academie des Sciences* **1960**; 250:1727-1729.
- Jacob, F., and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **1961**; 3:318-356.
- Jacobson, A., and Peltz, S. W. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annual Review of Biochemistry* **1996**; 65:693-739.
- Jeffery, W. R., Tomlinson, C. R., and Brodeur, R. D. Localization of actin messenger RNA during early ascidian development. *Developmental Biology* **1983**; 99(2):408-417.
- Jimenez, G. S., Khan, S. H., Stommel, J. M., and Wahl, G. M. p53 regulation by post-translational modification and nuclear retention in response to diverse stresses. *Oncogene* **1999**; 18(53):7656-7665.
- Johnson, R. S., Ohguro, H., Palczewski, K., Hurley, J. B., Walsh, K. A., and Neubert, T. A. Heterogeneous N-acylation is a tissue- and species-specific posttranslational modification. *Journal of Biological Chemistry* **1994**; 269(33):21067-21071.
- Johnstone, O., and Lasko, P. Translational regulation and RNA localization in *Drosophila* oocytes and embryos. *Annual Review of Genetics* **2001**; 35:365-406.
- Jones, K. A. Taking a new TAK on tat transactivation. *Genes and Development* **1997**; 11(20):2593-2599.
- Jones, K. A., and Kadonaga, J. T. Exploring the transcription-chromatin interface. *Genes and Development* **2000**; 14(16):1992-1996.
- Jongens, T. A., Hay, B., Jan, L. Y., and Jan, Y. N. The germ cell-less gene product: a posteriorly localized component necessary for germ cell development in *Drosophila*. *Cell* **1992**; 70(4):569-584.
- Karge, W. H., Schaefer, E. J., and Ordovas, J. M. Quantification of mRNA by polymerase chain reaction (PCR) using an internal standard and a nonradioactive detection method. *Methods in Molecular Biology* **1998**; 110:43-61.
- Katze, M. G., Kwieciszewski, B., Goodlett, D. R., Blakely, C. M., Nedderman, P., Tan, S. L., and Aebersold, R. Ser(2194) is a highly conserved major phosphorylation site of the hepatitis C virus nonstructural protein NS5A. *Virology* **2000**; 278(2):501-513.

- Kehoe, J. W., and Bertozzi, C. R. Tyrosine sulfation: a modulator of extracellular protein-protein interactions. *Chemistry and Biology* **2000**; 7(3):R57-R61.
- Kim, T. K., and Maniatis, T. The mechanism of transcriptional synergy of an *in vitro* assembled interferon- β enhanceosome. *Molecular Cell* **1997**; 1(1):119-129.
- Kingston, R. E., Bunker, C. A., and Imbalzano, A., N. Repression and activation by multiprotein complexes that alter chromatin structure. *Genes and Development* **1996**; 10(8):905-920.
- Kislauskis, E. H., Li, Z., Singer, R. H., and Taneja, K. L. Isoform-specific 3'-untranslated sequences sort α -cardiac and β -cytoplasmic actin messenger RNAs to different cytoplasmic compartments. *Journal of Cell Biology* **1993**; 123(1):165-172.
- Kohchi, T., Umesono, K., Ogura, Y., Komine, Y., Nakahigashi, K., Komano, T., Yamada, Y., Ozeki, H., and Ohyama, K. A nicked group II intron and trans-splicing in liverwort, *Marchantia polymorpha*, chloroplasts. *Nucleic Acids Research* **1988**; 16(21):10025-10036.
- Korner, C. G., and Wahle, E. Poly(A) tail shortening by a mammalian poly(A)-specific 3'-exoribonuclease. *Journal of Biological Chemistry* **1997**; 272(16):10448-10456.
- Kozak, M. Nucleotide sequences of 5'-terminal ribosome-protected initiation regions from two reovirus messages. *Nature* **1977**; 269(5627):391-439.
- Kozak, M. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* **1978**; 15(4):1109-1123.
- Kozak, M. Influence of mRNA secondary structure on binding and migration of 40S ribosomal subunits. *Cell* **1980a**; 19(1):79-90.
- Kozak, M. Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. *Cell* **1980b**; 22(1 Part 1):7-8.
- Kozak, M. Role of ATP in binding and migration of 40S ribosomal subunits. *Cell* **1980c**; 22(2 Part 2):459-467.
- Kozak, M. Translation of insulin-related polypeptides from messenger RNAs with tandemly reiterated copies of the ribosome binding site. *Cell* **1983**; 34(3):971-978.
- Kozak, M. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin *in vivo*. *Nature* **1984**; 308(5956):241-246.
- Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **1986a**; 44(2):283-292.
- Kozak, M. Bifunctional messenger RNAs in eukaryotes. *Cell* **1986b**; 47(4):481-483.
- Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research* **1987**; 15(20):8125-8148.
- Kozak, M. The scanning model for translation: an update. *Journal of Cell Biology* **1989**; 108(2):229-241.
- Kozak, M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *Journal of Biological Chemistry* **1991**; 266(30):19867-19870.
- Kozak, M. Regulation of translation in eukaryotic systems. *Annual Review of Cell Biology* **1992**; 8:197-225.
- Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **1999**; 234(2):187-208.
- Kozak, M. Do the 5' untranslated domains of human cDNAs challenge the rules for initiation of translation (or is it *vice versa*)? *Genomics* **2000**; 70(3):396-406.
- Kozak, M. New ways of initiating translation in eukaryotes? *Molecular and Cellular Biology* **2001a**; 21(6):1899-1907.
- Kozak, M. A progress report on translational control in eukaryotes. *Science Signal Transduction Knowledge Environment* **2001b**; 2001(71):PE1.
- Kozak, M., and Shatkin, A. J. Identification of features in 5' terminal fragments from reovirus mRNA which are important for ribosome binding. *Cell* **1978**; 13(1):201-212.
- Kreuzer, K. A., Lass, U., Landt, O., Nitsche, A., Laser, J., Ellerbok, H., Pauli, G., Huhn, D., and Schmidt, C. A. Highly sensitive and specific fluorescence reverse transcription-PCR assay for the pseudogene-free detection of beta-actin transcripts as quantitative reference. *Clinical Chemistry* **1999**; 45(2):297-300.
- Krussel, J. S., Huang, H. Y., Simon, C., Behr, B., Pape, A. R., Wen, Y., Bielfeld, P., and Polan, M. L. Single blastomeres within human preimplantation embryos express different amounts of messenger ribonucleic acid for beta-actin and interleukin-1 receptor type I. *Journal of Clinical Endocrinology and Metabolism* **1998**; 83(3):953-959.

- Lander, E. S., Linton, L. M., Birren, B., Nussbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., Le Vine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendt, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramsier, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., and Szustakowski, J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**; 409(6822): 860-921.
- Landweber, L. F., and Gilbert, W. RNA editing as a source of genetic variation. *Nature* **1993**; 363(6425):179-182.
- Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Science of the USA* **1997**; 94(24):13057-13062.
- Lee, M. G., and van der Ploeg, L. H. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. *Annual Review of Microbiology* **1997**; 51:463-489.
- Lee, T. I., and Young, R. A. Regulation of gene expression by TBP-associated proteins. *Genes and Development* **1998**; 12(10):1398-1408.
- Lee, N. S., Bertrand, E., and Rossi, J. mRNA localization signals can enhance the intracellular effectiveness of hammerhead ribozymes. *RNA* **1999**; 5(9):1200-1209.
- Lee, T. I., and Young, R. A. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics* **2000**; 34:77-137.
- Legrain, P., Wojcik, J., and Gauthier, J. M. Protein-protein interaction maps: a lead towards cellular functions. *Trends in Genetics* **2001**; 17(6):346-352.

- Lemon, B., and Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes and Development* **2000**; 14(10):2551-2569.
- Liang, P., and Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **1992**; 257(5072):967-971.
- Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics* **2001**; 28(4):327-334.
- Lindahl, L., and Hinnebusch, A. Diversity of mechanisms in the regulation of translation in prokaryotes and lower eukaryotes. *Current Opinion in Genetics and Development* **1992**; 2(5):720-726.
- Lis, J., and Wu, C. Protein traffic on the heat shock promoter: parking, stalling, and trucking along. *Cell* **1993**; 74(1):1-4.
- Liu, M. C., Yu, S., Sy, J., Redman, C. M., and Lipmann, F. Tyrosine sulfation of proteins from the human hepatoma cell line HepG2. *Proceedings of the National Academy of Sciences of the USA* **1985**; 82(21):7160-7164.
- Liu, L. F., and Wang, J. C. Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences of the USA* **1987**; 84(20):7024-7027.
- Lockhart, D. J., and Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **2000**; 405(6788):827-836.
- Long, R. M., Singer, R. H., Meng, X., Gonzalez, I., Nasmyth, K., and Jansen, R. P. Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. *Science* **1997**; 277(5324):383-387.
- Lutz, M., Burke, L. J., Barreto, G., Goeman, F., Greb, H., Arnold, R., Schultheiss, H., Brehm, A., Kouzarides, T., Lobanenko, V., and Renkawitz, R. Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Research* **2000**; 28(8):1707-1713.
- Maier, R. M., Zeltz, P., Kossel, H., Bonnard, G., Gualberto, J. M., and Grienemberger, J. M. RNA editing in plant mitochondria and chloroplasts. *Plant Molecular Biology* **1996**; 32(1-2):343-365.
- Maillet, I., Lagniel, G., Perrot, M., Boucherie, H., and Labarre, J. Rapid identification of yeast proteins on two-dimensional gels. *Journal of Biological Chemistry* **1996**; 271(17):10263-10270.
- Maniatis, T. A ubiquitin ligase complex essential for the NF-kappaB, Wnt/wingless, and hedgehog signaling pathways. *Genes and Development* **1999**; 13(5):505-510.
- Mansur, N. R., Meyer-Siegler, K., Wurzer, J. C., and Sirover, M. A. Cell cycle regulation of the glyceraldehyde-3-phosphate dehydrogenase/uracil DNA glycosylase gene in normal human cells. *Nucleic Acids Research* **1993**; 21(4):993-998.
- Maquat, L. E., and Carmichael, G. G. Quality control of mRNA function. *Cell* **2001**; 104(2):173-176.
- Marshall, A., and Hodgson, J. DNA chips: an array of possibilities. *Nature Biotechnology* **1998**; 16(1):27-31.
- Marton, M. J., DeRisi, J. L., Bennett, H. A., Iyer, V. R., Meyer, M. R., Roberts, C. J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D. E., Jr., Hartwell, L. H., Brown, P. O., and Friend, S. H. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine* **1998**; 4(11):1293-1301.
- Marzluff, W. F., Jr., Murphy, E. C., Jr., and Huang, R. C. Transcription of ribonucleic acid in isolated mouse myeloma nuclei. *Biochemistry* **1973**; 12(18):3440-3446.
- Marzluff, W. F., Jr. Transcription of RNA in isolated nuclei. *Methods in Cellular Biology* **1978**; 19:317-332.
- Mattaj, I. W., and Englmeier, L. Nucleocytoplasmic transport: the soluble phase. *Annual Review of Biochemistry* **1998**; 67:265-306.
- McCarthy, J. E. Posttranscriptional control of gene expression in yeast. *Microbiology and Molecular Biology Reviews* **1998**; 62(4):1492-1553.
- Medzihradsky, K. F., Campbell, J. M., Baldwin, M. A., Falick, A. M., Juhasz, P., Vestal, M. L., and Burlingame, A. L. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Analytical Chemistry* **2000**; 72(3):552-558.
- Meleforts, O., and Hentze, M. W. Translational regulation by mRNA/protein interactions in eukaryotic cells: ferritin and beyond. *Bioessays* **1993**; 15(2):85-90.
- Merika, M., and Thanos, D. Enhanceosomes. *Current Opinion in Genetics and Development* **2001**; 11(2):205-208.

- Misteli, T. Cell biology of transcription and pre-mRNA splicing: nuclear architecture meets nuclear function. *Journal of Cell Science* **2000**; 113(Part 11):1841-1849.
- Modrek, B., and Lee, C. A genomic view of alternative splicing. *Nature Genetics* **2002**. 30(1):13-19.
- Molete, J. M., Petrykowska, H., Bouhassira, E. E., Feng, Y. Q., Miller, W., and Hardison, R. C. Sequences flanking hypersensitive sites of the beta-globin locus control region are required for synergistic enhancement. *Molecular and Cellular Biology* **2001**; 21(9): 2969-2980.
- Mowry, K. L., and Cote, C. A. RNA sorting in *Xenopus* oocytes and embryos. *FASEB Journal* **1999**; 13(3):435-445.
- Murtaugh, T. J., Wright, L. S., and Siegel, F. L. Posttranslational modification of calmodulin in rat brain and pituitary. *Journal of Neurochemistry* **1986**; 47(1):164-172.
- Myers, L. C., and Kornberg, R. D. Mediator of transcriptional regulation. *Annual Review of Biochemistry* **2000**; 69:729-749.
- Nakamura, A., Amikura, R., Mukai, M., Kobayashi, S., and Lasko, P. F. Requirement for a noncoding RNA in *Drosophila* polar granules for germ cell establishment. *Science* **1996**; 274(5295):2075-2079.
- Nalivaeva, N. N., and Turner, A. J. Post-translational modifications of proteins: acetylcholinesterase as a model system. *Proteomics* **2001**; 1(6):735-747.
- Neuman-Silberberg, F. S., and Schüpbach, T. The *Drosophila* dorsoventral patterning gene *gurken* produces a dorsally localized RNA and encodes a TGF α -like protein. *Cell* **1993**; 75(1):165-174.
- Niehrs, C., Beisswanger, R., and Huttner, W. B. Protein tyrosine sulfation, 1993 — an update. *Chemico-Biological Interactions* **1994**; 92(1-3):257-271.
- Nilsen, T. W. Trans-splicing in protozoa and helminths. *Infectious Agents and Disease* **1992**; 1(4):212-218.
- Nilsen, T. W. Trans-splicing of nematode pre-messenger RNA. *Annual Review of Microbiology* **1993**; 47:413-440.
- Nilsen, T. W. Evolutionary origin of SL-addition trans-splicing: still an enigma. *Trends in Genetics* **2001**; 17(12):678-680.
- Nudler, E. Transcription elongation: structural basis and mechanisms. *Journal of Molecular Biology* **1999**; 288(1):1-12.
- Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Science of the USA* **1999**; 96(12): 6591-6596.
- Ogbourne, S., and Antalis, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal* **1998**; 331(Part 1):1-14.
- Oh, S. K., and Sarnow, P. Gene regulation: translational initiation by internal ribosome binding. *Current Opinion in Genetics and Development* **1993**; 3(2):295-300.
- Ohno, M., Fornerod, M., and Mattaj, I. W. Nucleocytoplasmic transport: the last 200 nanometers. *Cell* **1998**; 92(3):327-336.
- Paik, W. K., Cho, Y. B., Frost, B., and Kim, S. Cytochrome c methylation. *Biochemistry and Cell Biology* **1989**; 67(9):602-611.
- Palacios, I. M., and St. Johnston, D. Getting the message across: the intracellular localization of mRNAs in higher eukaryotes. *Annual Review of Cell and Developmental Biology* **2001**; 17:569-614.
- Pallen, C. J., Tan, Y. H., and Guy, G. R. Protein phosphatases in cell signalling. *Current Opinion in Cell Biology* **1992**; 4(6):1000-1007.
- Pandey, A., and Mann, M. Proteomics to study genes and genomes. *Nature* **2000**; 405(6788):837-846.
- Parekh, R. B., and Rohlf, C. Post-translational modification of proteins and the discovery of new medicine. *Current Opinion in Biotechnology* **1997**; 8(6):718-723.
- Park, H., Baek, K., Jeon, C., Agarwal, K., and Yoo, O. Characterization of the gene encoding the human transcriptional elongation factor TFIIS. *Gene* **1994**; 139(2):263-267.
- Paule, M. R. Transcription of ribosomal RNA genes by eukaryotic RNA polymerase I. Springer-Verlag, Berlin, Germany **1998**.
- Paule, M. R., and White, R. J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Research* **2000**; 28(6):1283-1298.
- Pereira-Leal, J. B., Hume, A. N., and Seabra, M. C. Prenylation of Rab GTPases: molecular mechanisms and involvement in genetic disease. *FEBS Letters* **2001**; 498(2-3):197-200.
- Peri, S., and Pandey, A. A reassessment of the translation initiation codon in vertebrates. *Trends in Genetics* **2001**; 17(12):685-687.

- Perier, R. C., Junier, T., and Bucher, P. The Eukaryotic Promoter Database EPD. *Nucleic Acids Research* **1998**; 26(1):353-357.
- Perier, R. C., Junier, T., Bonnard, C., and Bucher, P. The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Research* **1999**; 27(1):307-309.
- Perier, R. C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. The eukaryotic promoter database (EPD). *Nucleic Acids Research* **2000**; 28(1):302-303.
- Petersen, B. H., Rapaport, R., Henry, D. P., Huseman, C., and Moore, D. P. Effect of treatment with biosynthetic human growth hormone (GH) on peripheral blood lymphocyte populations and function in growth hormone-deficient children. *Journal of Clinical Endocrinology Metabolism* **1990**; 70(6):1756-1760.
- Phi-Van, L., and Strätling, W. H. Association of DNA with nuclear matrix. *Progress in Molecular and Subcellular Biology* **1990**; 11:1-11.
- Plempner, R. K., and Wolf, D. H. Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends in Biochemical Sciences* **1999**; 24(7):266-270.
- Powell, L. M., Wallis, S. C., Pease, R. J., Edwards, Y. H., Knott, T. J., and Scott, J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **1987**; 50(6):831-840.
- Praz, V., Perier, R., Bonnard, C., and Bucher, P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Research* **2002**; 30(1):322-324.
- Preiss, T., and Hentze, M. W. From factors to mechanisms: translation and translational control in eukaryotes. *Current Opinion in Genetics and Development* **1999**; 9(5):515-521.
- Prives, C., and Manley, J. L. Why is p53 acetylated? *Cell* **2001**; 107(7):815-818.
- Puissant, C., Bayat-Sarmadi, M., Devinoy, E., and Houdebine, L. M. Variation of transferrin mRNA concentration in the rabbit mammary gland during the pregnancy-lactation-weaning cycle and in cultured mammary cells. A comparison with the other major milk protein mRNAs. *European Journal of Endocrinology* **1994**; 130(5):522-529.
- Raeymaekers, L. General principles of quantitative PCR. In: Quantitative PCR Protocols. Kochanowski, B., and Reischl, U., editors. Humana Press, Totowa, New Jersey, USA **1999**:31-41.
- Raff, J. W., Whitfield, W. G., and Glover, D. M. Two distinct mechanisms localise cyclin B transcripts in syncytial *Drosophila* embryos. *Development* **1990**; 110(4):1249-1261.
- Rain, J.-C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., and Legrain, P. The protein-protein interaction map of *Helicobacter pylori*. *Nature* **2001**; 409(6817):211-215.
- Rajagopalan, L. E., and Malter, J. S. Regulation of eukaryotic messenger RNA turnover. *Progress in Nucleic Acid Research and Molecular Biology* **1997**; 56:257-286.
- Ramachandran, C., and Melnick, S. J. Multi-drug resistance in human tumors – molecular diagnosis and clinical significance. *Molecular Diagnosis* **1999**; 4(2):81-94.
- Reines, D., Conaway, J. W., and Conaway, R. C. The RNA polymerase II general elongation factors. *Trends in Biochemical Sciences* **1996**; 21(9):351-355.
- Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. Genome-wide location and function of DNA binding proteins. *Science* **2000**; 290(5500):2306-2309.
- Rhoads, R. E. Regulation of eukaryotic protein synthesis by initiation factors. *Journal of Biological Chemistry* **1993**; 268(5):3017-3020.
- Ripple, M. O., and Wilding, G. Alteration of glyceraldehyde-3-phosphate dehydrogenase activity and messenger RNA content by androgen in human prostate carcinoma cells. *Cancer Research* **1995**; 55(19):4234-4236.
- Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **2000**; 287(5454):873-880.
- Rolland, V., Dugail, I., Le Liepvre, X., and Lavau, M. Evidence of increased glyceraldehyde-3-phosphate dehydrogenase and fatty acid synthetase promoter activities in transiently transfected adipocytes from genetically obese rats. *Journal of Biological Chemistry* **1995**; 270(3):1102-1106.

- Ross, J. mRNA stability in mammalian cells. *Microbiological Reviews* **1995**; 59(3):423-450.
- Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalton, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **2000**; 24(3):227-235.
- Rouault, T., and Klausner, R. Regulation of iron metabolism in eukaryotes. *Current Topics in Cellular Regulation* **1997**; 35:1-19.
- Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vossball, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. Comparative genomics of the eukaryotes. *Science* **2000**; 287(5461): 2204-2215.
- Rusche, L. N., Huang, C. E., Piller, K. J., Hemann, M., Wirtz, E., and Sollner-Webb, B. The two RNA ligases of the *Trypanosoma brucei* RNA editing complex: cloning the essential band IV gene and identifying the band V gene. *Molecular and Cellular Biology* **2001**; 21(4):979-989.
- Ruvolo, P. P., Deng, X., and May, W. S. Phosphorylation of Bcl2 and regulation of apoptosis. *Leukemia* **2001**; 15(4):515-522.
- Russo, V. E. A., Martienssen, R. A., and Riggs, A. D. Epigenetic mechanisms of gene regulation. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York, USA **1996**.
- Ruvkun, G. Molecular biology. Glimpses of a tiny RNA world. *Science* **2001**; 294(5543): 797-799.
- Sachs, A. B. Messenger RNA degradation in eukaryotes. *Cell* **1993**; 74(3):413-421.
- Sachs, A. B., and Buratowski, S. Common themes in translational and transcriptional regulation. *Trends in Biochemical Sciences* **1997**; 22(6):189-192.
- Santamarina-Fojo, S., Remaley, A. T., Neufeld, E. B., and Brewer, H. B., Jr. Regulation and intracellular trafficking of the ABCA1 transporter. *Journal of Lipid Research* **2001**; 42(9): 1339-1345.
- Sarre, T. F. The phosphorylation of eukaryotic initiation factor 2: a principle of translational control in mammalian cells. *Biosystems* **1989**; 22(4):311-325.
- Saunders, C., and Cohen, R. S. The role of oocyte transcription, the 5'UTR, and translation repression and derepression in *Drosophila* gurken mRNA and protein localization. *Molecular Cell* **1999**; 3(1):43-54.
- Scheffner, M., Smith, S., and Jentsch, S. The ubiquitin conjugation system. In: Ubiquitin and the biology of the cell. Peters, J. M., Harris, J. R., and Finley, D., editors. Plenum Press, New York, New York, USA **1998**.
- Schena, M., Shalton, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**; 270(5235):467-470.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* **2000**; 24(3):236-244.
- Schneider, R., Agol, V. I., Andino, R., Bayard, F., Cavener, D. R., Chappell, S. A., Chen, J. J., Darlix, J. L., Dasgupta, A., Donze, O., Duncan, R., Elroy-Stein, O., Farabaugh, P. J., Filipowicz, W., Gale, M. Jr., Gehrke, L., Goldman, E., Groner, Y., Harford, J. B., Hatzoglou, M., He, B., Hellen, C. U., Hentze, M. W., Hershey, J., Hershey, P., Hohn, T., Holcik, M., Hunter, C. P., Igarashi, K., Jackson, R., Jagus, R., Jefferson, L. S., Joshi, B., Kaempfer, R., Katze, M., Kaufman, R. J., Kiledjian, M., Kimball, S. R., Kimchi, A., Kirkegaard, K., Koromilas, A. E., Krug, R. M., Kruys, V., Lamphear, B. J., Lemon, S., Lloyd, R. E., Maquat, L. E., Martinez-Salas, E., Mathews, M. B., Mauro, V. P., Miyamoto, S., Mohr, I., Morris, D. R., Moss, E. G., Nakashima, N., Palmenberg, A., Parkin, N. T.,

- Pe'ery, T., Pelletier, J., Peltz, S., Pestova, T. V., Pilipenko, E. V., Prats, A. C., Racaniello, V., Read, G. S., Rhoads, R. E., Richter, J. D., Rivera-Pomar, R., Rouault, T., Sachs, A., Sarnow, P., Scheper, G. C., Schiff, L., Schoenberg, D. R., Semler, B. L., Siddiqui, A., Skern, T., Sonenberg, N., Tahara, S. M., Thomas, A. A., Toulme, J. J., Wilusz, J., Wimmer, E., Witherell, G., and Wormington, M. New ways of initiating translation in eukaryotes. *Molecular and Cellular Biology* **2001**; 21(23):8238-8246.
- Schübeler, D., Groudine, M., and Bender, M. A. The murine beta-globin locus control region regulates the rate of transcription but not the hyperacetylation of histones at the active genes. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(20):11432-11437.
- Shang, Y., Hu, X., DiRenzo, J., Lazar, M. A., and Brown, M. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell* **2000**; 103(6):843-852.
- Shilatifard, A., Conaway, J. W., and Conaway, R. C. Mechanism and regulation of transcriptional elongation and termination by RNA polymerase II. *Current Opinion in Genetics and Development* **1997**; 7(2):199-204.
- Shine, J., and Dalgarno, L. Determinant of cistron specificity in bacterial ribosomes. *Nature* **1975a**; 254(5495):34-38.
- Shine, J., and Dalgarno, L. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *European Journal of Biochemistry* **1975b**; 57(1):221-230.
- Shpakovski, G. V., Acker, J., Wintzerith, M., Lacroix, J.-F., Thuriaux, P., and Vigneron, M. Four subunits that are shared by the three classes of RNA polymerase are functionally interchangeable between *Homo sapiens* and *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* **1995**; 15(9):4702-4710.
- Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., Loerch, P. M., Leonardson, A., Lum, P. Y., Cavet, G., Wu, L. F., Altschuler, S. J., Edwards, S., King, J., Tsang, J. S., Schimmack, G., Schelter, J. M., Koch, J., Ziman, M., Marton, M. J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M. R., Mao, M., Burchard, J., Kidd, M. J., Dai, H., Phillips, J. W., Linsley, P. S., Stoughton, R., Scherer, S., and Boguski, M. S. Experimental annotation of the human genome using microarray technology. *Nature* **2001**; 409(6822):922-927.
- Shyu, A. B., and Wilkinson, M. F. The double lives of shuttling mRNA binding proteins. *Cell* **2000**; 102(2):135-138.
- Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **2001**; 106(6):697-708.
- Sinensky, M. Recent advances in the study of prenylated proteins. *Biochimica et Biophysica Acta* **2000a**; 1484(2-3):93-106.
- Sinensky, M. Functional aspects of polyisoprenoid protein substituents: roles in protein-protein interaction and trafficking. *Biochimica et Biophysica Acta* **2000b**; 1529(1-3):203-209.
- Singh, R., and Green, M. R. Sequence-specific binding of transfer RNA by glyceraldehyde-3-phosphate dehydrogenase. *Science* **1993**; 259(5093):365-368.
- Sirover, M. A. New insights into an old protein: the functional diversity of mammalian glyceraldehyde-3-phosphate dehydrogenase. *Biochimica et Biophysica Acta* **1999**; 1432(2):159-184.
- Sloof, P., and Benne, R. RNA editing in trypanosome mitochondria: guidelines for models. *FEBS Letters* **1993**; 325(1-2):146-151.
- Smith, H. C., and Sowden, M. P. Base-modification mRNA editing through deamination—the good, the bad and the unregulated. *Trends in Genetics* **1996**; 12(10):418-424.
- Smith, C. W., and Valcarel, J. Alternative premRNA splicing: the logic of combinatorial control. *Trends in Biochemical Sciences* **2000**; 25(8):381-388.
- Sollner-Webb, B. RNA editing. *Current Opinion in Cell Biology* **1991**; 3(6):1056-1061.
- Sollner-Webb, B. Trypanosome RNA editing: resolved. *Science* **1996**; 273(5279):1182-1183.
- Sommerville, J., and Ladomery, M. Masking of mRNA by Y-box proteins. *FASEB Journal* **1996**; 10(4):435-443.
- Soutoglou, E., and Talianidis, I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science* **2002**; 295(5561):1901-1904.
- Spanakis, E. Problems related to the interpretation of autoradiographic data on gene expression using common constitutive tran-

- scripts as controls. *Nucleic Acids Research* **1993**; 21(16):3809-3819.
- Spellmann, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* **1998**; 9(12):3273-3297.
- St. Johnston, D. RNA localization. Getting to the top. *Current Biology* **1994**; 4(1):54-56.
- Stallcup, M. R. Role of protein methylation in chromatin remodeling and transcriptional regulation. *Oncogene* **2001**; 20(24):3014-3020.
- Takizawa, P. A., Sil, A., Swedlow, J. R., Herskowitz, I., and Vale, R. D. Actin-dependent localization of an RNA encoding a cell-fate determinant in yeast. *Nature* **1997**; 389(6646):90-993.
- Tatton, W. G., Chalmers-Redman, R. M. E., Elstner, M., Leesch, W., Jagodzinski, F. B., Stupak, D. P., Sugrue, M. M., and Tatton, N. A. Glyceraldehyde-3-phosphate dehydrogenase in neurodegeneration and apoptosis signaling. *Journal of Neural Transmission Supplements* **2000**; 8(Supplement 60):77-100.
- Thanos, D., and Maniatis, T. Virus induction of human IFN β gene expression requires the assembly of an enhanceosome. *Cell* **1995**; 83(7):1091-1100.
- Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology* **1999**; 75(2-3):291-295.
- Theodorakis, N. G., and Cleveland, D. W. Physical evidence for cotranslational regulation of beta-tubulin mRNA degradation. *Molecular Cell Biology* **1992**; 12(2):791-799.
- Towbin, H., Staehelin, T., and Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the USA* **1979**; 76(9):4350-4354.
- Tuboi, S., Taniguchi, N., and Katunuma, N. The post-translational modification of proteins: Roles in molecular and cellular biology. Tuboi, S., Taniguchi, N., and Katunuma, N., editors. Japan Scientific Societies Press, Tokyo, Japan **1992**.
- Turner, B. M. Histone acetylation as an epigenetic determinant of long-term transcriptional competence. *Cellular and Molecular Life Sciences* **1998**; 54(1):21-31.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **2000**; 403(6770):623-627.
- Uptain, S. M., Kane, C. M., and Chamberlin, M. J. Basic mechanisms of transcript elongation and its regulation. *Annual Review of Biochemistry* **1997**; 66:117-172.
- Vagner, S., Galy, B., and Pyronnet, S. Irresistible IRES. Attracting the translation machinery to internal ribosome entry sites. *EMBO Reports* **2001**; 2(10):893-898.
- Van Waarde, A. What is the function of protein carboxyl methylation? *Comparative Biochemistry and Physiology B* **1987**; 86(3):423-438.
- Varshavsky, A. The ubiquitin system. *Trends in Biochemical Sciences* **1997**; 22(19):383-387.
- Varshavsky, A., Turner, G., Du, F., and Xie, Y. Felix Hoppe-Seyler Lecture 2000. The ubiquitin system and the N-end rule pathway. *Biological Chemistry* **2000**; 381(9-10):779-789.
- Varro, A., and Dockray, G. J. Post-translational processing of progastrin: inhibition of cleavage, phosphorylation and sulphation by brefeldin A. *Biochemical Journal* **1993**; 295(Part 3):813-819.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **1995**; 270(5235):484-487.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K. W. Characterization of the yeast transcriptome. *Cell* **1997**; 88(2):243-251.
- Venkateswaran, A., Laffitte, B. A., Joseph, S. B., Mak, P. A., Wilpitz, D. C., Edwards, P. A., and Tontonoz, P. Control of cellular cholesterol efflux by the nuclear oxysterol receptor LXR α . *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(22):12097-12102.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L.,

- Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yoosheph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu X. The sequence of the human genome. *Science* **2001**; 291(5507):1304-1351.
- Vidal, M., and Legrain, P. Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Research* **1999**; 27(4):919-929.
- Vosseller, K., Wells, L., and Hart, G. W. Nucleocytoplasmic O-glycosylation: O-GlcNAc and functional proteomics. *Biochimie* **2001**; 83(7):575-581.
- Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **2000**; 287(5450):116-22.
- Wang, A. H., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., and Rich, A. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **1979**; 282(5740):680-686.
- Wang, C., Dickinson, L. K., and Lehmann, R. Genetics of nanos localization in *Drosophila*. *Developmental Dynamics* **1994**; 199(2): 103-115.
- Wang, N., Silver, D. L., Thiele, C., and Tall, A. R. ATP-binding cassette transporter A1 (ABCA1) functions as a cholesterol efflux regulatory protein. *Journal of Biological Chemistry* **2001**; 276(26):23742-23747.
- Warrington, J. A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological Genomics* **2000**; 2(3):143-147.
- Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. The functioning of higher eukaryotic genes. In: Molecular

- biology of the gene. Benjamin, New York, New York **1965**; 1:704.
- Weckwerth, W., Willmitzer, L., and Fiehn, O. Comparative quantification and identification of phosphoproteins using stable isotope labeling and liquid chromatography/mass spectrometry. *Rapid Communications in Mass Spectrometry* **2000**; 14(18):1677-1681.
- Weinzierl, R. O. J., Dynlacht, B. D., Tijan, R. Largest subunit of *Drosophila* transcription factor II-D directs assembly of a complex containing TBP and a coactivator. *Nature* **1993**; 362(6420):511-517.
- Weinzierl, R. O. J. Mechanisms of gene expression: structure, function and evolution of the base transcription machinery. Imperial College Press, London, United Kingdom **1999**.
- Weiss, R. B. Ribosomal frameshifting, jumping and readthrough. *Current Opinion in Cell Biology* **1991**; 3(6):1051-1055.
- White, R. J. RNA polymerase III transcription. Springer-Verlag, Berlin, Germany **1998**.
- Wickens, M. In the beginning is the end: regulation of poly(A) addition and removal during early development. *Trends in Biochemical Sciences* **1990**; 15(8):320-324.
- Wickner, S., Maurizi, M. R., and Gottesman, S. Posttranslational quality control: folding, re-folding, and degrading proteins. *Science* **1999**; 286(5446):1888-1893.
- Wilkinson, K. D. Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB Journal* **1997**; 11(14):1245-1256.
- Wilson, G. M., and Brewer, G. The search for trans-acting factors controlling messenger RNA decay. *Progress in Nucleic Acid Research and Molecular Biology* **1999**; 62:257-291.
- Wolffe, A. P., and Hayes, J. J. Chromatin disruption and modification. *Nucleic Acids Research* **1999**; 27(3):711-720.
- Woychik, N. A. Fractions to functions: RNA polymerase II thirty years later. *Cold Spring Harbor Symposia on Quantitative Biology* **1998**; 63:311-317.
- Yanofsky, C. Transcription attenuation. *Journal of Biological Chemistry* **1988**; 263(2):609-612.
- Yie, J., Senger, K., and Thanos, D. Mechanism by which the IFN- β enhanceosome activates transcription. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(23):13108-13113.
- Yudkovsky, N., Ranish, J. A., and Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **2000**; 408(6809):225-229.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. *Science* **1997**; 276(5316):1268-1272.
- Zhong, H., and Simons, J. W. Direct comparison of GAPDH, β -Actin, cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. *Biochemical and Biophysical Research Communications* **1999**; 259(3):523-526.
- Zhu, H., and Snyder, M. Protein arrays and microarrays. *Current Opinions in Chemistry and Biology* **2001**; 5(1):40-45.
- Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M., and Snyder, M. Global analysis of protein activities using proteome chips. *Science* **2001**; 293(5537):2101-2105.

2

Sample preparation and supplementary tools

2.1

Introduction

Scientists who wish to analyse the expression of genes must first define (i) the material of interest and (ii) the targets of interest. Traditionally, both tissues and isolated or cultured cells are the material of first choice for analysing gene expression, whereas RNA and proteins comprise the targets of interest within these samples.

Gene expression studies on healthy and pathological tissues have been used to identify genes involved in the pathogenesis of a disease or which are of interest as drug targets. Expression studies on tissues have also been used to predict the outcome of disease. In particular, many such studies have been performed in the field of cancer (*e.g.*, Blick *et al.*, 1984; Escot *et al.*, 1991; Groner, 1992; Kononen *et al.*, 1998; Lehr *et al.*, 2001; Pillai, 1991; Rosson & Tereba, 1983; van 't Veer *et al.*, 2002), but have also been carried out in the field of arteriosclerosis (*e.g.*, Lawn *et al.*, 2001; O'Brien *et al.*, 1992; Oemar *et al.*, 1997), hepatitis (*e.g.*, Yamashita *et al.*, 2001) and others using serial analysis of gene expression (SAGE), DNA or tissue microarrays, northern blotting, quantitative RT-PCR, *in situ* hybridisation (ISH), immunohistochemistry (IHC) or enzyme-linked immunosorbent assays (ELISAs).

Expression profiling has also contributed to our understanding of tissue differentiation and development (*e.g.*, Simon *et al.*, 1985), tissue regeneration (*e.g.*, Fausto, 1984) and to our knowledge of tissue functions such as growth or transport processes (*e.g.*, Müller *et al.*, 1983).

The main method currently used to characterise tissue-specific expression of a novel gene of interest is the tissue-specific northern blot. In the past years, however, much energy has been devoted to analyse the expression of all known genes in different kinds of tissues of an organism. A well known example is the Mouse Gene Expression Database which forms part of the framework of the Mouse Genome Informatics Project (<http://www.informatics.jax.org/>) (Ringwald *et al.*, 1994, 1997, 1999, 2000a, 2000b and 2001). As part of this project a database has been constructed that contains data on the expression of all known genes during the development of mice. The expression data have been generated using immunohistochemistry, *in situ* hybridisation, northern and western blotting, RNase protection assays and RT-PCR. A further important example is the tissue-based gene expression database with its appendant tools (cDNA xProfiler; Differential

Gene Expression Displayer, DGED; Gene Library Summarizer, GLS) at the National Cancer Institute (available at <http://cgap.nci.nih.gov/Tissues>; see chapter 7 for more details). The latter databases contain information about the content of cDNA libraries derived from, for example, several whole tissue specimens and microdissected tissues (*e.g.*, adrenal cortex, cerebellum, heart, kidney). Several gene expression databases are described in chapter 7 of this book.

It is now recognised that whole tissue specimens have important disadvantages when they are used for analysing gene expression. First, tissues normally consist of different kinds of cells. For example, a section of liver contains hepatocytes, Ito cells, Kupffer's cells, vascular cells, bile duct epithelium and so on. This is of particular importance with gene expression analysis methods such as SAGE, DNA microarrays, or differential display RT-PCR that require large amounts of homogeneous material for RNA extraction. For this reason, methods such as *in situ* hybridisation or immunohistochemistry have generally been used to investigate gene expression *in vivo*. The amount of tissue required may limit analysis, and the tissue of interest may be hard to obtain (*e.g.*, tissues from living humans). Much interest has therefore focused on the use of cell lines derived from healthy and pathological tissues that allow the analysis of gene expression in tissue-specific cells *in vitro*. Frequently, such cell lines serve as cell culture models to study the pathology of a disease. For example, several cell culture models have been generated from healthy and pathological cells (Braun *et al.*, 2000) to study immunology (Abraham, 2000), cancer and leukemia (Chaganti & Houldsworth, 2000; Drexler & Matsuo, 2000; Drexler *et al.*, 2000; Pageot *et al.*, 2000); neurology (Gokhan *et al.*, 1998), reproduction (Fleming, 2000) and infectious diseases such as spongiform encephalopathy (Beranger *et al.*, 2001). A thorough listing of the cell lines that are currently available may be obtained from the American Type Culture Collection (ATCC) at <http://www.atcc.org> (Hay, 1996; Phelps *et al.*, 1996) and from the German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, DSMZ) at <http://www.dsmz.de/> (Dirks *et al.*, 1999). In addition, several gene-deficient or stably transfected cell lines are available (Bunz, 2002; Gumperz, 2000; Winding & Berchtold, 2001; Zhong & Minneman, 2000).

A well-known example for near-genome-wide expression studies on cell lines is the SAGE gene to tag mapping approach at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/SAGE/>). As part of this project SAGE-map, a searchable SAGE library database, derived from several normal and pathological cell lines and tissues (*e.g.*, prostate carcinoma, normal prostate tissue, glioblastoma multiform cell lines and tumours, epithelial cells) from a variety of organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Bos taurus* and *Danio rerio*) has been created. This database contains data on the frequency of SAGE tags of UniGene clusters in the SAGE libraries (Lal *et al.*, 1999). Further details may be found in chapter 7.

Central areas of gene expression profiling using cell lines have been the study of cancer (Alizadeh *et al.*, 2000; Gemma *et al.*, 2001; Golub *et al.*, 1999; Mills *et al.*, 2001a; Mok *et al.*, 2001; Nacht *et al.*, 1999; Peale & Gerritsen, 2001; Polyak & Riggin, 2001; Velculescu *et al.*, 2000; Waghray *et al.*, 2001), apoptosis (Hofmann *et al.*,

2001), arteriosclerosis (de Waard *et al.*, 1999; Mikita *et al.*, 2001; Shiffman *et al.*, 2000), development and differentiation (Wigle *et al.*, 2001), and inflammation and infection (Hoffmann *et al.*, 2001; Huang *et al.*, 2001).

The phenotype of cells may change during cell culture. For example, smooth muscle cells may switch from the synthetic to the contractile type during which the expression of genes such as of smoothelin and calponin may alter (Samaha *et al.*, 1996; Sobue *et al.*, 1999; van der Loop *et al.*, 1997; Worth *et al.*, 2001). For this reason, regular checks of the identities and the purity of cell lines are required (Kaplan & Hukku, 1998; MacLeod, 1992; Stacey *et al.*, 1991; Webb & Debenham, 1992). Although there may be difficulties in transferring results from cell culture experiments to the *in vivo* situation (Carlson & Chadwick, 1987), the ready availability of such cultures often makes them the first choice for experimenters.

Recently, several attempts have been made to overcome the disadvantages of cell culture experiments. These have included the use of laser-assisted tissue microdissection to isolate single cells or cell populations from whole tissue specimens. This technique is described in detail in a later section of this chapter. Several quantification and *in vitro* amplification protocols have been published to allow use of the RNA in such samples for gene expression profiling (Baugh *et al.*, 2001; Fink *et al.*, 2001; Harsch *et al.*, 2001; Karsten *et al.*, 2002; Luzzi *et al.*, 2001; Ohyama *et al.*, 2000; Phillips & Eberwine, 1996; Wang *et al.*, 2000).

This chapter also describes several methods for the preparation of homogeneous cell populations including immunopurification, countercurrent elutriation, density-gradient centrifugation, flow cytometry and tissue microdissection. The most commonly used methods for the isolation of RNAs and proteins are also described.

2.2

Preparation of cells and tissues

2.2.1

Immunopurification of cells

The need to separate cells arises from the need to isolate specific cell types from heterogeneous populations such as tissue or blood. Cell separation techniques mediated by antibody/antigen reactions are widely used in both research and clinical laboratories. This has been facilitated by the generation of better monoclonal antibodies with improved specificity for defined antigen epitopes.

Immunopurification of homogeneous cell populations from heterogeneous cell populations has been widely used for the isolation of a range of different cell types (blood cells, immune cells, endothelial cells, neural cells, *etc.*) and is well documented (Molday *et al.*, 1977; Flø *et al.*, 1991; Hadjivassiliou *et al.*, 2000; Hansel *et al.*, 1989; Hewett & Murray, 1993; Luxembourg *et al.*, 1998; Manyonda *et al.*, 1992; Murphy *et al.*, 1992). Cell isolation by immunopurification has been also termed immunomagnetic purification or immunoprecipitation of cells.

Immunomagnetic purification is often performed using Dynabeads (DynaL Biotech GmbH, Hamburg, Germany). This technology allows the simple, rapid and reliable separation of cells requiring little more than a specific antibody coupled to a magnetic bead and a magnet. Dynabeads are uniform, spherical, superparamagnetic, monodisperse polymer particles, coated with cell type-specific antibodies. These beads have an even dispersion of magnetic material ($\delta\text{-Fe}_2\text{O}_3$ and Fe_3O_4) throughout the bead. They can be added to a heterogeneous cell suspension in which they bind to the desired target cells. The resulting complex consisting of beads and cells can be removed from the suspension using special magnetic particle concentrators. The principle of immunomagnetic purification of cells is illustrated in figure 2.1.

The method of immunomagnetic cell isolation makes two strategies feasible: (i) positive and (ii) negative cell isolation procedures. Both strategies are shown in figure 2.2.

Positive isolation is a procedure by which predefined cellular subsets can be isolated directly from complex mixtures of cells such as blood. Usually, positive cell isolation is performed using beads coated with specific antibodies in the direct method (see figure 2.1.A). Antibody-coupled beads are directly mixed with the cell suspension. After formation of the cell/antibody/bead complex, supernatant is removed by pelleting the bead-containing complexes using a magnet. The remaining pelleted complexes are washed and a population of well-defined target cells with high purity is revealed (DynaL A. S., 1996).

In negative isolation, as the name suggests, the cell population is purified by removing all unwanted cell types. The procedure of negative selection can be performed by using either the direct or the indirect method. In the direct method, a mixture of beads coated with primary antibody specific to the contaminating subsets is directly applied to the cell suspension. In the indirect method, primary antibodies reacting with the contaminating subsets are first added to the cell suspension. After incubation, the excess of unbound antibodies is removed and the cell/antibody complexes are mixed with beads coated with the secondary antibody. The rosetted cells are subsequently removed using the magnet pelleting procedure. Thus, cells of interest can be easily separated. In general, purity obtained using negative isolation procedures is lower than for positive isolation and the indirect method is more efficient than the direct method in removing target cells from a heterogeneous suspension.

After cell isolation with antibody-coupled magnetic beads, cells bound to the antibody-coated beads must usually be detached from the beads. Several methods have been developed for this purpose. The detachment is possible using polyclonal antibodies that react with the F_{ab} fragments of monoclonal antibodies (DynaL Biotech GmbH; Geretti *et al.*, 1993), by overnight incubation of positively isolated cells at 37 °C (Berardi *et al.*, 1995; Puri *et al.*, 1994), enzymatic cleavage with the proteolytic enzymes pronase (Anderson *et al.*, 1993), chymopapain (Nolta *et al.*, 1995; Traycoff *et al.*, 1995) or trypsin (Curtis *et al.*, 1992; Galileo *et al.*, 1991), using competing molecules (Jackson *et al.*, 1990) or disruption of a linker interaction sensitive to tonicity change (Scouten & Konecny, 1992).

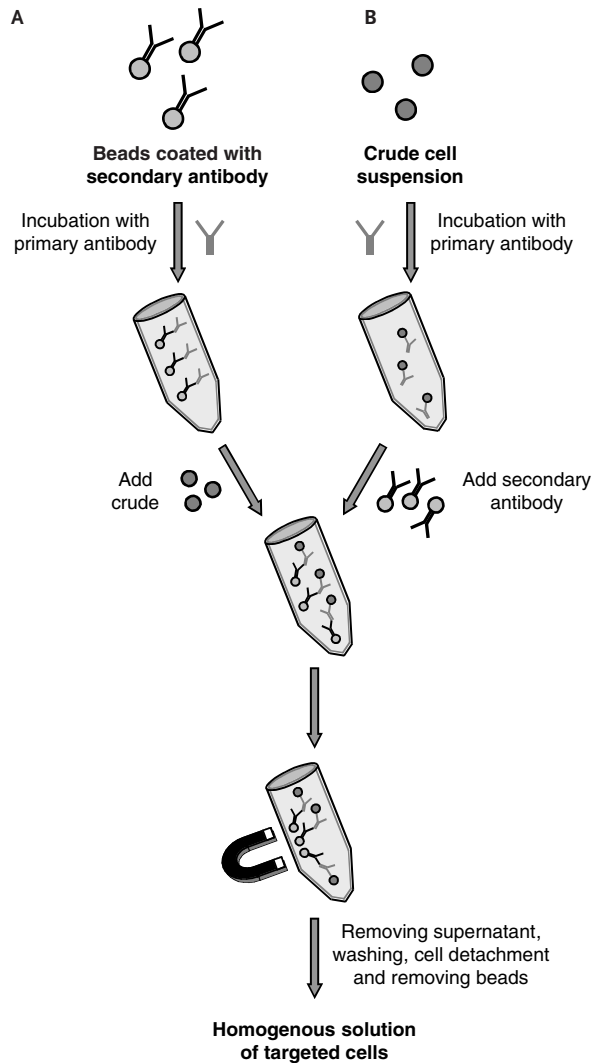


Figure 2.1 Direct and indirect isolation technique using immunomagnetic beads. A) Direct isolation technique. Magnetic beads coated with a specific secondary antibody are directly applied to a solution of a target-specific primary antibody. After binding of the primary antibody to the secondary antibody, the resulting complex is directly applied to the heterogeneous cell suspension. During incubation the target cells bind to the primary antibodies coupled to the beads *via* binding to the primary antibodies. After formation of the cell/antibody/bead complex the target cells can be recovered or removed using a magnet. **B)** Indirect isolation technique. In the first step, the target-specific

primary antibody is added to the cell suspension. After washing to remove the unbound antibody, the antibody/cell complex is mixed with magnetic beads coated with a secondary antibody. The complex consisting of primary antibodies and cells binds to the antibodies of the beads. Using a magnet the resulting complex of target cells, antibodies and beads can be removed from the cell suspension. The direct method requires less time, whereas the indirect method provides higher flexibility for the user. The figure is adapted from 'Cell separation and protein purification. Technical handbook' (Dynal A. S., Oslo, Norway).

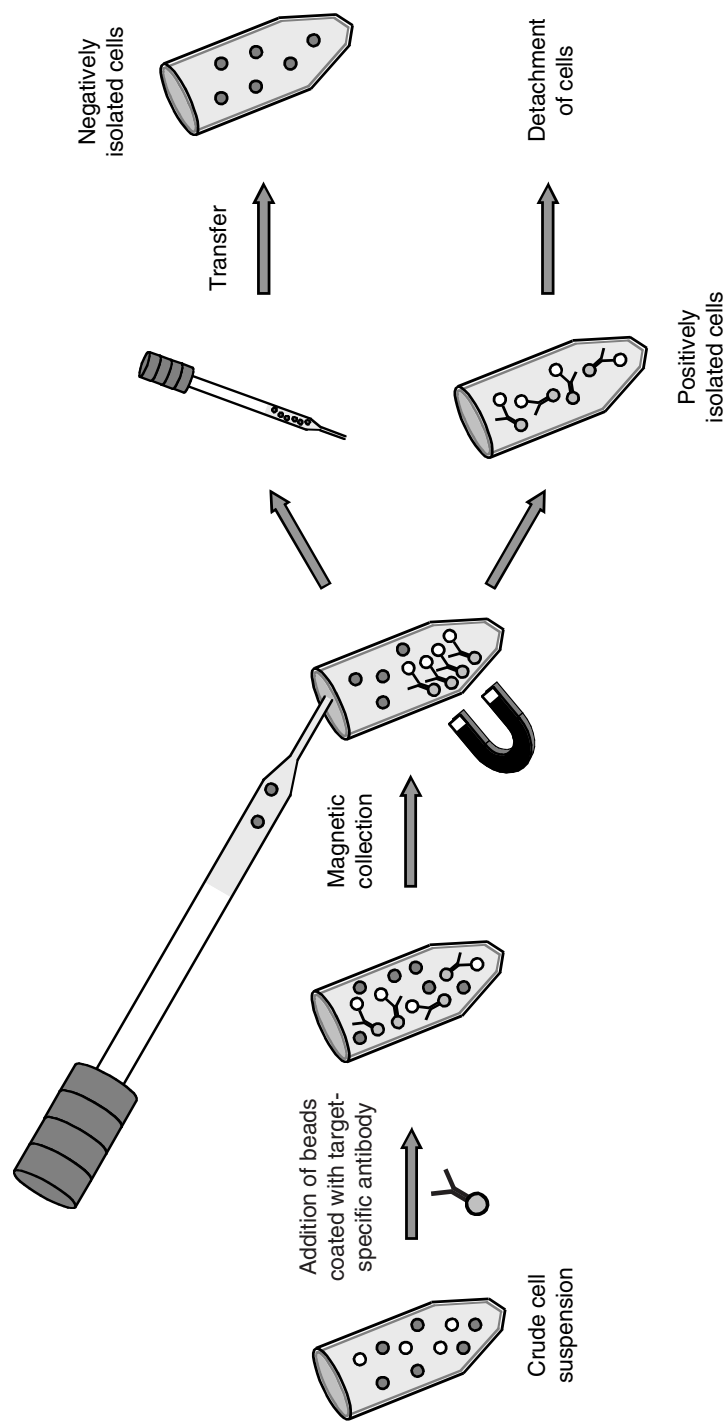


Figure 2.2 Positive and negative isolation procedures using magnetic beads. The differences between the isolating strategies are described in detail in the text. The figure is adapted from 'Cell separation and protein purification. Technical handbook' (Dynal A. S., Oslo, Norway).

2.2.2

Differential-velocity centrifugation/countercurrent elutriation

A method which is used for the non-traumatic isolation of living cells in suspension is differential-velocity centrifugation. This principle is employed both in therapeutic blood cell separation machines and in the technique of countercurrent elutriation (Ito *et al.*, 1975; Stevenson, 1984).

The principle of these methods is that a cell in a flowing solution is subjected to two forces, the shear force imparted by the surrounding liquid and the gravitational, or in a centrifuge, centrifugal force, which is a function of the cellular mass. The cell moves along a vector which is the sum of these two forces.

In practice, most of these systems operate using a centrifuge rotor which contains a cell separation chamber. The rotation of the centrifuge produces a centrifugal force which tends to move the cells in the separation centre away from the axis of rotation. At the same time, an isotonic buffer is pumped through the separation chamber from the edge of the rotor to the centre. This tends to move the cells towards the axis of rotation. Under these circumstances, the cells in solution separate into discrete layers depending on their density, surface characteristics and shape. Each of these layers can be selectively eluted by altering the speed at which the suspension fluid is pumped through the chamber (see figures 2.3 and 2.4).

The advantages of this method are that the *in vivo* characteristics of the cells are largely preserved because the isolation is in an isotonic liquid phase and only one

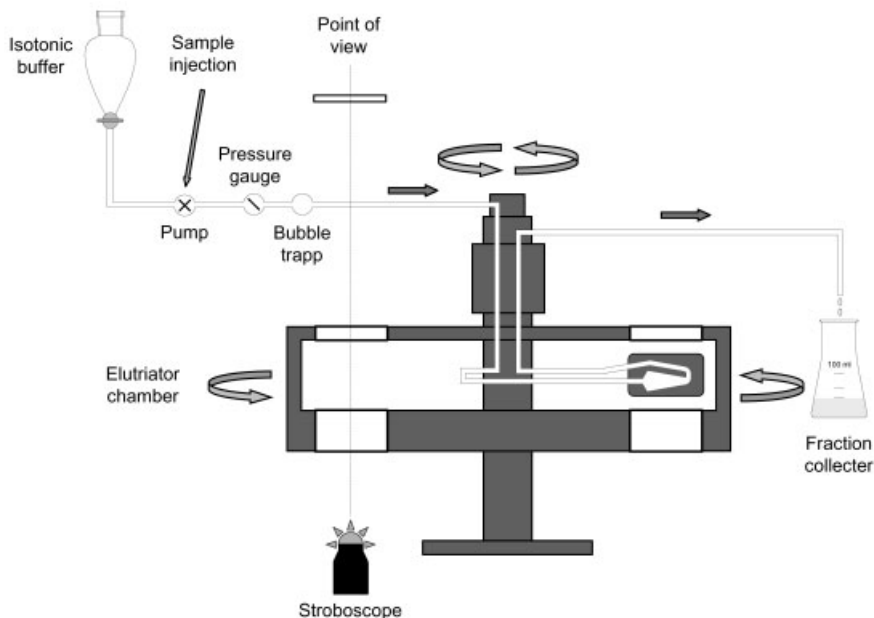


Figure 2.3 Scheme of a cell elutriator for countercurrent centrifugation. Adapted from Beckman Instruments, Inc. (Palo Alto, California, USA).

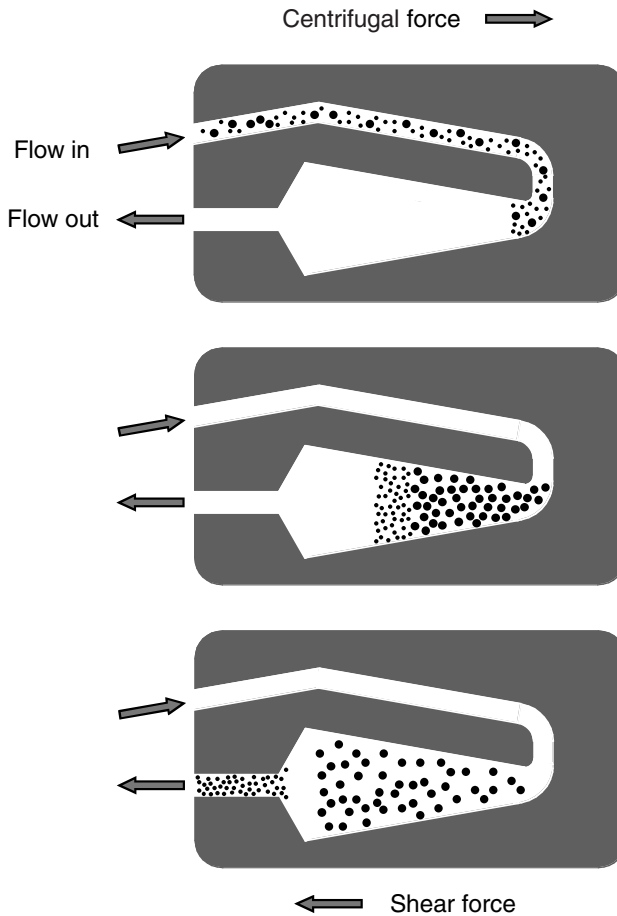


Figure 2.4 Principle of countercurrent centrifugation for the isolation of a homogenous cell population from a heterogenous cell suspension.

centrifugation step (at the end of separation) is required. Moreover, this method offers a high degree of purity of the isolated fractions and isolation of large numbers of cells. For example, up to 1×10^8 monocytes can be isolated from a single human donor in one setting using this method .

The main disadvantage of this method is the large logistic effort and relatively high cost associated with its use. Nevertheless, this method is particular useful for isolating subpopulations of primary cells from human blood, for example, monocytes or other cells of the immune system (Buckley *et al.*, 1984; Cullen *et al.*, 1997; Engel *et al.*, 2001; Lorkowski *et al.*, 2001; O'Keefe *et al.*, 1989; Stevenson, 1984; Stevenson *et al.*, 1981; Yasaka *et al.*, 1981).

2.2.3

Surface affinity chromatography

Heterogeneous cell populations can be separated or partitioned by exploiting differences in their affinity to specific materials. This process is called affinity chromatography and in the case of cell partitioning is often performed as column chromatography.

An early application was the separation of blood cells according to their adherence characteristics. The technique, described as adherence chromatography, used a column of siliconised glass beads of about 200 micrometres in diameter (Rabinowitz, 1964). Although, this method was unable to separate human lymphocytes from erythrocytes, it provided proof of principle of affinity chromatography for cell separation. Since then, several reports have described the chromatographic behaviour of certain types of mammalian cells on columns packed with agarose beads (Tangen *et al.*, 1971), polymethyl-methacrylate beads (Fricova *et al.*, 1974) or cross-linked dextran beads (Fricova *et al.*, 1974; Hudson *et al.*, 1973) and described the adherence of the cells to different column packing materials (Fricova *et al.*, 1974; Grinnell *et al.*, 1972).

Affinity column chromatography methods are aqueous/polymer two-phase systems and have been successfully used for the partitioning of biological macromolecules including several different kinds of cells. Although chromatographic methods have been carried out successfully by the use of various aqueous polymer phase systems for partitioning of viruses (Bengtsson *et al.*, 1962; Norrby & Albertsson, 1960), bacteria (Albertsson, 1965; Albertsson & Baird, 1962) and cell organelles (Ericson, 1974), in the field of mammalian cells, they have been exclusively used for partitioning of different kinds of blood cells (Walter & Selby, 1966 and 1967; Walter *et al.*, 1967, 1968, 1969a and 1969b).

Among the many types of aqueous polymer phase systems available, the dextran-polyethylene glycol system has been the most commonly used for the partition of the cells. In addition of polyethylene glycols, polypropylene glycols have also been described. Several studies on column packings with chemically bonded polyethylene glycols and polypropylene glycols as the stationary phase and dextran constituting the aqueous mobile phase have been performed with peripheral blood cells (Shibusawa, 1999 and 2001, and the references cited therein).

Methoxyethoxymethyl groups are similar to the two oxymethylene groups of polyethylene glycols and have also been used as column packing materials for cell partitioning. Commonly used compounds are cellulose beads (Cellulofine) and cross-linked dextran beads (Sephadex) (Shibusawa, 1999).

Affinity-based approaches have been described using, for example, dielectric affinity separation of human breast cancer cells (Becker *et al.*, 1995) or the different binding affinities of cells to adhesion molecules such as α -selectin for the enrichment of populations of hematopoietic stem and progenitor cells from bone marrow cell preparations (Greenberg & Hammer, 2001).

Although affinity column chromatography methods seems to have the potential for cell fractionation, in particular for the isolation of peripheral blood cells such as

erythrocytes, platelets, granulocytes, lymphocytes and leukocytes, many practical difficulties are experienced, including the absorption of a large number of cells on the column and the ability only to enrich cells but not to separate them quantitatively (Shibusawa, 1999). Therefore, these methods are not the first choice for isolating specific cell types from heterogenous cell populations to perform gene expression studies in the isolated cells.

2.2.4

Density-gradient centrifugation

Centrifugation is probably the oldest and most widely used method of separating cellular and sub-cellular particles in solution. Centrifugation exploits differences in the size and density of components of a mixture in order to separate those components. The terminal velocity of a particle is determined by Stoke's law:

$$v_t = \frac{2 \times R^2 \times (\rho_s - \rho) \times a}{9 \times \mu}$$

where v_t is the terminal velocity of the particle, R the radius of the particle, a the centrifugal acceleration of the centrifuge, μ the viscosity of the medium, ρ_s the density of the particle, and ρ the density of the medium.

From this equation it is apparent that, on average, bigger and heavier particles will travel through the media faster and settle at the bottom of a centrifuge tube in a shorter time.

In practice, three variables in the above equation are altered to isolate particles of interest: a , the centrifugal acceleration of the centrifuge, μ , the viscosity of the medium, and ρ , the density of the medium. The density of the medium can be altered by simple addition of sucrose to the bottom of the tube, or by placing layer after layer of sucrose solutions of different concentrations, and therefore of different densities, in a test tube with the heaviest layer at the bottom and the lightest layer at the top. The cell fraction to be separated is placed on top of the uppermost layer (figure 2.5). A particle will sink if its density is greater than that of the surrounding solution and will continue to sink until it reaches a position where its density equals that of the surrounding solution. A centrifuge is used to speed up the process of reaching this equilibrium. However, in contrast to usual centrifugation steps, the time of centrifugation is not critical as long as the system is permitted to come to equilibrium.

For separation of components of human blood cells a variety of isolation media is commercially available. These are generally solutions of dextran or polysucrose that have been optimised for the isolation of particular blood components such as mononuclear cells or neutrophils based on the original method described by Bøyum (1968, 1977 and 1984). In addition, recent progress was made with the introduction of a continuous flow method that separates cells entirely based on cell density (Ito & Shinomiya, 2001). The method is gentle and capable of processing a large number of cells. The potential capability of the method was demonstrated on separation of lymphocytes and granulocytes from human buffy coat.

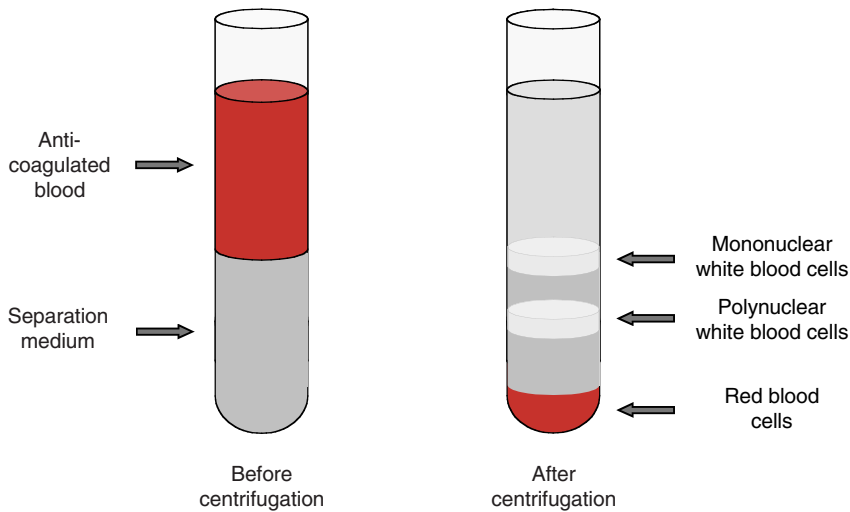


Figure 2.5 Principle of a system for separating blood components using density gradient centrifugation.

The advantages of the density-gradient centrifugation method are that it is rapid, relatively cheap, and can be performed on small sample volumes. The main disadvantage is a less-than-perfect isolation of specific components. Moreover, the use of hypertonic solutions may produce substantial alterations in cell function and cell expression profiles.

2.2.5

Flow cytometry

2.2.5.1 Introduction

After more than three decades of development, flow cytometry can today be considered a mature and effective technique for analysis of multiple cellular and non-cellular parameters.

While the term 'cytometry' is broadly used to describe the measurement of physical and/or chemical characteristics of cells or other biological particles (Shapiro, 1995), 'flow cytometry' is the technique involved in the process. By the application of a precise differential pressure between two solutions, the sheath fluid and the sample suspension, cells and/or particles are forced in a common stream which passes in single line fashion in front of a measuring device. 'Cell sorting' is a natural evolution of the process. After acquisition and identification of the population of interest, an electrical charge is applied to the stream and the cells or particles of interest can be filtered off for collection.

2.2.5.1.1 Applications and abilities of flow cytometry

Flow cytometry is widely involved in many research and clinical applications in the fields of immunology, anticancer research, biology and genetics. Today, flow cytometry is a fundamental tool for drug discovery and high-throughput screening thanks to its ability to analyse several thousand events per second.

The instruments available today can analyse over to 100,000 cells per second, collecting up to 32 parameters for every single event. Full digitalisation of the acquisition process and the fast distribution of the information allow multiparametric four-way sorting at a rate exceeding 100,000 cells per second.

2.2.5.1.2 Origins of flow cytometry

Flow cytometry can be considered the natural evolution of fluorescence microscopy, which was developed in the 1930s. Developments in electronics and computing after the end of the Second World War made new instruments available for new research fields. The electron microscope in the 1950s and the conjugation of a fluorescent compound (fluorescein) with an antibody (Coons & Kaplan, 1950) can be considered two of the milestones in the progression toward modern flow cytometry. The Coulter Counter® developed by Wallace H. Coulter (Coulter, 1956) in response to the need for a reliable, fast and accurate screening method to detect blood cells, opened the way to instruments that could differentiate cell clusters and to the first cell sorters. The introduction of the first FACS (fluorescence activated cell sorter) by Becton Dickinson in 1974, which was due to the efforts of Len Herzenberg, gave a fillip to the diffusion of flow cytometry (Bonner *et al.*, 1972). Since then flow cytometry has constantly gained in importance.

2.2.5.2 Analysers and cell sorters

Analysers measure samples, cells or beads, while cell sorters are cytometers that both analyse and separate populations from a heterogeneous sample.

Most analysers are closed systems, where the stream flows enclosed into a transparent quartz flow cell. High-speed cell sorters are, however, ‘open’ (jet-in-air) systems.

2.2.5.2.1 Fluorescence

Cytometers evaluate external and internal cellular and non-cellular parameters measuring light scattering and fluorescence emission. Fluorescence takes place in molecules, called fluorochromes. Electrons on external orbitals, after being excited by light of a specific wavelength, return to the relaxed state (ground state), emitting fluorescence with longer wavelength. As a result, an emission and an absorption spectra is received; the difference between the excitation and absorption maxima is called *Stoke's shift*. The amount of light absorbed at a given wavelength is called the *absorption coefficient*. The *quantum efficiency* is the number of photons emitted for every photon absorbed.

The aim of the research in the fluorescence field is to find compounds with an absorption coefficient as close as possible to the most important laser lines, but at

the same time the highest quantum efficiency. Extensive catalogues of fluorescent compounds are now available. Coupled with the flexibility of modern multi-laser analysers and sorters, these provide virtually unlimited possibilities for researchers and operators. (see, for example, the Molecular Probes, Inc., (Eugene, Oregon, USA) website at <http://www.molecularprobes.com>).

2.2.5.2.2 Lasers

The acronym 'laser' stands for 'light amplification by stimulated emission of radiation' and though the theory behind the laser is beyond the aim of this chapter, it is fundamental to define basic properties.

All lasers produce an intense beam of light that is monochromatic, coherent and collimated. The term monochromatic refers to the fact that the wavelength (colour) of the laser light is extremely pure compared with other light sources. Coherence means that all the photons (energy) that make up a laser beam have a fixed phase, which causes the light to form a narrow beam with very low divergence that can travel long distance or can be focused on a very small spot transferring a high amount of energy.

All lasers include a gain medium, the source of laser light (*e.g.*, argon gas, helium-neon gas), an excitation source (power supply) and a resonator structure (mirrors that can reflect the light through the gain).

While analysers usually have a fixed laser optical design, cell sorters can accommodate up to three lasers in every combination (argon-ion, krypton, helium-neon or diode) covering all the possible applications. Table 2.1 may be used as a starting point for a choice of lasers.

2.2.5.3 Fluidics

2.2.5.3.1 Flow cells

In most flow cells, the suspension of cells or other particles is guided inside a saline sheath in a process called hydrodynamic focusing. Cytometers are generally equipped with four main flow cell types (Melamed *et al.*, 1990). The choice for a specific cell type is always a compromise between optical sensitivity and sorting performances. These flow cell types are as follows:

- (i) *Jet-in-air*: best for sorting but has inferior optical properties.
- (ii) *Flow-through cuvette*: first-rate for optical properties and can also be used for sorting.
- (iii) *Closed cross flow*: excellent optical properties but has no sorting capabilities.
- (iv) *Open flow across surface*: excellent optical properties but also has no sorting capabilities.

Table 2.1 Applications and possible laser selection. Data of this table are contributed by Dr. W. Roy Overton, Cytomation, Inc., Fort Collins, Colorado, USA (HeCd = helium-cadmium; HeNe = helium-neon).

Application	Fluorescent dye	HeCd		Argon		Laser				HeNe	Dye	Diode
		UV	UV	458	488	UV	406	Krypton				
Immu-no-fluorescence	Cascade Blue	•	•			•						
	AMCA	•	•			•						
	Fluorescein (FITC)				•							
	Phycoerythrin (PE)				•							
	PerCP				•							
	Red 613 (PE:TxR)				•							
	Red 670 (PE:Cy5)				•							
	Rhodamine (TRITC)						•			•		
	Texas Red (TxR)						•			•		
	Allophycocyanin (APC)									•		
DNA	Cy5											•
	APC-Cy7											•
	Hoechst 33342	•	•					•				
	DAPI	•	•					•				
	Propidium Iodide (PI)	•	•		•			•		•		
	7-AAD				•					•		
	Acridine Orange				•							
	Hoechst 33258	•	•					•				
	Chromomycin A3			•								
	Indo-1	•	•									
Calcium	Fluo-3 & Fura Red				•							
	GFP								•			
Reporter genes	enhanced-GFP			•	•							
	enhanced-BFP										•	
	enhanced-CFP			•								
	enhanced-YFP			•								

2.2.5.3.2 Jet-in-air and hydrodynamic focusing

The jet-in-air is the best configuration for high-speed cell sorters due to the possibility to apply high pressures of up to 80 pounds per square inch (PSI; about 5.6 kilograms per square centimetre), reaching a sorting rate of more than 100,000 events per second (figure 2.6).

In the jet-in-air system, the sample suspension (core) is funnelled in a sheath of particle-free isotonic saline solution. Regulators accurately control the differential pressure between the two solutions in order to maintain laminarity of the resulting flow. The velocity gradient created at the interphase between the two solutions, faster in the centre of the core and slower on the edge, forces the particles to flow in a single line. Just below the nozzle, one or more laser beams intercept the streams providing excitation.

2.2.5.3.3 Detectors

Light scatter and fluorescence are transformed into electrical pulses by the use of photomultipliers and photodiodes.

The photodiode is a silicon layer; it produces current when photons simply hit the surface. Photodiodes do not need any external voltage source and they are utilised mainly for light scatter detection. Photodiodes are relatively inexpensive; they show a lower sensitivity compared with photomultipliers, but have wider spectral characteristics.

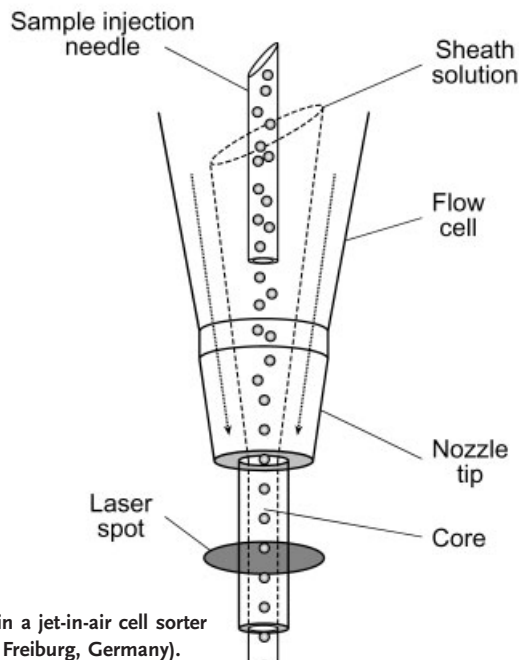


Figure 2.6 Hydrodynamic focusing in a jet-in-air cell sorter (Cytomation Bioinstruments GmbH, Freiburg, Germany).

Photomultipliers, by contrast, are electron tubes. They do require an external electrical source in order to amplify the electron emission through several secondary emissions inside the tube itself. Compared with photodiodes, photomultipliers exhibit narrower spectral characteristics, higher sensitivity, and are the first choice for the detection of fluorescence. Due to the thermionic effect, photomultipliers show evidence of ‘noise’: at high voltages an increase of temperature inside the photomultiplier results in emissions of electrons even in the absence of light emission.

2.2.5.4 Parameters and optics

2.2.5.4.1 Technical terms of detection

Once the laser beam excites the particles, light scattering and fluorescence emission appear at the same time. Forward scatter (FSC) and side scatter (SSC) are the two main parameters of light scatter.

Forward scatter, detected by a photodiode, and in some application by a photomultiplier, correlates with the size, shape and optical homogeneity of the particle. The assessment of forward scatter is, however, influenced by the angles over which the light is collected, by shadows and by the refractive index between the cells and the medium.

Side scatter, detected by a photomultiplier, assesses internal morphology and tends to be more sensitive to inclusions within cells than forward scatter. Side scatter is used to distinguish granulated from non-granulated cells.

Fluorescence is always detected by a photomultiplier.

In the jet-in-air system the optical pathway is designed with an orthogonal geometry and comprises two main sections: the interrogation table and the detection table connected through an objective lens (figure 2.7).

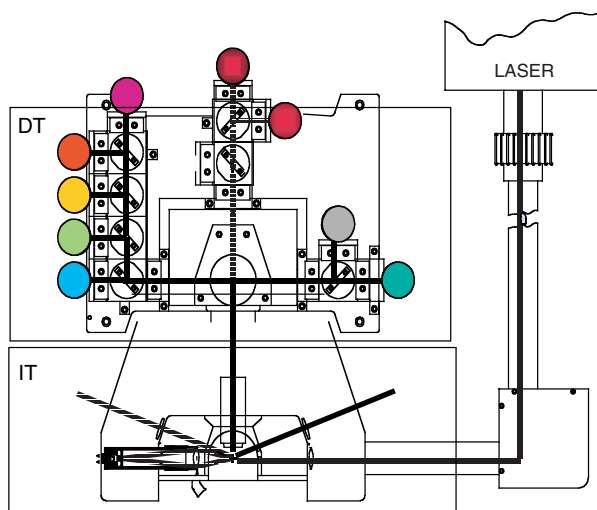


Figure 2.7 Cytoation MoFlo® interrogation table (IT) and detection table (DT), with three laser pathways (Cytoation, Inc., Fort Collins, Colorado, USA).

2.2.5.4.2 Interrogation table

An interrogation table is where the particle intersects the exciting light source, usually a 488 nanometres laser beam (for some instrument a mercury-arc lamp is used instead). The light scattered is collected through a lens in front of the forward scatter detector, and through an objective at right angles to the laser beam and the stream.

2.2.5.4.3 Detection table

The light collected through the 90° objective is redirected with mirrors or prisms, then divided into several bandwidths. Dichroic mirrors, either long-pass or short-pass, positioned at 45° along the light pathway, split the light at a precise wavelength into two bandwidths (figure 2.8). Bandpass filters either transmit or reject a bandwidth and they are usually positioned in front of a photomultiplier. The combinations of filters and dichroic mirrors divide light into precise bandwidths. Only the emitted light centred on the emission maximum of each fluorochrome is detected by the specific photomultiplier. The combination also reduces spectral overlap of one fluorochrome into the bandpass filter of another (see compensation and figure 2.10).

2.2.5.4.4 Electronics and signal processing

Photons travelling into photomultipliers are transformed into electrical current, which, after a preamplification step, is displayed as pulse with a voltage between zero and ten volts (Carter & Meyer, 1995).

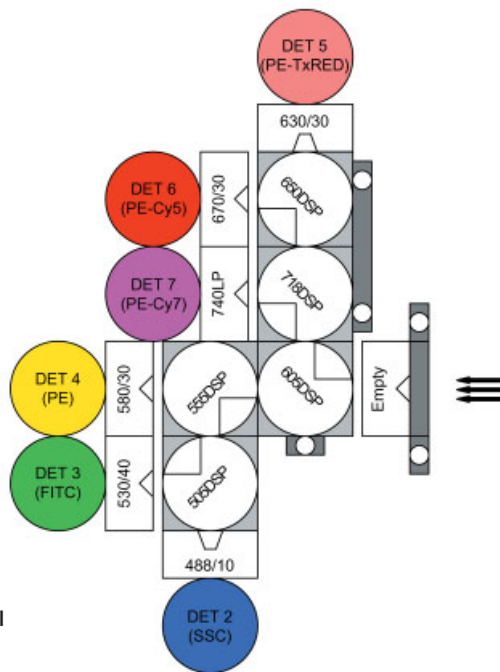


Figure 2.8 CytoFACS MoFlo® z-con-figuration, with filters and dichroic mirrors for the 488 nanometres laser line (CytoFACS, Inc.). The following abbreviations are used in this figure: DET, detection table; DSP, digital signal processor; FITC, fluorescein; PE, phycoerythrin; TxRED, Texas Red.

The amplitude and the shape of the pulse are proportional to the number of photons collected, the rate of the sample and the size of the particle.

An increase of the differential pressure, between the sample and the sheath increases the rate but produces narrower pulses and lower amplitude. It is therefore important that the design of the preamplifier take into consideration emission from particles with different sizes at different pressures.

Since the preamplifier is a source of electronic 'noise' a threshold is always present. Therefore, only signals above the preset threshold value are processed further. The electronics require a trigger signal in order to start the acquisition process. In most cases, there is a single trigger linked to the forward scatter. However, it is also possible to use multiple signals, with multiple thresholds and to combine these using Boolean operations. A common multi-trigger application in flow cytometry is marine biology (Andreatta *et al.*, 2001).

The pulses outputs from the preamplifier are set so as to be proportional to the height, the area or the width of the pulse. The combination of area and height, for example, is useful for doublet discrimination (Ormerod, 1995).

2.2.5.4.5 Dead-time

The so-called dead-time is the time required to electronically process and discriminate the signal between two different inputs. Today, the shortest dead-time is 5.5 microseconds. Two events falling within a 5.5 microseconds window will be classified as a coincidence and will not be processed further. Though a long dead-time is not serious concern in the design of an analyser due to the low acquisition rate, high-speed cell sorters need to operate with the shortest possible dead-time in order to produce an acceptable yield.

2.2.5.4.6 Linear circuits and logarithmic amplifiers

Cytometers are equipped with linear circuits and logarithmic amplifiers. Linear circuitry is used mainly for fluorescence compensation. Logarithmic amplification allows the contemporaneous display on the same axis of very dim and very intense signals without losing the resolution. Modern instruments work on a four-decades dynamic range and it is possible to resolve on a 1,024-channel scale a 1,000-fold difference in fluorescence (see figure 2.9).

2.2.5.5 Compensation

To evaluate properly the emission spectra of two or more fluorochromes together with the associated bandpass filters, the Scripps Research Institute (San Diego, California, USA) offers a useful tool, the Spectrum Viewer (<http://facs.scripps.edu/spectra/>).

Figure 2.10 shows the emission spectra for fluorescein (FITC) and phycoerythrin along with the bandpass filters typically used for each fluorochrome. It is common for the emission distribution to have a steep intensity rise on the shorter wavelength side and a gradual tail declining in intensity on the longer wavelength side.

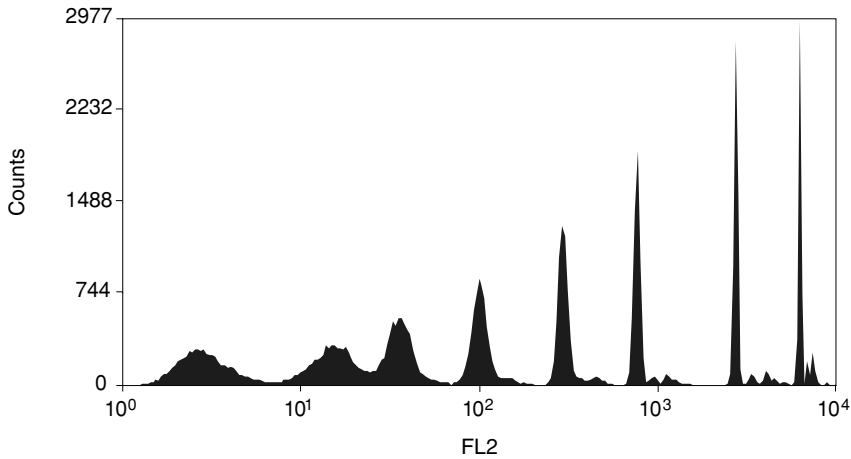


Figure 2.9 Sphero® eight peak ‘Rainbow’ beads (Spherotech, Inc., Libertyville, Illinois, USA), analysed on a Cytomation MoFlo® (Cytomation, Inc.). The distribution of the eight peak intensities in FL2, filter 570/20, over four decades is shown.

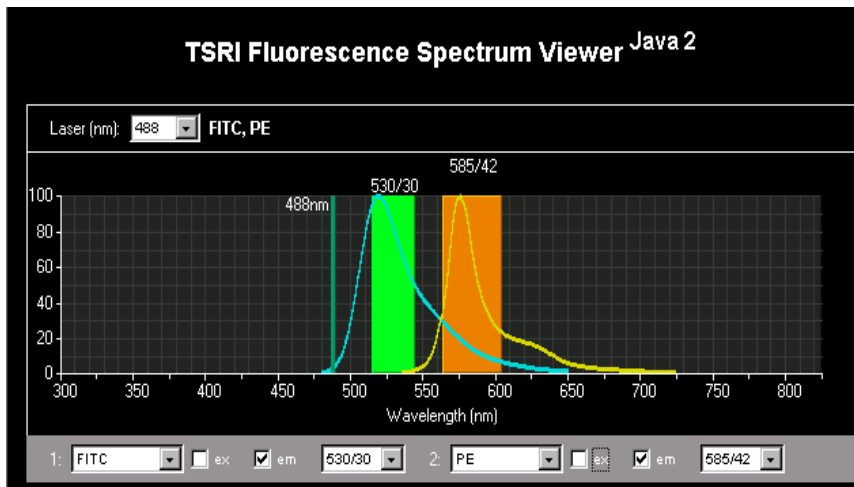


Figure 2.10 Bandpass filters and emission spectra for fluorescein (FITC) and phycoerythrin (PE). The laser line at 488 nanometres is shown as well as the two bandpass filters 530/30 and 585/42 for FITC and phycoerythrin, respectively (TSRI Spectrum Viewer; The Scripps Research Institute, San Diego, California, USA; available at <http://facs.scripps.edu/spectra/>).

Spectral overlap is usually not a big concern in the shorter wavelength. Problems may however occur when emissions from the shorter wavelength fluorochrome (*i.e.*, FITC) spill into the bandpass filter (and photomultiplier) of the longer wavelength one (*i.e.*, phycoerythrin). This spectral overlap produces a spurious increase in the phycoerythrin signal because a small percentage of the total signal in the

phycoerythrin band pass is due to FITC. In order to measure the real value of the phycoerythrin fluorescence, the observed phycoerythrin intensity has to be 'compensated' by subtracting out a percentage of the observed FITC intensity (Roederer, 1999; Stewart & Stewart, 1999).

Digital signal processing allows software correction of the data. After acquisition, original data can be stored together with the corrected data in the same file.

In recent years, interest in the pathologic relevance of small lymphocyte subsets variations have advanced multicolour flow cytometry applications (Baumgarth & Roederer, 2000; Bigos *et al.*, 1999; de Rosa *et al.*, 2001). For these multiparametric applications, which utilise nine to eleven colours, proper compensation is absolutely crucial. New specialised software, such as the FlowJo (Tree Star, Inc., San Carlos, California, USA) has been developed for this purpose.

2.2.5.6 Cell sorting

Cell sorters isolate highly purified populations out of a heterogeneous sample. The populations of interest, identified during the acquisition on the basis of any combination of scatter and fluorescence, are then sorted into separate tubes (figure 2.11).

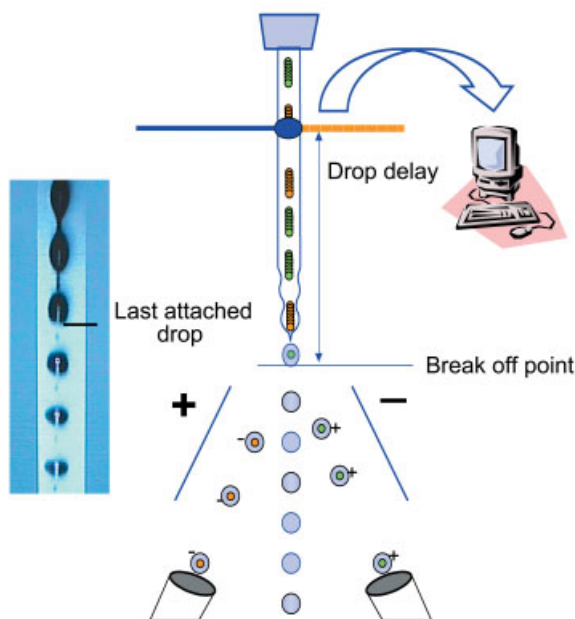


Figure 2.11 Particle sorting in a jet-in-air cell sorter. Following the acquisition and a positive identification by the electronics, the cell of interest reaches the break-off point. Immediately before separation, a charge is applied to the entire stream. The target cell leaves the stream

inside a charged droplet that will be subsequently deflected into separate tubes by means of an electrical field. On the left, the break-off point and droplets are shown as they appear on the droplet camera of the cytometer.

The most common way to sort cells is through the formation of droplets, which are deflected by means of an electrical field. The physics of droplet formation is well-known since the first report by Lord Rayleigh more than a century ago (Rayleigh, 1878). At a given pressure and frequency, any stream emerging through a nozzle will break into droplets at a precise distance from the nozzle (break-off point).

In a cell sorter working at 60 pounds per square inch (4.2 kilograms per square centimetre), with a nozzle of 70 micrometres and a frequency of 100 kilohertz, the stream velocity is about 25 metres per second, and the break-off point is about twelve millimetres from the nozzle. The frequency, obtained with a piezoelectric crystal enclosed into the nozzle assembly, corresponds to the number of droplets generated per second.

The time delay that occurs between the acquisition point (laser intercept) and the break-off point is called the 'drop delay'.

When the target particle, after acquisition and positive identification, reaches the break-off point, the entire stream is charged and the particle leaves the stream inside a charged droplet. The droplet can then be deflected into a tube by means of an electrical field applied between two deflection plates. Empty droplets and all other unwanted particles are discarded.

It is clear that stability of the fluidics and a constant drop delay are the most important factors governing the performance of cell separators. In order to sort only the target particle the stream must be charged at exactly the right delay. Any change in the drop delay, due to pressure variations, air bubbles, modification of the laminar flow or temperature will dramatically affect purity, recovery and yield of the sorted sample.

2.2.5.7 Practical considerations

2.2.5.7.1 Sort logic and drop envelope

When setting up an experiment involving sorting, it is important to take into consideration the definitions given in table 2.2.

Table 2.2 Definitions to be taken in consideration for setting up an experiment involving particle sorting (Corcoran & Lopez).

<i>Term</i>	<i>Definition</i>
Purity (%)	$\frac{\text{Number of particles of interest in the sorted sample}}{\text{Total number of particles in the sorted sample}}$
Recovery (%)	$\frac{\text{Number of particles of interest in the sorted sample}}{\text{Number of particles of interest as indicated on the sorter counter}}$
Yield (%)	$\frac{\text{Number of particles of interest in the sorted sample}}{\text{Number of particles of interest in the original sample}}$

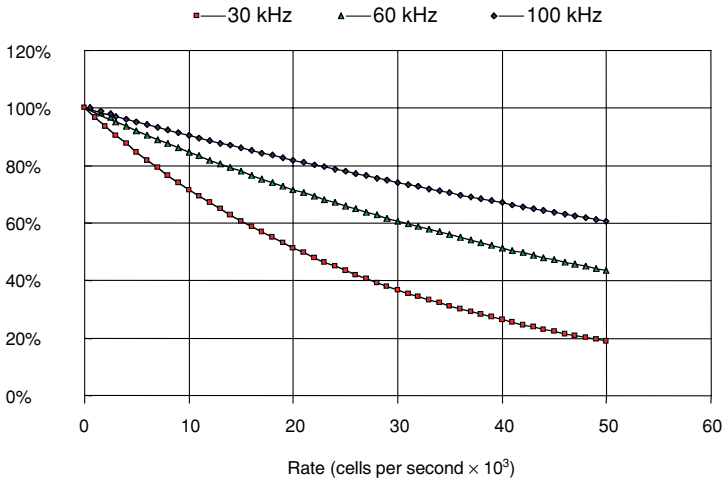


Figure 2.12 Yield and Poisson statistics. With a fixed 5.5 microseconds dead-time the yield decreases progressively with the increase of the rate. At higher rate the electronics will discard more particles, including positive ones, due to the higher number of coincidences. Increasing the frequency, and consequently the number of droplets per second, it is possible to enhance the yield.

During sorting, the probability that two events occur within a 5.5 microseconds time frame (coincidence), is directly proportional to the sorting rate. The electronics, unable to discriminate the two events, will discard the coincidence (hard aborted) wasting not only undesired particles but also potential positive ones, thus compromising the yield.

When yield is the critical factor in the isolation of very rare populations, purity will be sacrificed in a way that most of the particles of interest will be sorted, as well as some undesired ones. In statistical terms, the yield follows the Poisson distribution of rare events. At a given rate and dead-time, an increase of the frequency, in other words generating more droplets, will also increase the yield (figure 2.12).

When purity is the decisive factor, *i.e.* single cell isolation, both recovery and yield will be penalised. Anytime there is a probability to sort contaminants the electronics will discard the event, increasing the total abort rate.

Therefore, cell sorters are equipped with different algorithms, also called 'sorting logic' or 'abort logic', selectable by the operator in order to maximise the target particles. Beside the sort logic it is also possible to sort more than one droplet per unit time. The sorting envelope can be selected in a way that one, two or three droplets are charged which increases the probability of sorting the target particle.

The right combination of sort logic and drop envelope depends on the application (see table 2.3).

Table 2.3 MoFlo® (Cytomation, Inc., Fort Collins, Colorado, USA) sorting logic as an example.

Sorting logic	Purity	Recovery	Yield	Abort
Single	•••	•	•	•••
Purify	••	••	••	••
Enrich	•	•••	•••	•

2.2.5.7.2 Cell size and nozzle tip

As a general rule the nozzle tip installed should be three to five times bigger than the average particle size. This choice directly influences the sorting rate. Several nozzles are available, ranging from 50 to 400 micrometers, with pressures ranging from 80 pounds per square inch (5.6 kilograms per square centimetre) down to ten pounds to square inch (0.7 kilograms per square centimetre), respectively.

2.2.5.7.3 Sample concentration

Concentration is critical for high-speed sorting. In order to achieve a 'single cell line' in front of the laser beam, concentrations of up to 1×10^8 particles per millilitre are required. Unfortunately, only particles with a diameter of less than 25 micrometres can be processed due to physical limitations in the packing geometry.

2.2.5.8 Discussion

New and traditional applications in the field of immunology, cell biology and genetic engineering can be served by flow cytometry. State-of-the-art analysers support eleven parameters, nine colours, three laser configurations, backed up by digital data acquisition and compensation and acquisition rates exceeding 50,000 cells per second.

High-speed cell sorters can sort four different populations, with independent sorting logic, at a rate up to 100,000 cells per second, in a completely automatic fashion.

Nowadays, analysers and sorters allow prolonged walk-away procedures that run in fully automatic fashion for several hours, improving at the same time purity, recovery and yield, rescuing samples and alerting the operator in case of malfunctions.

In the last years, the expansion of biotechnology companies enhanced the need of screening thousands of samples per day, in multi-well plates. Modern instruments can sort cells on any plate, from the microscope slide up the 1,536 multi-well plate, with extreme precision and purity. Such systems coupled with a robotic tray mover permit high-throughput screening, collecting up to 32 parameters for every single event in full automatic mode (Ashcroft & Lopez, 2000).

2.2.6

Tissue microdissection techniques**2.2.6.1 Introduction**

Modern biological methods such as DNA microarray techniques for RNA analysis and mass spectrometric investigations of proteins require a high amount of pure samples, which must be processed quickly after harvesting.

In former times scraping off coarse pieces of histological material using a needle or a glass pipette was the only way to obtain selected sections from a histological section mounted onto a glass slide. Even when performed under a microscope, this method did not allow adequate isolation of cell areas of interest.

This changed with the advent of methods that use the principle of light amplification by stimulated emission of radiation (laser) for pure sample capture and in particular for tissue microdissection. The method to generate monochromatic light using the laser principle was invented in 1960 by Theodore H. Maiman (Maiman, 1960).

2.2.6.2 The force of focused light

A unique manipulating tool was created when a laser was coupled through the epifluorescence path into a research microscope and focused through the objective lens. The narrow spot of focused laser light generates a physical force that allows dissection of microscopically small objects without any mechanical contact (figure 2.13).

The laser-based microdissection of single cells was first used by Meier-Ruge *et al.* (1976). Since then, different approaches have been developed to retrieve pure samples from various sources. Although physical contact with the specimen should be avoided during dissection of the sample, this is not the case for most methods of specimen retrieval (*e.g.*, needle extraction or application of adhesive transfer caps after laser microdissection). These retrieval methods were all associated with a risk of contamination with neighbouring or non-selected material (Bonner *et al.*, 1997; Zhuang *et al.*, 1995).

The principle of laser cutting with a pulsed ultraviolet A laser is a phenomenon called 'ablative photo-decomposition' which is a photochemical process without measurable heat formation in the selected sample and with no collateral damage to the adjacent material (Srinivasan, 1986). This technique allows one to safely cre-

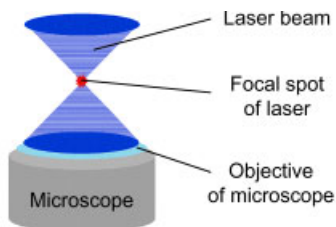


Figure 2.13 Principle of laser cutting.

A tool for tissue microdissection is created, when a laser is coupled through the epifluorescence path into a research microscope. Single laser pulses are focused through an objective lens. Within the narrow laser focal spot, the focused light generates forces that allow microdissection of microscopically small objects without any mechanical contact.

ate incisions, perform cell fusions, ablate cells or microdissect cells, subcellular organelles and other microparticles.

Most of the known laser microdissection systems work with a nitrogen laser, which produces radiation at a wavelength of 337 nanometres in the range of ultraviolet light. There is the general perception that ultraviolet light damages DNA. However, the ultraviolet region of the spectrum is quite broad. The 337 nanometres nitrogen laser works within the ultraviolet A range, where no damage to biological matter occurs due to absorption. The absorption wavelengths of 260 nanometres for DNA and 280 nanometres for proteins are sufficiently distant from the applied 337 nanometres laser wavelength. There is no measurable heat formation and the applied laser wavelength does not affect the biological information, so that recovery of DNA, RNA or protein is not impaired.

2.2.6.3 LPC – Laser pressure catapulting

A breakthrough in laser-based microdissection and sample recovery came with the invention of laser pressure catapulting (LPC) technology that enables the transfer of the selected cells without any mechanical contact, but solely by the force of focused laser light (Schütze & Lahr, 1998) (figure 2.14).

The power of focused laser light is sufficient to lift selected specimens out of the object plane and to transport them even over distances of several centimetres along the direction of the incident laser light. Like a ball being kicked into goal, the specimens are catapulted directly into an appropriate collection vial, for example into the cap of a standard micro-centrifuge tube (figure 2.15). This procedure has the potential to entirely automate the sampling process.

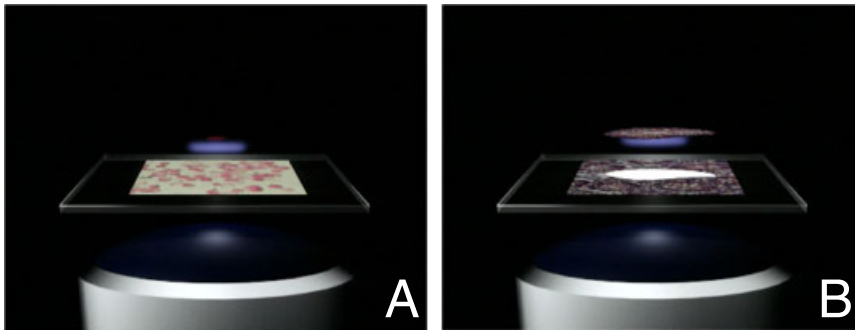


Figure 2.14 Schematic view of the laser pressure catapulting (LPC) procedure. A dissected single cell **A)** or an entire dissected cell area **B)** is lifted from the slide by the action of a single laser shot without mechanical contact. The power of focused laser light is sufficient to lift selected specimens out of the object plane

and to transport them even over distances of several centimetres along the direction of the incident laser light. Like a ball that is kicked into goal the specimens are catapulted directly into an appropriate collection vial, for example, into the cap of a standard micro-centrifuge tube (see also figure 2.15).

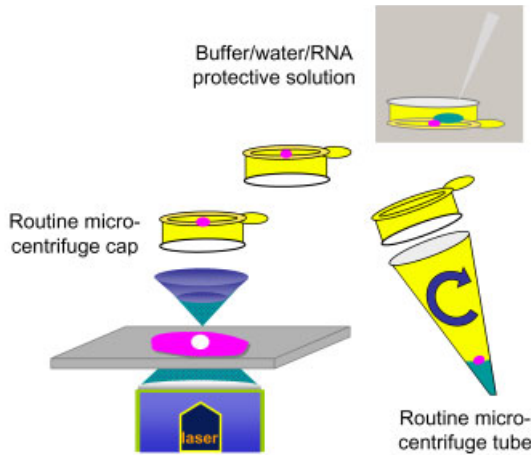


Figure 2.15 Schematic view of laser pressure catapulting (LPC) sample collection. The specimen is catapulted into a micro-centrifuge cap that contains some liquid. By simply spinning down the sample into the tube the required analyses can be performed immediately without any step in between.

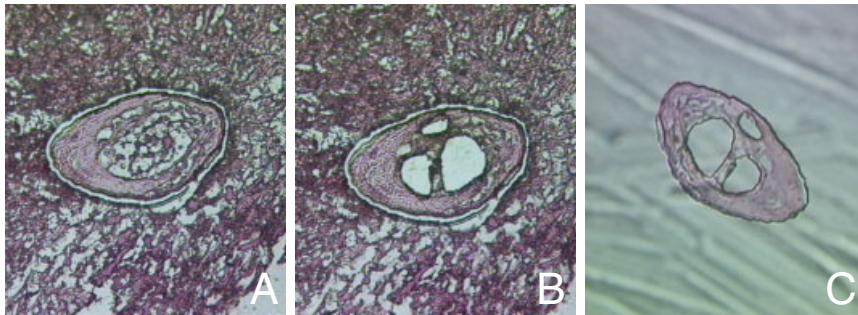


Figure 2.16 Laser pressure catapulting (LPC) for pure and homogeneous sample preparation. **A)** A clear gap is cut around the selected area to avoid contamination with neighbouring tissue. **B)** Ablation of any unwanted material within the selected cell area enables homogeneous preparations. **C)** Only the selected area is found in the cap after dissection.

Unwanted cell contamination dramatically reduces the detection level of genetic alterations. For example, if a normal cell of a tumor section is mixed with a tumor cell carrying a heterozygous gene mutation, the gene mutation signal is reduced to 25 percent. Pure target cell samples are therefore very important for molecular genetic analysis.

However, tissue preparations are usually inhomogeneous and consist of a mixture of different cells. The ultraviolet laser cuts a clear gap between selected and non-selected regions (figure 2.16.A). Thus, sample recovery can be performed without contamination by non-selected material. Furthermore, unwanted specimen within larger areas can selectively be destroyed by laser ablation (figure 2.16.B). Due to the fact that the catapulting process is performed against gravity, only the selected samples will be found in the cap (figure 2.16.C) without any contam-

ination by surrounding material. This guarantees contamination-free isolation of morphologically defined pure cell populations.

2.2.6.3.1 Laser settings and software

Laser settings and the mode of laser operations are dependent on the selected sample. For example, with the PALM® MicroBeam laser system (P. A. L. M. Microlaser Technologies AG, Bernried, Germany) the optimal parameters are conveniently adjusted on the computer screen and may be stored in personal files for laser applications with similar sample types.

Different cells or cell areas are outlined on the screen and marked with different colours using computer graphic tools. It is possible to collect several areas within distinct groups (figure 2.17.A). The grouped specimen (figure 2.17.B) are displayed within a list of elements and can be highlighted. The selected laser action as well as immediate catapulting is automatically performed on a simple mouse click (figure 2.17.C and 2.17.D).

If desired, all operations may be carried out without looking into the microscope, *i. e.* using only the computer screen for sample selection and starting the laser procedures.

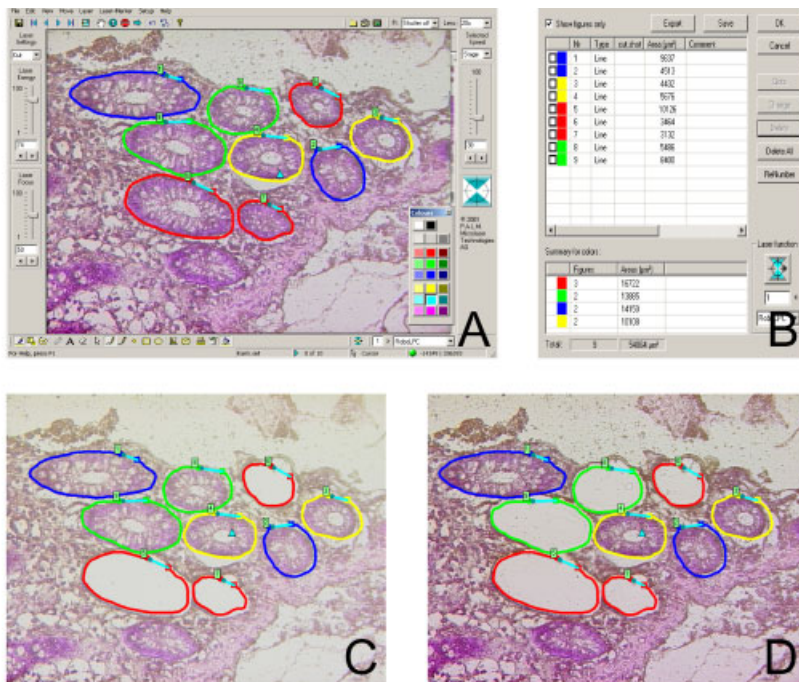


Figure 2.17 Laser pressure catapulting (LPC) elements showing selected groups to be microdissected. A) Outlining and pre-selection of desired cell areas using colour-coding and computer graphic tools. **B)** List of elements showing selected groups to be microdissected. **C)** Section after catapulting the red marked areas. **D)** Section after catapulting the red and green marked areas.

2.2.6.3.2 What kind of specimen can be used?

Any kind of tissue from different sources, human, animal, plant, bacteria and even archival glass mounted tissue sections can be captured using this laser method. Paraffin sections, frozen sections, blood smears, cytopins, stained and immunolabelled specimens can be cut and catapulted.

So far, no restrictions have been reported regarding the origin of the selected specimen or the applied preparation and staining procedures. Samples of any shape and size from one micrometre up to one millimetre in diameter have been captured. Even tumour cells retained on a 20 micrometres thick Teflon filter routinely used for blood filtration have successfully been collected (Vona *et al.*, 2000).

It is possible to handle even wet samples, as is proven by successful live cell catapulting (Mayer *et al.*, 2002). Single cells can be cloned after catapulting and may be used for microarray techniques.

Due to the precision of laser pressure catapulting systems, small cells, single nuclei, chromosomes, chromosomal parts and even filaments can be microdissected and catapulted. Another possibility is to discriminate between the nuclei and the cytoplasm of selected cells. Collection of nuclei in one vial and cytoplasm in another vial allows study, for example, of the kinetics of drug diffusion or virus infection within single cells and also allows investigation of differential expression of protein or mRNA.

2.2.6.3.3 Specimen collection, recovery and investigation

The catapulted specimens are collected within a micro-centrifuge cap or any other appropriate collection device placed above the objective. The cap may contain a small droplet of water, buffer (*e.g.*, PCR buffer) or a RNA protective solution to preserve RNA from degradation. It also may be covered with a fine film of routine PCR (lightweight mineral) oil. After catapulting, the samples are easily spun down into the tip of the tube for direct processing (Schütze *et al.*, 2002) (see figure 2.15).

Cells isolated by this method have been characterised by a wide variety of qualitative and quantitative molecular assays. Single cell PCR (Hahn *et al.*, 2000) and RT-PCR are best performed in catapulted specimens. The recently introduced real-time PCR technology allows the reliable quantification of very small amounts of nucleic acids. These techniques have been successfully applied to the quantification of DNA and RNA isolated from microdissected tissue sections or single cells. The exact analysis of quantitative changes of nucleic acids during the course of pathological alterations has thus become possible.

RT-PCR of archival sections has been used to analyse gene expression in a single cell (Lahr, 2000). The successful combination of laser microdissection with microarrays has also been demonstrated (Becker *et al.*, 2001; Nilsson *et al.*; Schütze *et al.*, 2002). Gene expression can therefore now be related to morphology which will give new insights into cell growth and differentiation but also into the development of cancer and other diseases.

2.2.6.4 μ CUT – Laser microdissection ultraviolet cut

Laser microdissection ultraviolet cutting (μ CUT; SL Microtest GmbH/MMI AG, Zürich-Glattbrugg, Switzerland) is performed with another type of ultraviolet laser as in the LPC system; sample collection is performed in a contact-mode by using needles or by manually placing the lid of a tube filled with adhesive material onto the specimen. The specimen is mounted on a special transfer membrane. The laser then cuts both the tissue and the membrane. The dissected sample adheres to the adhesive cap and the user has to remove the lid (figure 2.18.A).

2.2.6.5 Laser scissors

The laser scissors method (Cell Robotics International, Inc., Albuquerque, New Mexico, USA) also uses an ultraviolet laser. The samples are collected by contact mode with a special cap or a needle (see figure 2.18.B). The stage is motorised, but can only move in slow motion. Special software for automatic laser capture has not yet been reported.

2.2.6.6 LMD – Laser microdissection

Like the LPC system, the laser microdissection (LMD) technique (Leica Microsystems AG, Wetzlar, Germany) uses an ultraviolet A laser. However, the arrangement of laser and microscope is different. LMD is performed using an upright microscope. The section is placed into the slide holder in an upside down direction. The region of interest is cut out by laser activation and subsequently transferred into the inverted cap that is mounted below the specimen (figure 2.19).

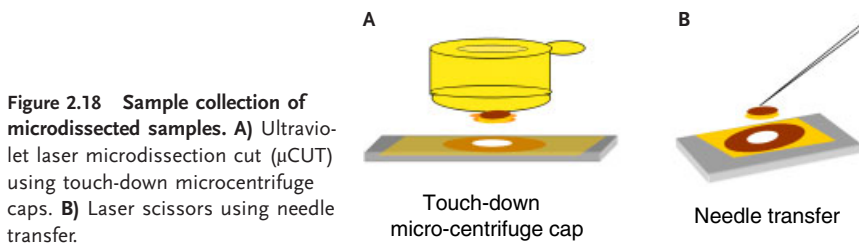


Figure 2.18 Sample collection of microdissected samples. A) Ultraviolet laser microdissection cut (μ CUT) using touch-down microcentrifuge caps. **B)** Laser scissors using needle transfer.

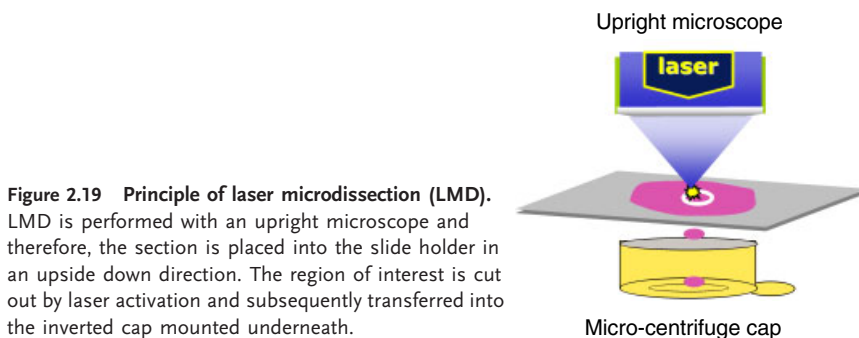


Figure 2.19 Principle of laser microdissection (LMD). LMD is performed with an upright microscope and therefore, the section is placed into the slide holder in an upside down direction. The region of interest is cut out by laser activation and subsequently transferred into the inverted cap mounted underneath.

Also unlike LPC technology, the laser scans while the stage remains fixed. Thus the working area is restricted to the actual displayed screen and pre-selection of specimen throughout the rest of the slide is not possible.

Specimen for the LMD technique are mounted onto a special membrane spanned between a plastic frame. The degree of adhesion to conventional glass slides is too great to allow the sample to fall into the collection cap.

2.2.6.7 LCM – Laser capture microdissection

The laser capture microdissection (LCM) technology (Arcturus Engineering, Inc., Mountain View, California, USA) is a widely used method (Fend *et al.*, 1999; Gordon *et al.*, 2000; Luo *et al.*, 1999; Luzzi *et al.*, 2000; Wong *et al.*, 2000) and a complete different technology compared to the other methods described above. The laser used is a continuous wave near-infrared laser. A special cap spanned with a transfer membrane is brought into contact with the tissue slice. The laser locally heats the sample and activates the heat sensitive adhesive coating which melts and adheres to a section of the specimen. On retraction of the cap, the sections of the specimen that have adhered to the cap are extracted. With each laser shot, material can be melted in a diameter of 7.5, 15 or 30 microns (figure 2.20). As the separation of the target cells is performed by adhesion, there is a risk of contamination with unwanted material (Heel & Dawkins, 2001; Vasmatazis *et al.*, 1998).

2.2.6.8 Ultrasonically oscillating needle

The ultrasonically oscillating needle technique does not employ lasers at all and can be compared to early mechanical needle microdissection methods with the addition of ultrasonic oscillation to facilitate cutting (Harsch *et al.*, 2001). Material capture is performed in liquid by soaking up the dissected material with a glass pipette. The precision of needle based techniques is much less than that of the laser methods, automation is not possible and there is a risk of contamination.

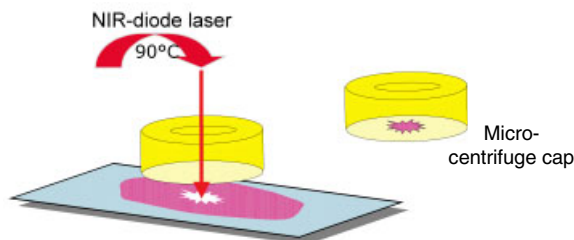


Figure 2.20 Principle of laser capture microdissection (LCM). By activating a near-infrared (NIR) laser the membrane mounted onto a cap adheres to the cell area of interest which is then collected in a micro-centrifuge cap.

2.2.6.9 Manual tissue dissection

Manual tissue dissection is the oldest and most inefficient way of isolation pathological samples from tissues and can be performed on routinely stained microscope slides using sections of five to fifteen micrometres on non-coated glass slides. Manual tissue dissection requires histologically homogeneous material and the areas to be dissected should have a diameter of at least one millimetre (Whetsell *et al.*, 1992). Using a sterile needle or a scalpel the selected sections can be procured (Perren *et al.*, 1998).

Table 2.4 gives an overview of the laser microdissection techniques mentioned above.

2.2.6.10 Discussion

Laser microdissection is established as a valuable method in a wide area of research. The non-contact procedures have become popular because of their accuracy and reliability in capturing specific target cells without contaminating material. There are some important criteria for the most suitable technique: it should be an entirely non-contact procedure, that has no restrictions regarding preparation methods and should allow for a high degree of automation. The combined micro-preparation techniques of laser microdissection and laser pressure catapulting allow reliable access to large structures in complex tissues, but also to single cells or subcellular fractions. Thousands of cells may be collected in a very short time with high precision and purity, as required for DNA microarray and proteomic evaluations. Microdissection combined with molecular methods like PCR, RT-PCR, real-time RT-PCR or DNA microarray techniques allows correlation of genetic and proteomic alterations in morphologically defined cell populations.

2.2.7

Miscellaneous cell isolation and culturing techniques

Several methods have been developed to prepare special cultures of cells derived from human or animal tissues. These methods have been described in detail in primary and secondary literature. For example, detailed protocols for creating cell cultures from heart muscle, endothelium and aorta have been described in Piper *et al.* (1990). Common cell culture techniques have been published in, for example, Boulton *et al.* (1992), Clynes (1998), Doyle *et al.* (1997) and Harrison & Rae (1997), whereas the primary literature comprises several hundreds of methods for cell isolation and culturing.

Table 2.4 A comparison of different laser microdissection systems.

	Laser pressure catalypuling (LPC)	Laser microdissection ultraviolet cut (μ CUT)	Laser scissors	Laser microdissection (LMD)	Laser capture microdissection (LCM)
Company	PALM Microlaser Technologies AG (Bernried, Germany)	SL Microtest GmbH/MMI AG (Zürich-Glatbrugg, Switzerland)	Cell Robotics International, Inc. (Albuquerque, New Mexico, USA)	Leica Microsystems AG (Wetzlar, Germany)	Arcturus Engineering, Inc. (Mountain View, California, USA)
Principle	• Non-contact procedure • No heat formation in the selected sample	• Contact procedure • No heat formation in the selected sample	• Contact procedure • No heat formation in the selected sample	• Non-contact procedure • No heat formation in the selected sample	• Contact procedure • Heat-activated melting of sample with contacting adhesive cap
Laser source	• Pulsed nitrogen ultra-violet A laser (wave-length 337 nanometres) coupled into the micro-scope <i>via</i> the epifluor-escence path. • Focussing of the laser through the objective.	• Ultraviolet laser coupled into the micro-scope <i>via</i> the epifluor-escence path. • Focussing of the laser through the objective.	• Pulsed nitrogen ultra-violet A laser (wave-length 337 nanometres) coupled into the micro-scope <i>via</i> the epifluor-escence path. • Focussing of the laser through the objective.	• Pulsed nitrogen ultra-violet A laser (wave-length 337 nanometres) coupled into the micro-scope <i>via</i> the epifluor-escence path. • Focussing of the laser through the objective.	• Continuous near-infrared diode laser (wavelength 980 nanometres)
Safety of laser wave-length	• No impact on DNA, RNA and protein recovery as the applied laser wavelength is well apart from the absorption maximum of nucleic acids and proteins.	• No impact to DNA, RNA and protein recovery as the applied laser wavelength is well apart from the absorption wavelength of nucleic acids and proteins.	• No impact to DNA, RNA and protein recovery as the applied laser wavelength is well apart from the absorption wavelength of nucleic acids and proteins.	• No impact to DNA, RNA and protein recovery as the applied laser wavelength is well apart from the absorption wavelength of nucleic acids and proteins.	• RNA and protein recovery is problematical due to heating of the sample

Table 2.4 (continuing) A comparison of different laser microdissection systems.

	<i>Laser pressure catalysing (L-PC)</i>	<i>Laser microdissection ultraviolet cut (μCUT)</i>	<i>Laser scissors</i>	<i>Laser microdissection (LMD)</i>	<i>Laser capture microdissection (LCM)</i>
Sample extraction	• Non-contact • Laser pressure catalysing	• Contact • Mechanical tools (needle or adhesive microcentrifuge cap)	• Contact • Mechanical tools (needle or adhesive microcentrifuge cap)	• Non-contact • Falling down after cutting by gravity	• Contact • Heat-activated melting of sample with contacting adhesive cap
Software	• Application specific software • Colour pre-selection of wanted cells • ‘List of elements’, grouped specimen • Area calculations • Photographic archiving system incorporated • Automatic memory function • Saving of pre-selected specimen for later microdissection	• Application specific software • Colour pre-selection of wanted cells • Automatic photographic archiving • Automatic memory function	• Industry standard software control, no microdissection-specific software	• Software for laser scanning control and image administration	• No microdissection specific software • Software predominantly designed for photographic records

Table 2.4 (continuing) A comparison of different laser microdissection systems.

	<i>Laser pressure catapulting (LPC)</i>	<i>Laser microdissection ultraviolet cut (μCUT)</i>	<i>Laser scissors</i>	<i>Laser microdissection (LMD)</i>	<i>Laser capture microdissection (LCM)</i>
Automation	• Motorised, computer-controlled microscope stage • Automatic cutting and catapulting features available in combination with computer graphic tools • Quick access to thousands of cells due to motorised stage-specific software and non-contact collection	• Motorised, computer-controlled microscope stage • No automatic procedures available	• Slow motion microscope stage (piezo-driven) • No automation possible	• Manual-driven microscope stage	• Manual stage control • No automatic procedures available
Object slides	• Routine glass object slides • Super-frosted or charged slides • Membrane mounted slides • Tissue culture vessels	• Membrane slides only	• Membrane slides only	• Membrane slides only	• Routine glass object slides • Not suitable for super-frosted or charged slides

Table 2.4 (continuing) A comparison of different laser microdissection systems.

	Laser pressure catalysing (LPC)	Laser microdissection ultraviolet cut (μCUT)	Laser scissors	Laser microdissection (LMD)	Laser capture microdissection (LCM)
Specimen	• Paraffin-embedded tissue • Frozen tissue sections • Cyto-centrifuged samples • Cell smears • Cells on filters • Chromosomes • Living cells • Wet and dry samples can be processed	• Paraffin-embedded tissue • Frozen tissue sections • Cyto-centrifuged samples • Cell smears • Chromosomes	• Paraffin-embedded tissue • Frozen tissue sections • Cyto-centrifuged samples • Cell smears	• Paraffin-embedded tissue • Frozen tissue sections • Cyto-centrifuged samples • Cell smears • Chromosomes	• Paraffin-embedded tissue • Frozen tissue sections • Cyto-centrifuged samples • Cell smears
Specimen shape and sample size	• Single cells or cell populations in complex areas • Tissue areas from one micrometre to one millimetre • Sub-cellular particles • Cell nuclei • Chromosomes • Fragments of chromosomes • Filaments	• Single cells or cell populations in complex areas • Tissue areas from one micrometre to one millimetre • Sub-cellular particles • Cell nuclei • Chromosomes • Fragments of chromosomes • Filaments	• Single cells or cell populations in complex areas • Tissue areas from one micrometre to one millimetre • Sub-cellular particles • Cell nuclei • Chromosomes • Fragments of chromosomes • Filaments	• Single cells or cell populations in complex areas • Tissue areas from one micrometre to one millimetre • Sub-cellular particles • Cell nuclei • Chromosomes • Fragments of chromosomes • Filaments	• Primarily cells equivalent to laser spot size • Suitable for single cell capture of large or spatially isolated cells only • Complex areas captured by multiple overlapping laser spots

Table 2.4 (continuing) A comparison of different laser microdissection systems.

	<i>Laser pressure catapulting (LPC)</i>	<i>Laser microdissection ultraviolet cut (μCUT)</i>	<i>Laser scissors</i>	<i>Laser microdissection (LMD)</i>	<i>Laser capture microdissection (LCM)</i>
Sample purity	• No contamination with unwanted material • Selective destruction of unwanted cells • Clear-cut separation from neighbouring tissue • Direct catapulting without any contact • Only the selected specimens fly against gravity into the collecting cap	• Danger of contamination due to mechanical contact • Selective destruction of unwanted cells • Clear-cut separation from neighbouring tissue	• Danger of contamination due to mechanical contact • Selective destruction of unwanted cells • Clear-cut separation from neighbouring tissue	• Danger of contamination as loose debris may fall down into the collection cap	• Selective destruction of unwanted cells not possible • No clear cut separation from neighbouring tissue
Versatility	• Non contact • Live-cell catapulting • Live-cell microsurgery • Live-cell microinjection • Live-cell fusions	• Live cell microsurgery • Live-cell microinjection • Live-cell fusions	• Live cell microsurgery • Live-cell microinjection • Live-cell fusions	• Live cell microsurgery • Live cell microinjection • Live-cell fusions	• No live-cell treatment possible; captures cells from frozen or fixed tissue samples only

2.3

Preparation of nucleic acids and proteins

2.3.1

Introduction

The analysis of gene expression is based on the investigation of ribonucleic acids and proteins which in turn requires the accurate and reliable isolation of these molecules. The isolated sample should represent the relative amounts of RNAs or proteins present in the tissue or in the cells at the moment of harvesting. The samples must be pure, of high-quality and undigested. All chemicals and solutions used must be free of RNases and proteases.

The following sections are devoted to the isolation of eukaryotic RNA. The analysis of prokaryotic RNA has its own characteristics and difficulties. Prokaryotic RNAs have no 5'-cap, rarely have poly(A) tails, and are highly unstable with an average half-life of about three minutes. Some bacterial mRNAs begin to degrade while they are still being translated. Therefore, special methods have been developed for the isolation of prokaryotic RNAs and the analysis of gene expression in prokaryotes, but these are beyond the scope of this book. The interested reader is referred to the primary literature (*e.g.*, Amara & Satchidanandam, 1997 and 2000; Coombes & Mahony, 2001; Fislage *et al.*, 1997; Handfield & Levesque, 1999; Loos *et al.*, 2001; Rivera-Marrero *et al.*, 1998; Wendisch *et al.*, 2001).

In isolating RNA and proteins from plant tissues, special attention must be paid to effectively grinding up the tissue (Gilman, 1994). Plant tissues are frequently fibrous and contain organic compounds that can make fresh tissue difficult to break up. Therefore, special protocols for the isolation of plant RNAs have also been developed that are not described in detail here. In this case also, we refer the interested reader to the specialist literature (*e.g.*, Bijli *et al.*, 2001; Cheng & Seemann, 1998; Dunn, 1997; Faggioli *et al.*, 1998; Suzuki *et al.*, 2000; Ye *et al.*, 1998).

For detailed reading on the isolation of ribonucleic acids and proteins, in particular isolation from blood, tissue, cell organelles, or single-procedure isolations of RNA and DNA, the reader is referred to the specialist literature (*e.g.*, Brown, 1996; Bynum & Ronzio, 1976; Carpenter & Simon, 1998; Coligan, 1995; Dennison, 1999; Farrell, 1998; Kaplan *et al.*, 1979; Krieg, 1996; Macfarlane & Dahle, 1998; Mrazek & Petrek, 1999; Quadroni & James, 1999; Rapley, 1998; Rothstein, 1994; Scopes & Smith, 1994; Stern & Newton, 1986).

2.3.2

Isolation of total RNA and mRNA

2.3.2.1 Introduction

While the genomic or mitochondrial genes of an organism are, to all intents and purposes, fixed, the mRNA population represents how these genes are expressed under any given set of conditions. Analysis of RNA by a number of technologies, including northern blotting and DNA microarray analysis, or by a variety of RT-

PCR-based methods, can provide a good reflection of an organism's gene expression profile.

This chapter deals with appropriate procedures for purification and analysis of RNA. General considerations for isolation of RNA are first described, followed by considerations for disruption and homogenisation of starting samples, and different RNA isolation technologies.

2.3.2.2 General considerations for isolation of RNA

2.3.2.2.1 Stabilisation of RNA in biological samples

In order to ensure accurate gene expression analyses, it is important that the RNA analysed truly represents the *in vivo* gene expression of the sample. This is complicated by the fact that changes can occur during handling of the sample and isolation of the RNA.

These changes, including down-regulation of genes and enzymatic degradation of RNA, can occur very rapidly, during sample collection and handling and before sample lysis can inhibit further changes in the RNA expression pattern. Immediate stabilisation before beginning RNA isolation is therefore a prerequisite for accurate gene-expression analysis.

2.3.2.2.2 Ribonucleases

Ribonucleases (RNases) are very stable and active enzymes that generally do not require cofactors to function. Since RNases are difficult to inactivate and since even minute amounts are sufficient to destroy RNA, use of any plasticware or glassware without first eliminating possible RNase contamination is recommended. Great care should be taken to avoid inadvertently introducing RNases into the RNA sample during or after the isolation procedure. In order to create and maintain an RNase-free environment, the following precautions must be taken during pre-treatment and use of disposable and non-disposable vessels and solutions while working with RNA.

2.3.2.2.3 General handling

Proper microbiological, aseptic technique should always be used when working with RNA. Hands and dust particles may carry bacteria and molds, which are the most common sources of RNase contamination. Wear of latex or vinyl gloves while handling reagents and RNA samples to prevent RNase contamination from the surface of the skin or from dusty laboratory equipment is recommended. Gloves should be changed frequently and tubes should be kept closed whenever possible. Isolated RNA should be kept on ice during pipetting of aliquots for downstream applications.

2.3.2.2.4 Use of plasticware

The use of sterile, disposable polypropylene tubes is recommended throughout all procedures. These tubes are generally RNase-free and do not require pre-treatment to inactivate RNases.

Non-disposable plasticware should be treated before use to ensure that it is RNase-free. Plasticware should be thoroughly rinsed with solution of 0.1 molar sodium hydroxide solution, one millimolar ethylene diaminetetraacetic acid (EDTA) solution followed by RNase-free water. Alternatively, chloroform-resistant plasticware can be rinsed with chloroform to inactivate RNases.

2.3.2.2.5 Glassware

Glassware should be treated before use to ensure that it is RNase-free. Glassware used for RNA work should be cleaned with a detergent, thoroughly rinsed, and oven baked at 240 °C for four or more hours before use. Autoclaving alone will not fully inactivate many RNases. Alternatively, glassware can be treated with diethyl pyrocarbonate (DEPC), which covalently modifies and inhibits RNases (Fedorcsak & Ehrenberg, 1966).

2.3.2.2.6 Solutions

Water and other aqueous solutions should be treated with DEPC and afterwards autoclaved. In addition, RNase inhibitors can be used where it is not practical to use DEPC (see next section).

2.3.2.2.7 RNase inhibitors

Besides DEPC, there are a number of RNase inhibitors that can be used during RNA isolation. The most commonly used are protein RNase inhibitors, 50 kilodalton proteins that strongly inhibit RNases A, B, and C, as well as human placental RNases and others. These proteins are competitive inhibitors of RNase, with very strong binding (Blackburn *et al.*, 1977). They are available from several suppliers.

2.3.2.3 Disruption and homogenisation of starting materials for isolation of RNA

Efficient disruption and homogenisation of the starting material is an absolute requirement for all total RNA isolation procedures. This is generally carried out in the presence of an organic solvent or a strong chaotropic agent in order to inhibit endogenous RNases that are released during the process. Nonetheless, changes in the RNA expression pattern may occur prior to or during disruption and homogenisation. For accurate gene expression analysis, the sample should be stabilised first (see following section on RNA stabilisation).

Disruption and homogenisation are two distinct steps.

Disruption: Complete disruption of cell walls and plasma membranes of cells and organelles is absolutely required to release all the RNA contained in the sample and to release and inactivate endogenous RNases. Different samples require different methods to achieve complete disruption. Incomplete disruption results in significantly reduced RNA yields.

Table 2.5 Guide to disruption and homogenisation methods for different samples.

Starting material	Disruption method	Homogenisation method
Cultured animal cells	Addition of lysis buffer	Rotorstator homogeniser, spin-column homogeniser or syringe and needle
Animal tissue	Rotorstator homogeniser	Rotorstator homogeniser
	Mortar and pestle	Spin-column homogeniser or syringe and needle
Bacteria	Bead mill	Bead mill
	Enzymatic (lysozyme) digestion followed by addition of lysis buffer	Vortexing
Yeast	Enzymatic (lyticase/zymolase) digestion of cell wall followed by lysis of spheroplasts by addition of lysis buffer	Vortexing
	Glass beads in a bead mill with lysis buffer	Glass beads in a bead mill with lysis buffer
Plants and filamentous fungi	Mortar and pestle	Spin-column homogeniser

Homogenisation: Homogenisation is necessary to reduce the viscosity of the cell lysates produced by disruption. Homogenisation shears the high-molecular weight genomic DNA and other high-molecular weight cellular components to create a homogeneous lysate. Incomplete homogenisation results in inefficient binding of RNA and therefore significantly reduced yields.

Some disruption methods simultaneously homogenise the sample while others require an additional homogenisation step. Table 2.5 gives an overview of different disruption and homogenisation methods suitable for various starting materials. It can be used as a guide to choose the appropriate method for the starting material with which you are working.

2.3.2.4 RNA isolation technologies

Different technologies and methods are available for isolation of RNA and/or cleanup of RNA from enzymatic reactions. In general, the methods involve disruption and lysis of the starting material (see section above), followed by removal of proteins, DNA, and other contaminants. In this section, different techniques are described. Typical protocols use a combination of several of these techniques. The choice of a particular protocol depends on the type of RNA (total RNA, cytoplasmic RNA, messenger RNA, low-molecular weight RNAs, ribosomal RNAs, *etc.*), the purity needed for downstream applications, the desired time and expense per sample, and whether intact RNA is essential or not.

2.3.2.4.1 Enzymatic digestion

A 'quick-and-dirty' method for isolation of RNA from cells involves cell lysis and digestion of proteins with proteinase K in the presence of sodium dodecylsulphate (SDS) and an RNase inhibitor. After inactivation of the proteinase K, by heating or organic extraction, the DNA is digested with DNase I (Favaloro *et al.*, 1980). Digestion times of 30 minutes to one hour, however, make this method more dirty than quick, and contaminants in the lysate may inhibit downstream applications or lead to RNA degradation. The method cannot be used with tissues since lysis with proteinase K is relatively slow, and it is difficult to prevent endogenous RNases from degrading the RNA.

Digestion with DNase I is often used at the end of a protocol to remove any DNA. Several standard techniques, such as caesium chloride density-gradients, silica-gel methods, and anion-exchange methods, can remove most of the DNA but removal of residual DNA may be required for RT-PCR and other RNA applications that are sensitive to the presence of small amounts of DNA. In such cases, treatment with RNase-free DNase is a good way to remove the remaining DNA. Following DNase treatment, the DNase can be removed by any of the techniques that separate proteins from nucleic acids, including organic extraction followed by alcohol precipitation or cleanup using silica-based or anion-exchange methods. The DNase treatment can also be included as part of the silica-based procedure so that the DNase is removed in subsequent wash steps (see section "Silica-based methods").

2.3.2.4.2 Organic extraction

Organic extraction is a classical technique that is often combined with proteinase K digestion, extraction with strong denaturants, alcohol or lithium chloride precipitation, and/or caesium chloride density-gradients.

Typically, the sample is mixed with phenol at acid pH. The phenol lyses cells and denatures the proteins in the sample (Stallcup & Washington, 1983). At acid pH, DNA in the sample is protonated, neutralising its charge and causing it to partition into the organic phase. RNA, by contrast, remains charged and partitions into the aqueous phase. The two phases are separated by centrifugation, and the aqueous phase is re-extracted with a mixture of phenol and chloroform, and then with chloroform to extract the remaining phenol. RNA in the aqueous phase is then precipitated with ethanol or isopropanol and/or separated on a caesium chloride gradient.

RNA isolated using organic extraction may contain residual phenol and/or chloroform, which can inhibit downstream reactions, such as RT-PCR, and which may also affect absorbance readings (Wilfinger *et al.*, 1997). A further drawback is that these reagents are toxic and teratogenic. In addition, the method is time-consuming and requires a degree of manual skill in the transfer of the aqueous phases.

2.3.2.4.3 Extraction with strong denaturants

Chaotropic agents, such as guanidine isothiocyanate and guanidine hydrochloride, are strong denaturants that immediately inactivate RNases to ensure isolation of intact RNA (Chirgwin *et al.*, 1979). Guanidine salts are also sufficient to disrupt cells. Extraction with chaotropic reagents is typically combined with subsequent organic extraction, alcohol or lithium chloride precipitation, caesium chloride density-gradients, anion-exchange methods, hybrid selection and/or silica-based methods.

2.3.2.4.4 Alcohol and lithium chloride precipitation

Both alcohol and lithium chloride precipitation rely on ‘salting out’ nucleic acids. Precipitation with ethanol or isopropanol, in the presence of sodium or ammonium acetate, is included in many RNA isolation protocols. The RNA must be partially purified before precipitation since proteins and DNA also precipitate. The precipitate is then washed with 70 percent alcohol to remove the remaining salts, dried, and redissolved. This technique allows concentration of the RNA and removal of salts, but it is also time-consuming. In general, this method works best with larger amounts of RNA. Otherwise, the precipitated pellet may be lost when removing the alcohol or during the drying procedure. This can occur with low amounts of RNA, when the pellet is very small, or when residual organic solvents, such as chloroform, partition to the bottom of the centrifuge tube with the RNA pellet.

Lithium chloride can be used to differentially precipitate RNA while DNA remains in solution (Cathala *et al.*, 1983). Lithium chloride precipitation provides higher purity than typical organic extraction methods. However, it typically takes several hours, thus requiring more time than alcohol precipitation. Furthermore, precipitation requires high molar concentrations of lithium chloride, which can interfere with downstream enzymatic reactions, such as RT-PCR. Lithium chloride precipitation is usually followed by alcohol precipitation to remove lithium chloride from the sample.

2.3.2.4.5 Caesium chloride and sucrose density-gradients

RNA can be purified by centrifugation through a caesium chloride density-gradient. The partially purified RNA is mixed with caesium chloride and ethidium bromide and centrifuged at a high *g*-force (*e.g.*, 36,000 to 40,000 × *g*) for twelve to twenty-four hours, typically overnight. RNA pellets to the bottom of the centrifuge tube, while DNA and proteins remain in the supernatant. The RNA pellet is collected and re-precipitated by alcohol precipitation to remove the remaining caesium chloride. This method allows isolation of high-quality RNA, but it is very time-consuming, labour-intensive, and expensive, making it difficult to use for preparation of multiple samples. The use of toxic ethidium bromide and caesium chloride is a further disadvantage.

Small RNA species, such as transfer RNAs (tRNAs) and 5S ribosomal RNAs (rRNAs), do not pellet in caesium chloride gradients. These small RNA species can be isolated by fractionation through sucrose gradients (Schweinfest *et al.*, 1982) or agarose gels (Sambrook *et al.*, 1989) containing methylmercuric hydro-

xide. However, methylmercuric hydroxide is volatile and extremely toxic, making it hazardous to use and difficult to dispose of. As an alternative, small RNA species can be separated using anion-exchange methods (see next section).

2.3.2.4.6 Anion-exchange methods

Solid-phase anion-exchange chromatography is based on the interaction between the negatively charged phosphates of nucleic acid and the positively charged surface molecules of the substrate. RNA binds to the substrate under defined salt conditions, and other contaminants, such as DNA, cellular proteins, and metabolites are eluted with different salt concentrations. This allows purification of RNA, DNA, and low-molecular weight RNA species from the same sample (QIAGEN GmbH, 1998). The eluted nucleic acids are recovered by alcohol precipitation, and are suitable for all downstream applications.

Anion-exchange technology yields RNA of a purity and biological activity equivalent to at least two rounds of purification in caesium chloride gradients, but in much less time. The isolated RNA also includes small RNA species that are excluded when using caesium chloride gradients or silica-based methods, and these small RNA species can be selectively isolated. In addition, this technology completely avoids the use of toxic substances, and can be applied for different throughput requirements as well as for different scales of purification. Some companies offer commercial kits for isolation of RNA and DNA based on anion-exchange technology; kits from different companies differ in their handling speeds and the quality and size of the isolated nucleic acids.

2.3.2.4.7 Hybrid selection for isolation of poly(A)⁺ mRNA

Hybrid selection is the most commonly used technique for isolation of poly(A)⁺ mRNA species. This technique relies on the fact that most eukaryotic mRNA species have a poly(A) tail at their 3'-termini. These mRNA species can then be selected by hybridisation of the poly(A) tail to an oligo(dT) sequence attached to a solid support (Aviv & Leder, 1972). DNA and RNA species that do not have a poly(A) tail are washed away, and the poly(A)⁺ mRNA is eluted by changing the temperature or the salt concentration.

Several companies offer kits for isolation of poly(A)⁺ mRNA, either from total RNA or directly from cells or tissues. Kits from different companies differ in their handling, depending on the solid support used. Different solid supports include columns and magnetic beads or polystyrene-latex beads (QIAGEN GmbH, 1999).

2.3.2.4.8 Silica-based methods

Silica-based technology provides a simple, reliable, fast, and inexpensive method for isolation of high-quality RNA. This method is based on the selective absorption of nucleic acids to silica in the presence of high concentrations of chaotropic salts such as guanidine hydrochloride, guanidine isothiocyanate, sodium iodide, and sodium perchlorate. The use of specific buffers in the lysis procedure ensures that only RNA is adsorbed while DNA, cellular proteins, and metabolites remain in solution. These contaminants are washed away, and high-quality RNA is eluted

from the silica using a low-salt buffer. With use of the appropriate materials or kits, the eluted RNA is ready for use in most downstream applications.

Several companies offer kits for isolation of RNA based on silica technology, which differ in handling and quality of the isolated RNA. For example, silica-gel particles can be used either in suspension or as a membrane in spin columns (QIAGEN GmbH, 2001a and 2001b) or multi-well units designed for high-throughput procedures (QIAGEN GmbH, 2002), including automated procedures. Use of spin columns is generally more convenient and avoids silica particle carryover in the eluted RNA.

Small RNA species, such as tRNAs and 5S rRNAs, are not isolated by silica-gelmembrane technology but can be efficiently isolated using anion-exchange methods.

2.3.3

Stabilisation of RNA prior to isolation

2.3.3.1 Introduction

RNA stabilisation is an absolute prerequisite for reliable gene expression analysis. It is important that the RNA analysed truly represents the *in vivo* gene expression of the sample. In this section, some of the possible methods for stabilisation of RNA in various types of samples are discussed.

2.3.3.2 RNA stabilisation in cells and tissues

Changes in gene expression can occur during sample harvest and handling and during RNA isolation. These changes can occur within seconds after harvesting. Stabilising samples minutes later, for example after centrifuging cell cultures, is often insufficient to prevent artefacts and preserve the gene-expression pattern.

There are two major types of artefacts that can occur. Cell death and enzymatic degradation of RNA result in artificial non-specific and specific reduction of mRNA species. At the same time, specific genes can be induced or down-regulated during handling and processing of the sample. The combination of these two effects can result in a transcription profile that differs considerably from the true *in vivo* gene-expression pattern.

2.3.3.3 RNA stabilisation in bacteria

RNA stabilisation in bacteria shares many of the problems associated with stabilisation in cells and tissues (see section “RNA stabilisation in cells and tissues”). However, these effects are even more problematic when working with bacteria. Prokaryotic mRNAs, which have no 5'-cap and rarely have poly(A) tails, are highly unstable, with an average half-life of about three minutes. Sometimes the bacterial mRNA even begins to degrade while it is still being translated.

For this reason, gene expression studies are even more difficult in prokaryotes than in eukaryotes. Ideally, the RNA should be stabilised prior to harvesting,

since changes in the gene expression pattern may occur during the harvesting procedure – not only due to rapid RNA degradation but also due to transcriptional induction. Even the relatively short times required for centrifugation to collect the bacteria can allow induction and other artificial changes in the gene expression profile to occur.

2.3.3.4 RNA stabilisation in blood

Collection of whole blood is the first step in many molecular assays used to study cellular RNA. However, a major problem in such experiments is the instability of the cellular RNA profile *in vitro*. For example, copy numbers of individual mRNA species in whole blood can change more than 1,000-fold during storage or transport at room temperature. This is caused both by rapid RNA degradation and by changes in the gene expression pattern which occur after the blood is drawn. Such changes in the RNA expression profile make reliable studies of *in vivo* gene expression patterns impossible. A method that immediately preserves the RNA expression profile both during and after blood collection is therefore essential for accurate analysis of gene expression in whole blood.

2.3.3.5 Methods for stabilisation of the RNA expression profile

2.3.3.5.1 Strong denaturing reagents

Many RNA isolation methods rely on chaotropic agents such as guanidine isothiocyanate and guanidine hydrochloride to immediately inactivate RNases and thus ensure isolation of intact RNA. Although very effective during RNA purification, these reagents have limited effect since they can be added only after sample harvesting. Significant changes in gene expression can therefore still occur during sample handling and harvest and lead to artefacts in gene expression analysis (figure 2.21).

2.3.3.5.2 Rapid freezing

Traditional methods for stabilising nucleic acids rely on rapid freezing of the sample in liquid nitrogen or on dry ice. These methods require special equipment such as insulated vacuum flasks for handling cold materials.

An important consideration is how rapidly the sample is frozen. The size of the sample is critical since larger samples freeze more slowly, and changes in the RNA expression pattern can occur while the sample is freezing. Both liquid nitrogen and dry ice have a relatively low heat capacity, so that even with small samples, freezing does not occur instantaneous by throughout the entire sample.

In addition, as with addition of strong denaturing reagents, rapid freezing can only be carried out after sample harvest and thus cannot prevent changes occurring during, for example, a centrifugation step performed to harvest bacteria or suspension cells. Furthermore, it is important to remember that freezing does not permanently stabilise the nucleic acids in the sample. Freezing generates ice crystals,

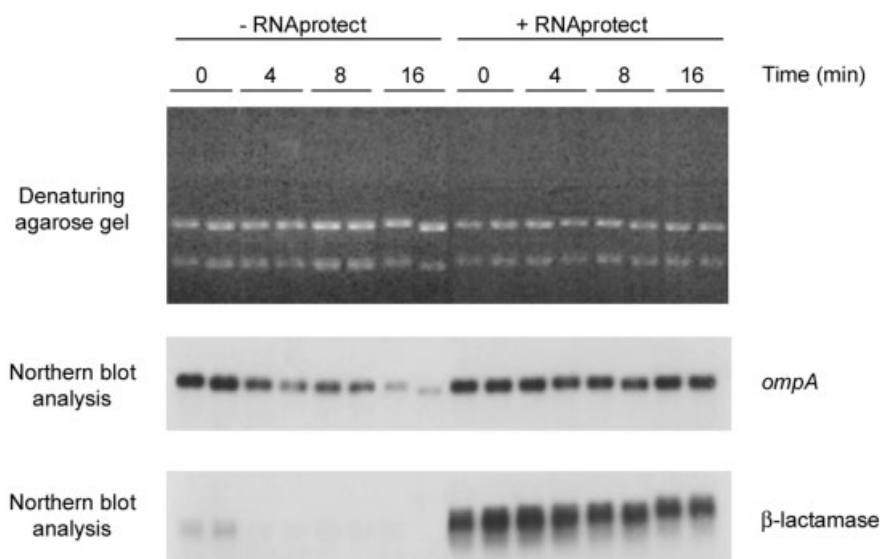


Figure 2.21 RNA stabilisation reagent for bacteria. In order to monitor mRNA degradation only, transcription was stopped by adding the RNA polymerase inhibitor rifampicin to a growing culture of *Escherichia coli*. The culture was split into two halves, and RNAprotect Bacteria Reagent (QIAGEN GmbH, Hilden, Germany) was added to one half. Samples were left at room temperature for zero, four, eight,

and sixteen minutes (min) before centrifugation and RNA isolation. The resulting RNA was analysed by denaturing agarose gel electrophoresis (**top panel**). Expression of two marker genes with different half lives was examined by northern blot analysis. **Middle panel:** outer membrane protein A (*ompA*) (half life of 15 minutes); **bottom panel:** β -lactamase (half life of two to five minutes).

which can break open organelles that segregate RNases and DNases *in vivo*. Subsequent thawing during sample processing can allow these endogenous RNases and DNases to act. Multiple freeze/thaw cycles can lead to further degradation. Therefore rapid freezing methods should be followed by disruption and homogenisation of the sample in strong denaturing reagents while taking care not to let the sample thaw.

2.3.3.5.3 Stabilisation of blood samples with anticoagulants

Blood samples are typically collected into a sealed tube containing an anticoagulant, such as citrate, heparin, or EDTA. While these prevent coagulation of the blood, their effect on nucleic acid stabilisation is limited. EDTA partially inhibits DNases by chelating Mg^{2+} , a cofactor for many DNases, but does little to stabilise RNA. In addition, commonly used anticoagulants may inhibit downstream applications such as PCR and RT-PCR.

2.3.3.6 Commercially available RNA stabilisation reagents

2.3.3.6.1 RNA stabilisation reagents for tissues

Reagents are now commercially available for stabilising RNA in tissues. The *RNA-later* RNA Stabilization Reagent (QIAGEN GmbH, Hilden, Germany) is a new technology that allows rapid and reliable analysis of gene expression patterns (QIAGEN GmbH, 2000c). Because of pending patents, the chemical composition of the *RNA-later* and PAXgene reagents cannot be given in the present article.

RNA in tissue samples stabilised in *RNA-later* RNA Stabilization Reagent is stable for one day at 37 °C, seven days at 18 °C to 25 °C, four weeks at 2 °C to 8 °C, or for extended periods at –20 °C or –80 °C. This allows cutting and handling of the sample at room temperature prior to RNA purification, avoiding the need for dry ice or liquid nitrogen. In addition, samples archived in *RNA-later* RNA Stabilization Reagent can be safely frozen, and, unlike samples simply frozen in liquid nitrogen, the sample can be thawed without affecting the RNA.

2.3.3.6.2 RNA stabilisation reagents for bacteria

The *RNAprotect* Bacteria Reagent (QIAGEN GmbH) is a unique reagent for immediate stabilisation of RNA in bacteria (QIAGEN GmbH, 2001). The reagent is added directly to the bacterial culture, thus preventing artificial changes in the gene expression profile that occur during bacterial harvest. RNA is stabilised before bacterial lysis, which allows more time for efficient disruption, while the bacterial RNA expression profiles are not affected by degradation or gene induction. *RNAprotect* bacteria reagent is suitable for use with both Gram-positive (*e.g.*, *Staphylococcus aureus* and *Mycobacterium avium*) and Gram-negative bacteria (*e.g.*, *Escherichia coli* and *Salmonella typhimurium*). The bacteria should be grown in minimal media for reliable and reproducible gene expression, although the *RNAprotect* Bacteria Reagent can also be used for bacteria grown in complex media.

2.3.3.6.3 RNA stabilisation reagents for blood

The *PAXgene* Blood RNA System (PreAnalytiX, Hombrechtikon, Switzerland) represents an integrated system for blood collection with RNA stabilisation and purification. The system requires the combined use of *PAXgene* Blood RNA tubes for blood collection and RNA stabilisation, followed by RNA isolation using the *PAXgene* Blood RNA kit. *PAXgene* Blood RNA tubes contain a proprietary blend of reagents based on a patented RNA stabilisation technology. Components of this blend protect RNA molecules from degradation by RNases and prevent induction of gene expression. The blend also promotes concentration of RNA during the first centrifugation step in the RNA isolation protocol, without needing to open the tube.

PAXgene Blood RNA tubes are intended for the collection of whole blood and stabilisation of the cellular RNA profiles. The cellular RNA profile remains stable for up to five days at room temperature.

2.3.4

Preparation of protein samples from cell cultures and tissues**2.3.4.1 Introduction**

Quantitative expression analysis or functional analysis of cellular proteins requires the isolation of these proteins. In most cases of analysis, preparations from cell cultures or tissues are used. Cell cultures usually contain a distinct cell type, while tissues represent a mixture of different cell types. Cell culture techniques can be applied to nearly all cell types from animals. In contrast to bacteria or yeast cells used in powerful recombinant DNA techniques, cultured cells from animals need the solid surface of a plastic vessel coated with a substance of the extracellular matrix for growth and division. Like in a living tissue, the cells of the resulting tissue culture are connected with neighbouring cells *via* macromolecular proteins of the so-called extracellular matrix. Furthermore, cultured cells are adhered to the surface of the coated culture vessel *via* extracellular protein components.

2.3.4.2 Separation and disruption of cells

The first step in sample preparation from cultured cells or tissues is commonly to break cell-cell contacts and cell-vessel contacts. Dissolving of the cell association is achieved by treatment with calcium (Ca^{2+})-chelating ethylenediaminetetraacetic acid (EDTA) and by addition of the proteinase trypsin, both of which promote the dissociation of cell-cell contacts and cell-vessel contacts. In the case of cell cultures, cells are then separated by gentle agitation (Barnes *et al.*, 1984). In a similar matter, tissues can be separated into their constituent cell types that can be sorted into cell type specific fractions using a fluorescence-activated cell sorter (FACS) (Herzenberg *et al.*, 1976).

Usually, if whole cells are not to be analysed, as is, for example, the case with Multi-epitope Ligand-Kartographie (MELK) technology (see chapter 5), the sorted or unsorted cells are subsequently disrupted into fragments that immediately reseal to form small vesicles. At this point of the preparation, proteinase inhibitors must be added to the suspension to prevent the activity of released cellular proteinases. Usually a cocktail of inhibitors, containing, for example, EDTA and phenylmethylsulfonyl fluoride are used to inhibit the activity of different kinds of proteinases (Alberts *et al.*, 1994). If the disruption step is applied carefully it leaves cellular organelles such as nuclei, mitochondria, the Golgi apparatus, lysosomes, and peroxisomes largely intact. Alternatively, tissues may be homogenised more thoroughly using a blender or mixer.

Careful disruption is achieved by mechanical procedures including (i) sonication; (ii) forcing cells through a small hole using high pressure ('French pressing'); (iii) shearing cells between a close fitting rotating plunger and the walls of a glass vessel; or (iv) using a mild detergent to cause holes in the plasma membrane. The resulting suspension is called a cell extract or cell homogenate. It contains soluble macromolecules and the membranous organelles (Alberts *et al.*, 1994).

2.3.4.3 Fractionation of cell extracts

The cell extract can be fractionated by centrifugation. As shown in figure 2.22, differential centrifugation separates cell components on the basis of size and density. The larger and denser components experience the greatest centrifugal force and move most rapidly. They sediment to form a pellet at the bottom of the tube, while smaller, less dense components remain in suspension above as the supernatant. A finer degree of separation can be achieved by velocity sedimentation or by equilibrium sedimentation.

Velocity sedimentation usually is performed in a sucrose density-gradient (typically five to twenty percent) with its dense end at the bottom of the centrifuge tube. The sample is applied as a small layer onto the top of the gradient. The rate at which a component sediments during subsequent ultra-centrifugation depends on its shape and size and can be described as its sedimentation coefficient or *S* value. In this way, even small molecules, as for example tRNAs and soluble proteins, can be separated from each other.

Equilibrium sedimentation by ultra-centrifugation uses sucrose or caesium chloride gradients at very high concentrations (*e.g.*, 20 to 70 percent sucrose). The sample is either layered on the top of, or dispensed within the steep gradient. During centrifugation, a component moves independently of its size or shape within a gradient until it forms a distinct band according to its buoyant density. This sensitive method can be used to separate macromolecules that contain heavy isotopes from the same macromolecules that do not. An overview to the methods described above is given by Alberts *et al.* (1994).

While intracellular soluble proteins are sufficient accessible to further applications membrane-associated proteins first have to be solubilised. Following a commonly used protocol, solubilisation of membrane proteins is achieved by the combined treatment of the sample with (i) the chaotropic agent urea that allows proteins to unfold and thus to expose their hydrophobic cores, (ii) surfactants (*e.g.*, Triton X-100) that solubilise the hydrophobic residues that are exposed as a result of denaturation by chaotropic salts, and (iii) reducing agents (*e.g.*, β -mercaptoethanol) to reduce disulfide bonds and thereby, promoting complete unfolding of a polypeptide. As a general problem, the required purification that can be performed either by centrifugation or by chromatography (see chapter 5) bears the risk to lose proteins, especially when several successive rounds of extraction and purification have to be performed (Herbert, 1999).

2.3.4.4 Sample preparation for downstream applications

In order to separate the proteins within a sample the sample must fulfil the specific demands (*e.g.*, volume, pH value, ion strength, or the absence of otherwise interfering components) of the method that is applied.

In the case of two-dimensional gel electrophoresis, still the most commonly used method for analysing whole proteomes or larger parts of it (see chapter 5), for example, interfering components such as lipids, nucleic acids, and salts must be removed from the fraction(s) of interest or the cell homogenate in order to obtain a

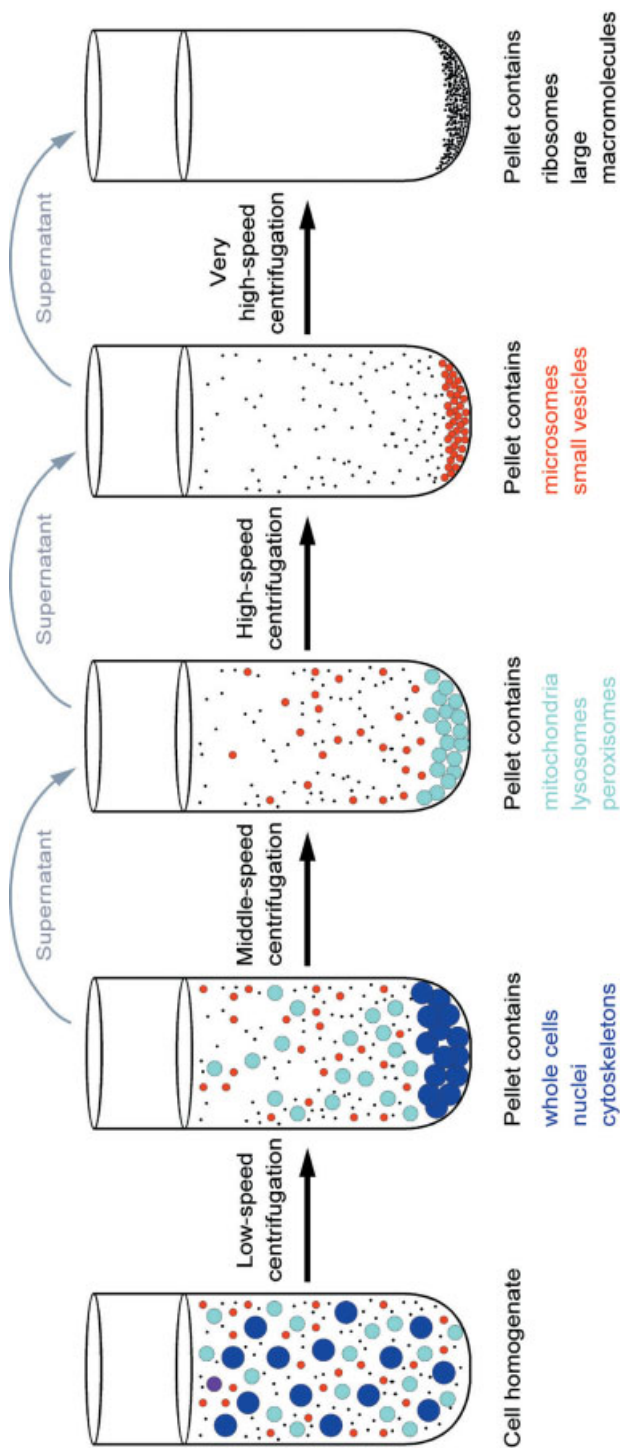


Figure 2.22 The different components of cell homogenates can be fractionated by successive centrifugation steps with increasing speeds (g values). Typical g values and centrifugation times for the mentioned centrifugation steps are (i) ten minutes at $1,000 \times g$ for low-speed centrifugation; (ii) 30 minutes at $20,000 \times g$ for middle-speed centrifugation; (iii) one hour at $75,000 \times g$ for very-high-speed centrifugation; and (iv) five hours at $150,000 \times g$ for very high-speed centrifugation. In general, the smaller and less dense the subcellular component, the greater the centrifugal force required for its sedimentation.

usable protein sample (Rabilloud, 1996). In contrast, crude samples containing whole cells including all subcellular components can be used for SDS polyacrylamide gel electrophoresis (see chapter 5). In this case, the protein samples are boiled in a suitable sample buffer providing the solubilisation of membrane-associated proteins (Laemmli, 1970).

Often, downstream applications require concentration of the sample in order to achieve a sufficient amount in a limited volume. Commonly, commercially available concentration tubes containing a porous membrane with a defined cut-off are used. The concentration is driven by centrifugal force. Alternatively, proteins can be precipitated. Most frequently, inorganic salts like ammonium sulphate or potassium phosphate are added to 'salt out' proteins.

For detailed information about protein purification the reader is referred to Scopes (1994).

2.3.4.5 Immunoprecipitation

The idea of using immunoprecipitation prior to SDS-based polyacrylamide gel electrophoresis (SDS-PAGE) and Western blotting is to isolate specific proteins (antigens) that are minor components of complex mixtures, such as a solution of all proteins of cultured cells or tissues (the lysate). This isolation also allows enrichment of the protein of interest to increase sensitivity of detection. This is achieved by incubation with antibodies directed against the protein of interest to guarantee high specific precipitation of single proteins. Although this method was developed to allow the enrichment of single proteins, immunoprecipitation combined with SDS-PAGE (maybe combined with radioactive metabolic labelling and detection by auto-radiography or immunological detection *via* Western blotting or silver staining of the gel for the detection of non-labelled proteins; Dustin *et al.*, 1986; Springer, 1994) can also be used to analyse the expression of genes at the protein level.

Immunoprecipitation consists of multiple ordered steps (figure 2.23). In detail, (i) cells of interest are first lysed with detergent if the antigen (the protein) to be precipitated is membrane-bound; (ii) homogenisation and centrifugation to remove cell debris; (iii) depending on the protocol, removal of proteins binding unspecific to the antibody used in the next step (*e.g.*, fibronectin is removed by incubation of the protein solution with homogenous protein G sepharose that is afterwards removed by precipitation during centrifugation); (iv) binding of a specific antigen to an antibody; (v) precipitation of the antibody/antigen complex; (vi) washing the precipitate; and (vii) dissociating the antigen from the immune complex, mostly by denaturation of the complex by adding SDS sample buffer for western blotting containing 2-mercaptoethanol or dithiothreitol (DTT). The dissociated antigen is then analysed by electrophoretic methods. The details of the immunoprecipitation protocols differ in the method of precipitation used and whether or not the antigen being precipitated is radiolabelled (Springer, 1994).

Commonly, radiolabelled antigens are immunoprecipitated with a specific antibody (polyclonal or monoclonal) covalently linked to sepharose or non-covalently

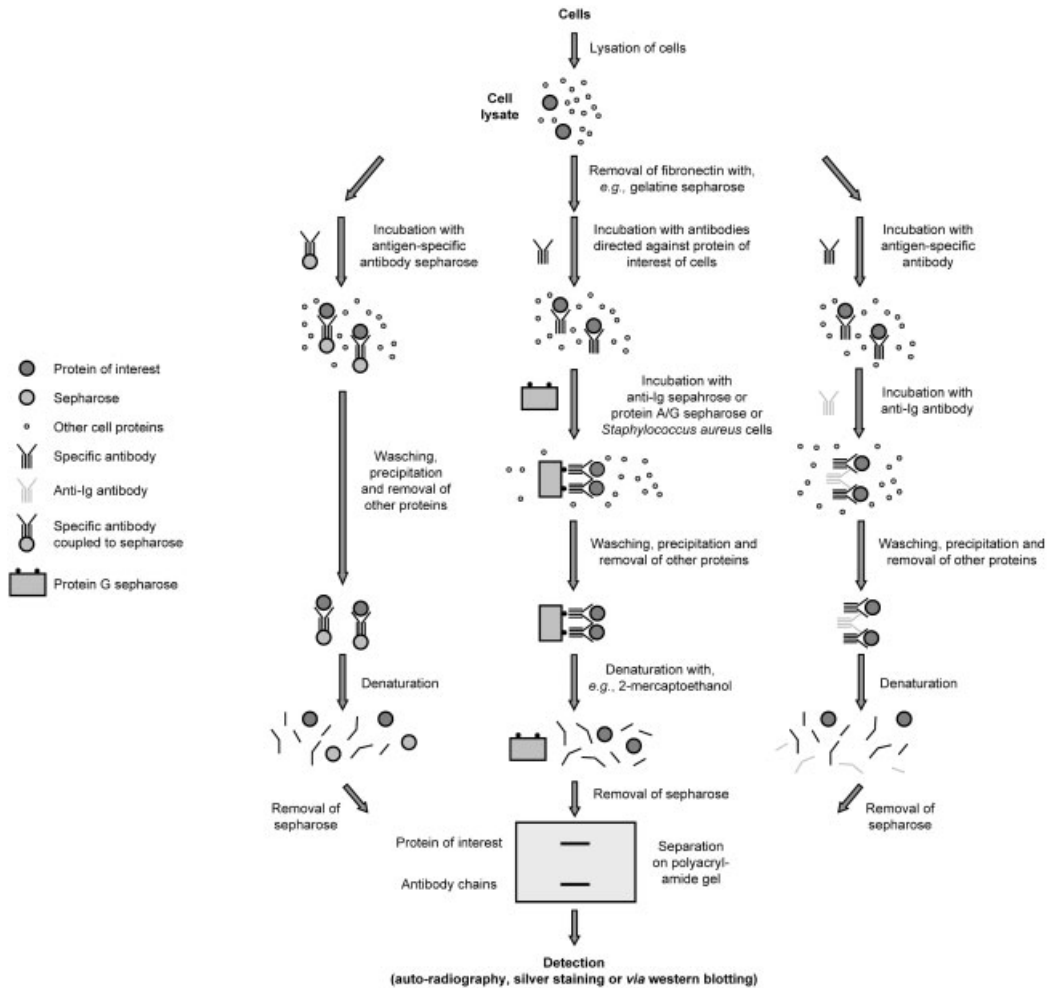


Figure 2.23 Schematic presentation of the principle procedures of immunoprecipitation for the isolation of protein antigens. See text and Springer (1994) for detailed information.

linked to protein G sepharose. Immunoprecipitation can be achieved with (i) polyclonal anti-immunoglobulin (Ig) serum, (ii) anti-Ig bound to sepharose (which is an insoluble, large-pore-size chromatographic matrix), (iii) *Staphylococcus aureus* protein A or (iv) *Streptococcus* protein G bound to sepharose, or (v) *Staphylococcus aureus* bacteria which contain protein A on the cell surface. In addition, modified protocols have been developed that should be used for immunoprecipitation of antigens that are non-specifically associated with other proteins, as well as protocols for the immunoprecipitation of non-labelled protein antigens (Springer, 1994).

Table 2.6 Relative affinities of proteins A and G for various IgG subclasses (adapted from Harlow & Lane, 1988, and Springer, 1994).

<i>Antibody</i>	<i>Affinity for protein A</i>	<i>Affinity for protein G</i>
Human IgG1	••••	••••
Human IgG2	••••	••••
Human IgG3	–	••
Human IgG4	••••	••••
Rat IgG1	• or ••	•
Rat IgG2a	–	••••
Rat IgG2b	–	••
Rat IgG2c	•• or ••••	••
Mouse IgG1	•	••••
Mouse IgG2a	•••	••••
Mouse IgG2b	•••	•••
Mouse IgG3	••	•••

Usually, monoclonal antibodies are non-precipitating, as are most polyclonal antibodies at the low antigen concentrations employed. Therefore, a reagent is used to precipitate or isolate the antibody/antigen complex. As mentioned above, this reagent may be anti-Ig serum, affinity-purified or monoclonal Ig, *Staphylococcus aureus* protein A or *Streptococcus* protein G coupled to sepharose, or *Staphylococcus aureus* Cowan strain II cells. Protein A methods are not generally applicable to either the rat (only its IgG1 and IgG2c subclasses bind) or the mouse (its IgG1 subclass binds too poorly for immunoprecipitation). In many cases, protein G can be substituted for protein A as a conjugate (see table 2.6). However, protein G sepharose appears to have a lower binding capacity.

Detailed standard protocols of immunoprecipitation for the isolation of protein antigens can be found in the 'Current Protocols in Molecular Biology' (Springer, 1994). Procedures for immunoprecipitation of antigens described in this series are similar to those described by Ho & Springer (1983), Kürzinger & Springer (1982), Sanchez-Madrid *et al.* (1983), Sastre *et al.* (1986), and Springer (1981).

2.3.4.6 Conclusions

In summary, the expense for protein sample preparation from cells and tissues depends largely on the downstream application chosen for protein separation. Another important selection criterion concerns the intracellular localisation of the protein(s) to be investigated, as soluble proteins are much easier to handle than membrane-associated ones. Therefore, standard procedures for the isolation of proteins must often be modified for specific applications.

2.4

Summary

A factor which is often underestimated in gene expression studies is the absolute necessity for clean starting materials. The results obtained will crucially depend on the material studied. This does not necessarily mean that relatively crude methods of isolation may not be sufficient for certain purposes. For example, useful information on gene expression in the atherosclerotic plaque has been obtained using relatively undifferentiated atherectomy (Faber *et al.*, 2001). However, such studies are likely to be of largely descriptive nature. Studies exploring functional aspects of gene expression will almost inevitably require the use of more differentiated methods allowing isolation of specific and narrowly circumscribed cell populations. Laser microdissection is a good example of such a method (Bonner *et al.*, 1997; Craven & Banks, 2001; Mills *et al.*, 2001b).

A further feature which is often overlooked, is the need to perform experiments with a sufficient number of replicates. This is necessary in order to deal with the problem of the large degree of inter-individual variation which may exist. This need for replicates should also be taken into account when deciding on the appropriate isolation method.

As the reader will by now have realised, many of the separation procedures are time and labour-intensive. This also applies to the methods of studying gene expression which are the subject of the following chapters. The experimenter should therefore take appropriate steps at the end of the isolation procedure to check that the purity and quality of the starting material is that required by the experimental design.

2.5

References

Recommended for further reading

- Aqueous two-phase systems. *Methods in Enzymology*. Walter, H., and Johansson, G., editors. Academic Press, New York, New York, USA 1994; 228.
- Braun, R. W., and Kumel, G. Separation of T cell subpopulations by monoclonal antibodies and affinity chromatography. *Methods in Enzymology* 1986; 121:737-748.
- Cell separation methods and applications. Recktenwald, D., and Radbruch, A., editors. Marcel Dekker, New York, New York, USA 1998.
- Eshhar, Z., Waks, T., and Bustin, M. Isolation of thymocytes by affinity chromatography. *Methods in Enzymology* 1974; 34:750-755.
- Guide to protein purification. *Methods in Enzymology*. Deutscher, M. P., editor. Academic Press, New York, New York, USA 1990; 182.
- Immunochemical techniques, part G: Separation and characterization of lymphoid cells. Di Sabato, G., Langone, J. J., and van Vunakis, H., editors. *Methods in Enzymology*. Academic Press, New York, New York, USA 1984; 108.
- Mills, K. An indirect rosette technique for the identification and separation of lymphocyte populations by monoclonal antibodies. *Methods in Enzymology* 1986; 121:726-737.
- RNA processing, part A: general methods. *Methods in Enzymology*. Dahlberg, J. E., and

- Abelson, J. N., editors. Academic Press, New York, New York, USA **1989**; 180.
- RNA processing, part B: specific methods. *Methods in Enzymology*. Dahlberg, J. E., and Abelson, J. N., editors. Academic Press, New York, New York, USA **1990**; 181.
- Internet resources**
- American Type Culture Collection (ATCC; Manassas, Vancouver, USA); accessible at <http://www.atcc.org>.
- cDNA xProfiler (National Cancer Institute, NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/Tissues>.
- Differential Gene Expression Displayer, DGED (NCI, Bethesda, Maryland, USA) ; accessible at <http://cgap.nci.nih.gov/Tissues>.
- German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, DSMZ; Braunschweig, Germany); accessible at <http://www.dsmz.de/>.
- Gene Library Summarizer, GLS (NCI, Bethesda, Maryland, USA) ; accessible at <http://cgap.nci.nih.gov/Tissues>.
- Molecular Probes, Inc. (Eugene, Oregon, USA); available at <http://www.molecularprobes.com/>.
- Mouse Genome Informatics Project (The Jackson Laboratory, Bar Harbor, Maine, USA); accessible at <http://www.informatics.jax.org/>.
- SAGE gene to tag mapping approach (National Center for Biotechnology Information, NCBI; National Library of Medicine, Bethesda, Maryland, USA); accessible at <http://www.ncbi.nlm.nih.gov/SAGE/>.
- TSRI Spectrum Viewer (The Scripps Research Institute, San Diego, California, USA); available at <http://facs.scripps.edu/spectra/>.
- Cited publications**
- Abraham, R. T. Mutant T cell lines as model systems for the dissection of T cell antigen receptor signaling pathways. *Immunologic Research* **2000**; 22(2-3):95-117.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. Molecular biology of the cell. 3rd edition. Garland Publishing, New York, New York, USA **1994**.
- Albertsson, P.-Å. Thin-layer countercurrent distribution. *Analytical Biochemistry* **1965**; 11(1):121-125.
- Albertsson, P.-Å. Partition of cell particles and macromolecules. 3rd edition, Wiley, New York, New York, USA **1986**.
- Albertsson, P.-Å., and Baird, G. D. Counter-current distribution of cells. *Experimental Cell Research* **1962**; 28(2):296-322.
- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wyndham, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**; 403(6769):503-511.
- Amara, R. R., and Satchidanandam, V. Specific polyadenylation and purification of total messenger RNA from *Escherichia coli*. *Nucleic Acids Research* **1997**; 25(17): 3465-3470.
- Amara, R. R., and Satchidanandam, V. Refinements of the differential display reverse transcription-polymerase chain reaction technique: use of oligo(dT)-based anchored primers with *Escherichia coli* messenger RNA identifies a salt-induced promoter in the *dcw* gene cluster. *Analytical Biochemistry* **2000**; 278(1):83-86.
- Anderson, G., Jenkinson, E. J., Moore, N. C., and Owen, J. J. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature* **1993**; 362(6415):70-73.
- Andreata, S., Wallinger, M. M., Posch, T., Psenner, R. Detection of subgroups from flow cytometry measurements of heterotrophic bacterioplankton by image analysis. *Cytometry* **2001**; 44(3): 218-225.
- Ashcroft, R. G., and Lopez, P. A. Commercial high speed machines open new opportunities in high throughput flow cytometry (HTFC). *Journal of Immunological Methods* **2000**; 243(1-2): 13-24.
- Aviv, H. and Leder, P. Purification of biologically active globin messenger RNA by chromatography on oligo thymidylic acid-cellulose. *Proceedings of the National Academy of Sciences of the United States of America* **1972**; 69(1):1498-1512.
- Baugh, L. R., Hill, A. A., Brown, E. L., and Hunter, C. P. Quantitative analysis of mRNA

- amplification by *in vitro* transcription. *Nucleic Acids Research* **2001**; 29(5):E29.
- Baumgarth, N., and Roederer, M. A practical approach to multicolor flow cytometry for immunophenotyping. *Journal of Immunological Methods* **2000**; 243(1-2): 77-97.
- Becker, F. F., Wang, X. B., Huang, Y., Pethig, R., Vykoukal, J., and Gascoyne, P. R. Separation of human breast cancer cells from blood by differential dielectric affinity. *Proceedings of the National Academy of Sciences of the United States of America* **1995**; 92(3):860-864.
- Becker, B., Vogt, T., Landthaler, M., and Stolz, W. Detection of differentially regulated genes in keratinocytes by cDNA array hybridization: Hsp27 and other novel players in response to artificial ultraviolet radiation. *Journal of Investigative Dermatology* **2001**; 116(6):983-988.
- Bengtsson, S., Phillipson, L., and Albertsson, P.-Å. Counter-current distribution of polio virus. *Biochemical and Biophysical Research Communications* **1962**; 9(4):318-322.
- Beranger, F., Mange, A., Solassol, J., and Lehmann, S. Cell culture models of transmissible spongiform encephalopathies. *Biochemical and Biophysical Research Communications* **2001**; 289(2):311-316.
- Berardi, A. C., Wang, A., Levine, J. D., Lopez, P., and Scadden, D. T. Functional isolation and characterization of human hematopoietic stem cells. *Science* **1995**; 267(5194):104-108.
- Bigos, M., Baumgarth, N., Jager, C. G., Herman, O. C., Nozaki, T., Stovel, R. T., Parks, D. R., and Herzenberg, L. A. Nine color eleven parameter immunophenotyping using three laser flow cytometry. *Cytometry* **1999**; 36(1): 36-45.
- Bijli, K. M., Singh, B. P., Sridhara, S., and Arora, N. Isolation of total RNA from pollens. *Preparative Biochemistry and Biotechnology* **2001**; 31(2):155-162.
- Blackburn, P., Wilson, G., and Moore, S. Ribonuclease inhibitor from human placenta. Purification and properties. *Journal of Biological Chemistry* **1977**; 252(16):5904-5910.
- Blick, M., Westin, E., Guterman, J., Wong-Staal, F., Gallo, R., McCredie, K., Keating, M., and Murphy, E. Oncogene expression in human leukemia. *Blood* **1984**; 64(6):1234-1239.
- Bonner, W. A., Hulett, H. R., and Sweet, R. G. Fluorescence activated cell sorting. *Review of Scientific Instruments* **1972**; 43(3): 404.
- Bonner, R. F., Emmert-Buck, M., Cole, K., Pohida, T., Chuaqui, R., Goldstein, S., and Liotta, L. A. Laser capture microdissection: molecular analysis of tissue. *Science* **1997**; 278(5342):1481-1483.
- Boulton, A. A. Practical cell culture techniques. Humana Press, Totowa, New Jersey, USA **1992**.
- Bøyum, A. Isolation of mononuclear cells and granulocytes from human blood. Isolation of mononuclear cells by one centrifugation, and of granulocytes by combining centrifugation and sedimentation at 1 g. *Scandinavian Journal of Clinical and Laboratory Investigation, Supplement* **1968**; 97:77-89.
- Bøyum, A. Separation of lymphocytes, lymphocyte subgroups and monocytes: A review. *Lymphology* **1977**; 10(2):71-76.
- Bøyum, A. Separation of lymphocytes, granulocytes, and monocytes from human blood using iodinated density gradient media. *Methods in Enzymology* **1984**; 108:88-102.
- Braun, A., Hammerle, S., Suda, K., Rothen-Rutishauser, B., Gunther, M., Kramer, S. D., and Wunderli-Allenspach, H. Cell cultures as tools in biopharmacy. *European Journal Pharmaceutical Sciences* **2000**; 11(Supplement 2):S51-S60.
- Brown, A. J. Preparation of total RNA. *Methods in Molecular Biology* **1996**; 53:269-276.
- Buckley, P. J., Beelen, R. H., Burns, J., Beard, C. M., Dickson, S. A., and Walker, W. S. Isolation of human splenic macrophages and lymphocytes by countercurrent centrifugal elutriation. *Journal of Immunological Methods* **1984**; 66(2):201-217.
- Bunz, F. Human cell knockouts. *Current Opinion in Oncology* **2002**; 14(1):73-78.
- Bynum, J. W., and Ronzio, R. A. An improved method for the rapid isolation of RNA from tissue with high ribonuclease content. *Analytical Biochemistry* **1976**; 73(1):209-214.
- Chaganti, R. S., and Houldsworth, J. Genetics and biology of adult human male germ cell tumors. *Cancer Research* **2000**; 60(6):1475-1482.
- Carlson, G. P., and Chadwick, R. W. Comparability of *in vitro* and *in vivo* methods for the determination of alterations in drug metabolism. *Toxicology Letters* **1987**; 39(2-3):125-138.
- Carpenter, C. D., and Simon, A. E. Preparation of RNA. *Methods in Molecular Biology* **1998**; 82:85-89.

- Carter, N. P., Meyer, E. W. μ Flow cytometry. A practical approach. Second Edition. Ormerod, M. G. IRL Press, Oxford, United Kingdom 1995.
- Cathala, G., Savoret, J.-F., Mendez, B., West, B. L., Karin, M., Martial, J. A., and Baxter, J. D. Laboratory methods: a method for isolation of intact, translationally active ribonucleic acid. *DNA* 1983; 2(4):329-335.
- Cheng, S. H., and Seemann, J. R. Extraction and purification of RNA from plant tissue enriched in polysaccharides. *Methods in Molecular Biology* 1998; 86:27-32.
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* 1979; 18(24):5294-5299.
- Clynes, M. Animal cell culture techniques. Springer-Verlag, Berlin, Germany 1998.
- Coligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W., and Wingfield P. T. In: Current protocols in protein science. Coligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W., and Wingfield P. T., eds. John Wiley & Sons, New York, New York, USA 1995.
- Coons, A. H., and Kaplan, M. H. Localization of antigen in tissue dells. Improvement in a method for the detection of antigens by means of fluorescent antibody. *Journal of Experimental Medicine* 1950; 91(1):1-13.
- Corcoran, R. M., and Lopez, P. A. Cells sorting at 50,000 events per second – Practical consideration. Cytomation, Inc., Application note. Cytomation, Inc., Fort Collins, Colorado, USA.
- Coulter, W. H. High speed automatic blood cell counter and cell size analyzer. *Proceedings of the Natural Electronics Conference* 1956; 12: 1034-1042
- Craven, R. A., and Banks, R. E. Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics* 2001; 1(10):1200-1204.
- Cullen, P., Mohr, S., Brennhause, B., Cignarella, A., and Assmann, G. Downregulation of the selectin ligand-producing fucosyltransferases Fuc-TIV and Fuc-TVII during foam cell formation in monocyte-derived macrophages. *Arteriosclerosis, Thrombosis and Vascular Biology* 1997; 17(8):1591-1598.
- Curtis, A. S., McGrath, M., and Gasmi, L. Localised application of an activating signal to a cell: experimental use of fibronectin bound to beads and the implications for mechanisms of adhesion. *Journal of Cell Science* 1992; 101(Part 2):427-436.
- Dennison, C. A guide to protein isolation. Kluwer Academic Publishing, Dordrecht, The Netherlands 1999.
- De Rosa, S. C., Herzenberg, L. A., Herzenberg, L. A., and Roederer, M. 11-color, 13-parameter flow cytometry: identification of human naïve T cells by phenotype, function, and T cell receptor diversity. *Nature Medicine* 2001; 7(2): 245-248.
- De Waard, V., van den Berg, B. M., Veken, J., Schultz-Heienbrok, R., Pannekoek, H., and van Zonneveld, A. J. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene* 1999; 226(1):1-8.
- Dirks, W., MacLeod, R. A., Jager, K., Milch, H., and Drexler, H. G. First searchable database for DNA profiles of human cell lines: sequential use of fingerprint techniques for authentication. *Cellular and Molecular Biology (Noisy-le-grand)* 1999; 45(6):841-853.
- Doyle, A. Mammalian cell culture. Essential techniques. Wiley-VCH, Weinheim, Germany 1997.
- Drexler, H. G., and Matsuo, Y. Malignant hematopoietic cell lines: *in vitro* models for the study of natural killer cell leukemia-lymphoma. *Leukemia* 2000; 14(5):777-782.
- Drexler, H. G., Matsuo, A. Y., and MacLeod, R. A. Continuous hematopoietic cell lines as model systems for leukemia-lymphoma research. *Leukemia Research* 2000; 24(11):881-911.
- Dunn, A. Isolation of messenger RNA from plant tissues. *Methods in Molecular Biology* 1997; 69:33-38.
- Dustin, M. L., Rothlein, R., Bhan, A. K., Dinarello, C. A., and Springer, T. A. Induction by IL 1 and interferon-gamma: tissue distribution, biochemistry, and function of a natural adherence molecule (ICAM-1). *Journal of Immunology* 1986; 137(1):245-254.
- Dynal A. S. *Cell separation and protein purification. Technical handbook*; Dynal A. S., Oslo, Norway 1996.
- Engel, T., Lorkowski, S., Lueken, A., Rust, S., Schlüter, B., Berger, G., Cullen, P., and Assmann G. The human ABCG4 gene is regulated by oxysterols and retinoids in monocyte-derived macrophages. *Biochemical and Biophysical Research Communications* 2001; 288(2):483-488.

- Ericson, I. Determination of the isoelectric point of rat liver mitochondria by cross-partition. *Biochimica et Biophysica Acta* **1974**; *356*(1):100-107.
- Escot, C., Le Roy, X., Chalbos, D., Joyeux, C., Simonsen, E., Daures, J. P., Soussaline, F., and Rochefort, H. Computer-aided quantification of RNA levels detected by *in situ* hybridization of tissue sections. *Analytical Cellular Pathology* **1991**; *3*(4):215-224.
- Faber, B. C., Cleutjens, K. B., Niessen, R. L., Aarts, P. L., Boon, W., Greenberg, A. S., Kitslaar, P. J., Tordoir, J. H., and Daemen, M. J. Identification of genes potentially involved in rupture of human atherosclerotic plaques. *Circulation Research* **2001**; *89*(6):547-554.
- Faggioli, F., Pasquini, G., and Barba, M. Comparison of different methods of RNA isolation for plum pox virus detection by reverse transcription-polymerase chain reaction. *Acta Virologica* **1998**; *42*(4):219-221.
- Farrell, R. E. In: RNA methodologies: a laboratory guide for isolation and characterization. Farrell, R. E., ed. Academic Press, San Diego, California, USA **1998**.
- Fausto, N. Messenger RNA in regenerating liver: implications for the understanding of regulated growth. *Molecular and Cellular Biochemistry* **1984**; *59*(1-2):131-147.
- Favaloro, J., Treisman, R., and Kamen, R. Transcription maps of polyoma virus-specific RNA: analysis by two-dimensional nuclease S1 gel mapping. *Methods in Enzymology* **1980**; *65*(1):718-749.
- Fedorcsak, I., and Ehrenberg, L. Effects of diethyl pyrocarbonate and methyl methane-sulfonate on nucleic acids and nucleases. *Acta Chemica Scandinavica* **1966**; *20*(1): 107-112.
- Fend, F., Emmert-Buck, M. R., Chuaqui, R., Cole, K., Lee, J., Liotta, L. A., and Raffeld, M. Immuno-LCM: laser capture microdissection of immunostained frozen sections for mRNA analysis. *American Journal of Pathology* **1999**; *154*(1):61-66.
- Fink, L., Kohlhoff, S., Stein, M. M., Hanze, J., Weissmann, N., Rose, F., Akkayagil, E., Manz, D., Grimminger, F., Seeger, W., and Bohle, R. M. cDNA array hybridization after laser-assisted microdissection from nonneoplastic tissue. *American Journal of Pathology* **2002**; *160*(1):81-90.
- Fislage, R., Berceanu, M., Humboldt, Y., Wendt, M., and Oberender, H. Primer design for a prokaryotic differential display RT-PCR. *Nucleic Acids Research* **1997**; *25*(9):1830-1835.
- Fleming, H. Structure and function of cultured endometrial epithelial cells. *Seminars in Reproductive Endocrinology* **1999**; *17*(1):93-106.
- Flø, R. W., Næss, A., Lund-Johansen, F., Mæhle, B. O., Sjørusen, H., Lehmann, V., and Solberg, C. O. Negative selection of human monocytes using magnetic particles covered by anti-lymphocyte antibodies. *Journal of Immunological Methods* **1991**; *137*(1):89-94.
- Fricova, V., Hrubá, A., and Pristoupil, T. I. Sephadex G-25 beads as chromatographic carrier for the fractionation of animal cells. *Journal of Chromatography* **1974**; *92*(2): 335-338.
- Galileo, D. S., Gee, A. P., and Linser, P. J. Neurons are replenished in cultures of embryonic chick optic tectum after immunomagnetic depletion. *Developmental Biology* **1991**; *146*(2):278-291.
- Gemma, A., Takenaka, K., Hosoya, Y., Matuda, K., Seike, M., Kurimoto, F., Ono, Y., Uematsu, K., Takeda, Y., Hibino, S., Yoshimura, A., Shibuya, M., and Kudoh, S. Altered expression of several genes in highly metastatic subpopulations of a human pulmonary adenocarcinoma cell line. *European Journal of Cancer* **2001**; *37*(12):1554-1561.
- Geretti, A. M., van Els, C. A., van Baalen, C. A., Poelen, M. C., and Osterhaus, A. D. Preservation of phenotype and function of positively selected virus-specific CD8⁺ T lymphocytes following anti-Fab detachment from immunomagnetic beads. *Journal of Immunological Methods* **1993**; *161*(1):129-133.
- Gilman, M. Phenol/SDS method for plant RNA preparation. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1994**:4.3.1-4.3.4.
- Gokhan, S., Song, Q., and Mehler, M. F. Generation and regulation of developing immortalized neural cell lines. *Methods* **1998**; *16*(3):345-358.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**; *286*(5439):531-537.

- Gordon, G. J., Coleman, W. B., and Grisham, J. W. Temporal analysis of hepatocyte differentiation by small hepatocyte-like progenitor cells during liver regeneration in retrorsine-exposed rats. *American Journal of Pathology* **2000**; 157(3):771-786.
- Greenberg, A. W., and Hammer, D. A. Cell separation mediated by differential rolling adhesion. *Biotechnology and Bioengineering* **2001**; 73(2):111-124.
- Grinnell, F., Milam, M., and Sreere, P. A. Studies on cell adhesion. II. Adhesion of cells to surfaces of diverse chemical composition and inhibition of adhesion by sulfhydryl binding reagents. *Archives of Biochemistry and Biophysics* **1972**; 153(1):193-198.
- Groner, B. Oncogene expression in mammary epithelial cells. *Journal of Cellular Biochemistry* **1992**; 49(2):128-136.
- Gumperz, J. E. Generation of HLA class I transfected target cell lines. *Methods in Molecular Biology* **2000**; 121:49-60.
- Hadjivassiliou, V., Green, M. H., and Green, I. C. Immunomagnetic purification of beta cells from rat islets of Langerhans. *Diabetologia* **2000**; 43(9):1170-1177.
- Hahn, S., Zhong, X. Y., Troeger, C., Burge-meister, R., Gloning, K., and Holzgreve, W. Current applications of single-cell PCR. *Cellular and Molecular Life Sciences* **2000**; 57(1):96-105.
- Handfield, M., and Levesque, R. C. Strategies for isolation of *in vivo* expressed genes from bacteria. *FEMS Microbiology Reviews* **1999**; 23(1):69-91.
- Hansel, T. T., Pound, J. D., Pilling, D., Kitas, G. D., Salmon, M., Gentle, T. A., Lee, S. S., and Thompson, R. A. Purification of human blood eosinophils by negative selection using immunomagnetic beads. *Journal of Immunological Methods* **1989**; 122(1):97-103.
- Harlow, E., and Lane, D. Antibodies: A laboratory manual. Cold Spring Harbor Laboratory. Cold Spring Harbor, New York, USA **1988**.
- Harrison, M. A., and Rae, I. F. General techniques of cell culture. Cambridge University Press, Cambridge, United Kingdom **1997**.
- Harsch, M., Bendrat, K., Hofmeier, G., Branscheid, D., and Niendorf, A. A new method for histological microdissection utilizing an ultrasonically oscillating needle. *American Journal of Pathology* **2001**; 158(6):1985-1990.
- Hay, R. J. Human cells and cell cultures: availability, authentication and future prospects. *Human Cell* **1996**; 9(3):143-152.
- Heel, K., and Dawkins, H. Laser microdissection and optical tweezers in research. *Today's Life Science* **2001**; 13(2):42-48.
- Herbert, B. Advances in protein solubilization for two-dimensional electrophoresis. *Electrophoresis* **1999**; 20(4-5):660-663.
- Herzenberg, L. A., Sweet, R. G., and Herzenberg, L. A. Fluorescence activated cell sorting. *Scientific American* **1976**; 234(3):108-116.
- Hewett, P. W., and Murray, J. C. Immunomagnetic purification of human microvessel endothelial cells using Dynabeads coated with monoclonal antibodies to PECAM-1. *European Journal of Cell Biology* **1993**; 62(2):451-454.
- Ho, M. K., and Springer, T. A. Biosynthesis and assembly of the alpha and beta subunits of Mac-1, a macrophage glycoprotein associated with complement receptor function. *Journal of Biological Chemistry* **1983**; 258(5):2766-2769.
- Hoffmann, K. F., McCarty, T. C., Segal, D. H., Chiamonte, M., Hesse, M., Davis, E. M., Cheever, A. W., Meltzer, P. S., Morse, H. C., 3rd, and Wynn, T. A. Disease fingerprinting with cDNA microarrays reveals distinct gene expression profiles in lethal type 1 and type 2 cytokine-mediated inflammatory reactions. *FASEB Journal* **2001**; 15(13):2545-2547.
- Hofmann, W. K., de Vos, S., Tsukasaki, K., Wachsmann, W., Pinkus, G. S., Said, J. W., and Koeffler, H. P. Altered apoptosis pathways in mantle cell lymphoma detected by oligonucleotide microarray. *Blood* **2001**; 98(3):787-794.
- Huang, Q., Liu, D., Majewski, P., Schulte, L. C., Korn, J. M., Young, R. A., Lander, E. S., and Hacohen, N. The plasticity of dendritic cell responses to pathogens and their components. *Science* **2001**; 294(5543):870-875.
- Hudson, L., and Phillips, B. The use of column bound B lymphocytes for absorption of antisera to T lymphocyte antigens. *Journal of Immunology* **1973**; 110(6):1663-1666.
- Ito, Y., Suaudeau, J., and Bowman, R. L. New flow-through centrifuge without rotating seals applied to plasmapheresis. *Science* **1975**; 189(4207):999-1000.
- Ito, Y., and Shinomiya, K. A new continuous-flow cell separation method based on cell density: Principle, apparatus, and preliminary application to separation of human buffy

- coat. *Journal of Clinical Apheresis* **2001**; 16(4):186-191.
- Jackson, C. J., Garbett, P. K., Nissen, B., and Schrieber, L. Binding of human endothelium to *Ulex europaeus* I-coated Dynabeads: application to the isolation of microvascular endothelium. *Journal of Cell Science* **1990**; 96(Part 2):257-262.
- Kaplan, B. B., Bernstein, S. L., and Gioio, A. E. An improved method for the rapid isolation of brain ribonucleic acid. *Biochemical Journal* **1979**; 183(1):181-184.
- Kaplan, J., and Hukku, B. Cell line characterization and authentication. *Methods in Cell Biology* **1998**; 57:203-216.
- Karsten, S. L., Van Deerlin, V. M., Sabatti, C., Gill, L. H., and Geschwind, D. H. An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis. *Nucleic Acids Research* **2002**; 30(2):E4.
- Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Thorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine* **1998**; 4(7):844-847.
- Krieg, P. A. In: A laboratory guide to RNA: isolation, analysis and synthesis. Krieg, P. A., ed. John Wiley & Sons, New York, USA **1996**.
- Kürzinger, K., and Springer, T. A. Purification and structural characterization of LFA-1, a lymphocyte function-associated antigen, and Mac-1, a related macrophage differentiation antigen associated with the type three complement receptor. *Journal of Biological Chemistry* **1982**; 257(20):12412-12418.
- Lahr, G. RT-PCR from archival single cells is a suitable method to analyse specific gene expression. *Laboratory Investigation* **2000**; 80(9):1477-1479.
- Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. A public database for gene expression in human cancers. *Cancer Research* **1999**; 59(21):5403-5407.
- Lawn, R. M., Wade, D. P., Couse, T. L., and Wilcox, J. N. Localization of human ATP-binding cassette transporter 1 (ABC1) in normal and atherosclerotic tissues. *Arteriosclerosis, Thrombosis and Vascular Biology* **2001**; 21(3):378-385.
- Lehr, H. A., Jacobs, T. W., Yaziji, H., Schnitt, S. J., and Gown, A. M. Quantitative evaluation of HER-2/neu status in breast cancer by fluorescence in situ hybridization and by immunohistochemistry with image analysis. *American Journal of Clinical Pathology* **2001**; 115(6):814-822.
- Loos, A., Glanemann, C., Willis, L. B., O'Brien, X. M., Lessard, P. A., Gerstmeier, R., Guillouet, S., and Sinskey, A. J. Development and validation of corynebacterium DNA microarrays. *Applied and Environmental Microbiology* **2001**; 67(5):2310-2318.
- Lorkowski, S., Kratz, M., Wenner, C., Schmidt, R., Weitkamp, B., Fobker, M., Reinhardt, J., Rauterberg, J., Galinski, E. A., and Cullen, P. Expression of the ATP-binding cassette transporter gene ABCG1 (ABC8) in Tangier disease. *Biochemical and Biophysical Research Communications* **2001**; 283(4):821-830.
- Luo, L., Salunga, R. C., Guo, H., Bittner, A., Joy, K. C., Galindo, J. E., Xiao, H., Rogers, K. E., Wan, J. S., Jackson, M. R., and Erlander, M. G. Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Medicine* **1999**; 5(1):117-122.
- Luxembourg, A. T., Borrow, P., Teyton, L., Brunmark, A. B., Peterson, P. A., and Jackson, M. R. Biomagnetic isolation of antigen-specific CD8⁺ T cells usable in immunotherapy. *Nature Biotechnology* **1998**; 16(3):281-285.
- Luzzi, V., Holtschlag, V., and Watson, M. A. Expression profiling of ductal carcinoma in situ by laser capture microdissection and high-density oligonucleotide arrays. *American Journal of Pathology* **2001**; 158(6):2005-2010.
- Macfarlane, D. E., and Dahle, C. E. Introduction to isolating RNA. *Methods in Molecular Biology* **1998**; 86:1-6.
- MacLeod, R. A. DNA fingerprints of cell lines. *Nature* **1992**; 359(6397):681-682.
- Maiman, T. Stimulated optical radiation in ruby lasers. *Nature* **1960**; 187(4736):493-494.
- Manyonda, I. T., Soltys, A. J., and Hay, F. C. A critical evaluation of the magnetic cell sorter and its use in the positive and negative selection of CD45RO⁺ cells. *Journal of Immunological Methods* **1992**; 149(1):1-10.
- Mayer, A., Stich, M., Broksch, D., Schütze, K., and Lahr, G. Going in vivo with laser micro-

- dissection. *Methods in Enzymology* **2002**; 356:in press.
- Meier-Ruge, W., Bielser, W., Remy, E., Hillenkamp, F., Nitsche, R., and Unsold, R. The laser in the Lowry technique for microdissection of freeze-dried tissue slices. *Histochemical Journal* **1976**; 8(4):387-401.
- Melamed, M. R., Lindmo, T., and Mendelsohn, M. L. Flow cytometry and sorting, 2nd edition. Wiley-Liss, New York, New York, USA **1990**.
- Mikita, T., Porter, G., Lawn, R. M., and Shiffman, D. Oxidized low density lipoprotein exposure alters the transcriptional response of macrophages to inflammatory stimulus. *Journal of Biological Chemistry* **2001**; 276(49):45729-45739.
- Mills, G. B., Bast, R. C., Jr., and Srivastava, S. Future for ovarian cancer screening: novel markers from emerging technologies of transcriptional profiling and proteomics. *Journal of the National Cancer Institute* **2001a**; 93(19):1437-1439.
- Mills, J. C., Roth, K. A., Cagan, R. L., and Gordon, J. I. DNA microarrays and beyond: completing the journey from tissue to cell. *Nature Cell Biology* **2001b**; 3(8):E175-E178.
- Mok, S. C., Chao, J., Skates, S., Wong, K., Yiu, G. K., Muto, M. G., Berkowitz, R. S., and Cramer, D. W. Prostatein, a potential serum marker for ovarian cancer: identification through microarray technology. *Journal of the National Cancer Institute* **2001**; 93(19):1458-1464.
- Molday, R. S., Yen, S. P., and Rembaum, A. Application of magnetic microspheres in labelling and separation of cells. *Nature* **1977**; 268(5619):437-438.
- Mrazek, F., and Petrek, M. Processing of mRNA from human leukocytes by biomagnetic separation: comparison with current methods of RNA isolation. *Acta Universitatis Palackianae Olomucensis Facultatis Medicae* **1999**; 142:23-28.
- Müller, R., Verma, I. M., and Adamson, E. D. Expression of c-onc genes: c-fos transcripts accumulate to high levels during development of mouse placenta, yolk sac and amnion. *EMBO Journal* **1983**; 2(5):679-684.
- Murphy, S. J., Watt, D. J., and Jones, G. E. An evaluation of cell separation techniques in a model mixed cell population. *Journal of Cell Science* **1992**; 102(Part 4):789-798.
- Nacht, M., Ferguson, A. T., Zhang, W., Petroziello, J. M., Cook, B. P., Gao, Y. H., Maguire, S., Riley, D., Coppola, G., Landes, G. M., Madden, S. L., and Sukumar, S. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Research* **1999**; 59(21):5464-5470.
- Nicholson, B. L. Techniques in fish cell culture. Elsevier Science, Limerick, The Netherlands **1985**.
- Nolta, J. A., Smogorzewska, E. M., and Kohn, D. B. Analysis of optimal conditions for retroviral-mediated transduction of primitive human hematopoietic cells. *Blood* **1995**; 86(1):101-110.
- Norrby, E. C. J., and Albertsson, P.-Å. Concentration of poliovirus by an aqueous polymer two-phase system. *Nature* **1960**; 188(4755):1047-1048.
- O'Brien, K. D., Gordon, D., Deeb, S., Ferguson, M., and Chait, A. Lipoprotein lipase is synthesized by macrophage-derived foam cells in human coronary atherosclerotic plaques. *Journal of Clinical Investigation* **1992**; 89(5):1544-1550.
- Oemar, B. S., Werner, A., Garnier, J. M., Do, D. D., Godoy, N., Nauck, M., Marz, W., Rupp, J., Pech, M., and Luscher, T. F. Human connective tissue growth factor is expressed in advanced atherosclerotic lesions. *Circulation* **1997**; 95(4):831-839.
- Ohyama, H., Zhang, X., Kohno, Y., Alevizos, I., Posner, M., Wong, D. T., and Todd, R. Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization. *Biotechniques* **2000**; 29(3):530-536.
- O'Keefe, R. J., Crabb, I. D., Puzas, J. E., and Rosier, R. N. Countercurrent centrifugal elutriation. High-resolution method for the separation of growth-plate chondrocytes. *Journal of Bone and Joint Surgery. American Volume* **1989**; 71(4):607-620.
- Ormerod, M. G. Analysis of DNA — general methods. A practical approach in Flow cytometry. 2nd edition. Ormerod, M. G., editor. IRL Press, Oxford, United Kingdom **1995**.
- Piper, H. M. Cell culture techniques in heart and vessel research. Piper, H. M., editor. Springer-Verlag, Berlin, Germany **1990**.
- Puri, R. K., Mehrotra, P. T., Leland, P., Kreitman, R. J., Siegel, J. P., and Pastan, I. A chimeric protein comprised of IL-4 and *Pseudomonas* exotoxin is cytotoxic for activated human lymphocytes. *Journal of Immunology* **1994**; 152(7):3693-3700.

- Quadroni, M., and James, P. Proteomics and automation. *Electrophoresis* **1999**; 20(4-5): 664-677.
- QIAGEN GmbH. QIAGEN RNA/DNA Handbook. QIAGEN GmbH, Hilden, Germany **1998**.
- QIAGEN GmbH. Oligotex Handbook, 1st edition. QIAGEN GmbH, Hilden, Germany **1999**.
- QIAGEN GmbH. RNeasy Protect Kits – the complete solution for stabilisation and isolation of intact RNA. *QIAGEN News* **2000**; 2000(5):1, 18-19.
- QIAGEN GmbH. RNeasy Mini Handbook, 3rd edition. QIAGEN GmbH, Hilden, Germany **2001a**.
- QIAGEN GmbH. RNeasy Midi/Maxi Handbook, 2nd edition. QIAGEN GmbH, Hilden, Germany **2001b**.
- QIAGEN GmbH. Bacterial RNA stabilisation and isolation with RNeasy Protect Bacteria Kits. *QIAGEN News* **2001c**; 2001(5):1, 18-19.
- QIAGEN GmbH. RNeasy 96 Handbook. QIAGEN GmbH, Hilden, Germany **2002**.
- Pageot, L. P., Perreault, N., Basora, N., Francoeur, C., Magny, P., and Beaulieu, J. F. Human cell models to study small intestinal functions: recapitulation of the crypt-villus axis. *Microscopy Research and Technique* **2000**; 49(4):394-406.
- Peale, F. V., Jr., and Gerritsen, M. E. Gene profiling techniques and their application in angiogenesis and vascular development. *Journal of Pathology* **2001**; 195(1):7-19.
- Perren, A., Roth, J., Muletta-Feurer, S., Saremaslani, P., Speel, E. J., Heitz, P. U., and Komminoth, P. Clonal analysis of sporadic pancreatic endocrine tumours. *Journal of Pathology* **1998**; 186(4):363-371.
- Phelps, R. M., Johnson, B. E., Ihde, D. C., Gazdar, A. F., Carbone, D. P., McClintock, P. R., Linnoila, R. I., Matthews, M. J., Bunn, P. A., Jr., Carney, D., Minna, J. D., and Mulshine, J. L. NCI-Navy Medical Oncology Branch cell line data base. *Journal of Cellular Biochemistry Supplement* **1996**; 24:32-91.
- Phillips, J., and Eberwine, J. H. Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells. *Methods* **1996**; 10(3): 283-288.
- Pillai, R. Oncogene expression and prognosis in cervical cancer. *Cancer Letters* **1991**; 59(2):171-175.
- Polyak, K., and Riggins, G. J. Gene discovery using the serial analysis of gene expression technique: implications for cancer research. *Journal of Clinical Oncology* **2001**; 19(11):2948-2958.
- Rabilloud, T. Solubilization of proteins for electrophoretic analyses. *Electrophoresis* **1996**; 17(5):813-829.
- Rabinowitz, Y. Separation of lymphocytes, polymorphonuclear leukocytes and monocytes on glass columns, including tissue culture observations. *Blood* **1964**; 23(6):811-828.
- Rapley, R. In: RNA isolation and characterization protocols. Rapley, R., editor. Humana Press, Totowa, New York, USA **1998**.
- Rayleigh, Lord, J. W. S. On the instability of jets. *Proceedings of the London Mathematical Society* **1878**; 10(4):4-13.
- Ringwald, M., Baldock, R., Bard, J., Kaufman, M., Eppig, J. T., Richardson, J. E., Nadeau, J. H., and Davidson, D. A database for mouse development. *Science* **1994**; 265(5181): 2033-2034.
- Ringwald, M., Davis, G. L., Smith, A. G., Trepanier, L. E., Begley, D. A., Richardson, J. E., and Eppig, J. T. The mouse gene expression database GXD. *Seminars in Cell and Developmental Biology* **1997**; 8(5):489-497.
- Ringwald, M., Mangan, M. E., Eppig, J. T., Kadin, J. A., and Richardson, J. E. GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. *Nucleic Acids Research* **1999**; 27(1):106-112.
- Ringwald, M., Eppig, J. T., Kadin, J. A., and Richardson, J. E. GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. The Gene Expression Database group. *Nucleic Acids Research* **2000**; 28(1):115-119.
- Ringwald, M., Eppig, J. T., and Richardson, J. E. GXD: integrated access to gene expression data for the laboratory mouse. *Trends in Genetics* **2000**; 16(4):188-190.
- Ringwald, M., Eppig, J. T., Begley, D. A., Corradi, J. P., McCright, I. J., Hayamizu, T. F., Hill, D. P., Kadin, J. A., and Richardson, J. E. The Mouse Gene Expression Database (GXG). *Nucleic Acids Research* **2001**; 29(1):98-101.
- Rivera-Marrero, C. A., Burroughs, M. A., Masse, R. A., Vannberg, F. O., Leimbach, D. L., Roman, J., and Murtagh, J. J., Jr. Identification of genes differentially expressed in *Mycobacterium tuberculosis* by differential

- display PCR. *Microbial Pathogenesis* **1998**; 25(6):307-316.
- Roederer, M. Compensation. Current protocols in cytometry. Robinson, J. P., Darzynkiewicz, Z., Dean, P. N., Dressler, L. G., Rabinovitch, P. S., Stewart, C. C., Tanke, H. J., and Wheelless, L. L., editors. John Wiley & Sons, New York, New York, USA **1999**.
- Rosson, D., and Tereba, A. Transcription of hematopoietic-associated oncogenes in childhood leukemia. *Cancer Research* **1983**; 43(8):3912-3918.
- Rothstein, F. Differential precipitation of proteins. Science and technology. *Bioprocess Technology* **1994**; 18:115-208.
- Samaha, F. F., Ip, H. S., Morrissey, E. E., Seltzer, J., Tang, Z., Solway, J., and Parmacek, M. S. Developmental pattern of expression and genomic organization of the calponin-h1 gene. A contractile smooth muscle cell marker. *Journal of Biological Chemistry* **1996**; 271(1):395-403.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. Molecular cloning: a laboratory manual, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA **1989**.
- Sanchez-Madrid, F., Simon, P., Thompson, S., and Springer, T. A. Mapping of antigenic and functional epitopes on the alpha- and beta-subunits of two related mouse glycoproteins involved in cell interactions, LFA-1 and Mac-1. *Journal of Experimental Medicine* **1983**; 158(2):586-602.
- Sastre, L., Kishimoto, T. K., Gee, C., Roberts, T., and Springer, T. A. The mouse leukocyte adhesion proteins Mac-1 and LFA-1: studies on mRNA translation and protein glycosylation with emphasis on Mac-1. *Journal of Immunology* **1986**; 137(3):1060-1065.
- Scheidl, S. J., Nilsson, S., Kalén, M., Hellström, M., Takemoto, M., Håkansson, J., and Lindahl, P. mRNA expression profiling of laser microbeam microdissected cells from slender embryonic structures. *American Journal of Pathology* **2002**; 160(3):801-813.
- Schütze, K., and Lahr, G. Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnology* **1998**; 16(8):737-742.
- Schweinfest, C. W., Kwiatkowski, R. W., and Dotti, R. P. Molecular cloning of a DNA sequence complementary to creatine kinase M mRNA from chickens. *Proceedings of the National Academy of Sciences of the United States of America* **1982**; 79(16):4997-5000.
- Scopes, R. K. Protein Purification, Principles and Praxis. 3rd edition. Springer-Verlag, Berlin, Germany **1994**.
- Scopes, R. K., and Smith, J. A. Analysis of proteins. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., eds. John Wiley & Sons, New York, USA **1994**:10.0.1-10.0.20.
- Scouten, W. H., and Konecny, P. Reversible immobilization of antibodies on magnetic beads. *Analytical Biochemistry* **1992**; 205(2):313-318.
- Shapiro, H. M. Practical flow cytometry. 3rd edition. Wiley-Liss, New York, New York, USA **1995**:43.
- Shibusawa, Y. Surface affinity chromatography of human peripheral blood cells. *Journal of Chromatography B, Biomedical Sciences and Applications* **1999**; 722(1-2):71-88.
- Shibusawa, Y. Separation and retention of human blood cells by surface-affinity chromatography. *Journal of Biochemical and Biophysical Methods* **2001**; 49(1-3):683-703.
- Shiffman, D., Mikita, T., Tai, J. T., Wade, D. P., Porter, J. G., Seilhamer, J. J., Somogyi, R., Liang, S., and Lawn, R. M. Large scale gene expression analysis of cholesterol-loaded macrophages. *Journal of Biological Chemistry* **2000**; 275(48):37324-37332.
- Simon, M. A., Drees, B., Kornberg, T., and Bishop, J. M. The nucleotide sequence and the tissue-specific expression of *Drosophila* c-src. *Cell* **1985**; 42(3):831-840.
- Sobue, K., Hayashi, K., and Nishida, W. Expressional regulation of smooth muscle cell-specific genes in association with phenotypic modulation. *Molecular and Cellular Biochemistry* **1999**; 190(1-2):105-118.
- Springer, T. A. Monoclonal antibody analysis of complex biological systems. Combination of cell hybridization and immunoadsorbents in a novel cascade procedure and its application to the macrophage cell surface. *Journal of Biological Chemistry* **1981**; 256(8):3833-3839.
- Springer, T. A. Immunoprecipitation. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, USA **1994**:10.16.1-10.16.11.

- Srinivasan, R. Ablation of polymers and biological tissue by ultraviolet lasers. *Science* **1986**; 234(4776):559-565.
- Stacey, G. N., Bolton, B. J., and Doyle, A. The quality control of cell banks using DNA fingerprinting. *EXS* **1991**; 58:361-370.
- Stallcup, M. R., and Washington, L. D. Region-specific initiation of mouse mammary tumor virus RNA synthesis by endogenous RNA polymerase II in preparations of cell nuclei. *Journal of Biological Chemistry* **1983**; 258(5):2802-2804.
- Stern, D. B., and Newton, K. J. Isolation of plant mitochondrial RNA. *Methods in Enzymology* **1986**; 118:488-496.
- Stevenson, H. C. Isolation of human mononuclear leukocyte subsets by countercurrent centrifugal elutriation. *Methods in Enzymology* **1984**; 108:242-249.
- Stevenson, H. C., Katz, P., Wright, D. G., Contreras, T. J., Jemionek, J. F., Hartwig, V. M., Flor, W. J., and Fauci, A. S. Human blood monocytes: characterization of negatively selected human monocytes and their suspension cell culture derivatives. *Scandinavian Journal of Immunology* **1981**; 14(3):243-256.
- Stewart, C. C., and Stewart, S. J. Four color compensation. *Cytometry* **1999**; 38(4):161-175.
- Suzuki, Y., Makino, A., and Mae, T. An efficient method for extraction of RNA from rice leaves at different ages using benzyl chloride. *Journal of Experimental Botany* **2001**; 52(360):1575-1579.
- Tangen, O., Berman, H. J., and Marfey, P. Gel filtration. A new technique for separation of blood platelets from plasma. *Thrombosis et Diathesis Haemorrhagica* **1971**; 25(2):268-278.
- Taylor, W. G., and Evans, V. J. Preparation and use of replicate mammalian cell cultures. *Methods in Cell Biology* **1974**; 8(0):47-73.
- Traycoff, C. M., Kosak, S. T., Grigsby, S., and Srour, E. F. Evaluation of *ex vivo* expansion potential of cord blood and bone marrow hematopoietic progenitor cells using cell tracking and limiting dilution analysis. *Blood* **1995**; 85(8):2059-2068.
- Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B., and Pastan, I. Discovery of three new genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences of the United States of America* **1998**; 95(1):300-304.
- Van der Loop, F. T., Gabbiani, G., Kohnen, G., Ramaekers, F. C., and van Eys, G. J. Differentiation of smooth muscle cells in human blood vessels as defined by smoothelin, a novel marker for the contractile phenotype. *Arteriosclerosis, Thrombosis and Vascular Biology* **1997**; 17(4):665-671.
- Van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**; 415(6871):530-536.
- Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. Analysing uncharted transcriptsomes with SAGE. *Trends in Genetics* **2000**; 16(10):423-425.
- Vona, G., Sabile, A., Louha, M., Sitruk, V., Romana, S., Schütze, K., Capron, F., Franco, D., Pazzagli, M., Vekemans, M., Lacour, B., Bréchet, C., and Paterlini-Bréchet, P. Isolation by size of epithelial tumor cells. A new method for the immunomorphological and molecular characterization of circulating tumor cells. *American Journal of Pathology* **2000**; 156(1):57-63.
- Waghray, A., Feroze, F., Schober, M. S., Yao, F., Wood, C., Puravs, E., Krause, M., Hanash, S., and Chen, Y. Q. Identification of androgen-regulated genes in the prostate cancer cell line LNCaP by serial analysis of gene expression and proteomic analysis. *Proteomics* **2001**; 1(10):1327-1338.
- Walter, H., and Selby, F. W. Counter-current distribution of red blood cells of slightly different ages. *Biochimica et Biophysica Acta* **1966**; 112(1):146-153.
- Walter, H., and Selby, F. W. Effects of DEAE-dextran on the partition of red blood cells in aqueous dextran-polyethylene glycol two-phase systems. *Biochimica et Biophysica Acta* **1967**; 148(2):517-524.
- Walter, H., Selby, F. W., and Garza, R. On the countercurrent distribution of red blood cells: an addendum. *Biochimica et Biophysica Acta* **1967**; 136(1):148-150.
- Walter, H., Krob, E. J., and Ascher, G. S. Factors in the partition of red blood cells in aqueous dextran-polyethylene glycol two-phase systems. *Biochimica et Biophysica Acta* **1968**; 165(3):507-514.

- Walter, H., Krob, E. J., and Ascher, G. S. Partition and countercurrent distribution of erythrocytes and leukocytes from different species. *Experimental Cell Research* **1969**; *55*(1):57-64.
- Walter, H., Krob, E. J., and Ascher, G. S. Separation of lymphocytes and polymorphonuclear leukocytes by countercurrent distribution in aqueous two-polymer phase systems. *Experimental Cell Research* **1969**; *55*(2):279-283.
- Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T., and Marincola, F. M. High-fidelity mRNA amplification for gene profiling. *Nature Biotechnology* **2000**; *18*(4):457-459.
- Webb, M. B., and Debenham, P. G. Cell line characterisation by DNA fingerprinting; a review. *Developments in Biological Standardization* **1992**; *76*:39-42.
- Wendisch, V. F., Zimmer, D. P., Khodursky, A., Peter, B., Cozzarelli, N., and Kustu, S. Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Analytical Biochemistry* **2001**; *290*(2):205-213.
- Westphal, G., Burgemeister, R., Friedemann, G., Wellmann, A., Wernert, N., Wollscheid, V., Becker, B., Vogt, T., Knüchel, R., Stolz, W. and Schütze, K. Noncontact laser catapulting: A basic procedure for functional genomics and proteomics. *Methods in Enzymology* **2002**; *356*:80-99.
- Whetsell, L., Maw, G., Nadon, N., Ringer, D. P., and Schaefer, F. V. Polymerase chain reaction microanalysis of tumors from stained histological slides. *Oncogene* **1992**; *7*(11):2355-2361.
- Wigle, D. A., Rossant, J., and Jurisica, I. Mining mouse microarray data. *Genome Biology* **2001**; *2*(7):REVIEWS1019.
- Wilfinger, W. W., Mackey, M., and Chomczynski, P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques* **1997**; *22*(3):474-476, 478-481.
- Winding, P., and Berchtold, M. W. The chicken B cell line DT40: a novel tool for gene disruption experiments. *Journal of Immunological Methods* **2001**; *249*(1-2):1-16.
- Wong, M. H., Saam, J. R., Stappenbeck, T. S., Rexer, C. H., and Gordon, J. I. Genetic mosaic analysis based on Cre recombinase and navigated laser capture microdissection. *Proceedings of the National Academy of Sciences of the United States of America* **2000**; *97*(23):12601-12606.
- Worth, N. F., Rolfe, B. E., Song, J., and Campbell, G. R. Vascular smooth muscle cell phenotypic modulation in culture is associated with reorganisation of contractile and cytoskeletal proteins. *Cell Motility and the Cytoskeleton* **2001**; *49*(3):130-145.
- Yamashita, T., Kaneko, S., Hashimoto, S., Sato, T., Nagai, S., Toyoda, N., Suzuki, T., Kobayashi, K., and Matsushima, K. Serial analysis of gene expression in chronic hepatitis C and hepatocellular carcinoma. *Biochemical and Biophysical Research Communications* **2001**; *282*(2):647-654.
- Yasaka, T., Mantich, N. M., Boxer, L. A., and Baehner, R. L. Functions of human monocyte and lymphocyte subsets obtained by countercurrent centrifugal elutriation: differing functional capacities of human monocyte subsets. *Journal of Immunology* **1981**; *127*(4):1515-1518.
- Ye, F., Abel, W. O., and Reski, R. Isolation of plant mitochondrial RNA from green leaves. *Methods in Molecular Biology* **1998**; *86*:33-38.
- Zhong, H., and Minneman, K. P. Use and pharmacological analysis of established and transfected cell lines expressing adrenergic receptors. *Methods in Molecular Biology* **2000**; *126*:221-234.
- Zhuang, Z., Bertheau, P., Emmert-Buck, M. R., Liotta, L. A., Gnarr, J., Linehan W. M., and Lubensky, I. A. A microdissection technique for archival DNA analysis of specific cell populations in lesions <1mm in size. *American Journal of Pathology* **1995**; *146*(3):620-625.

3

Methods for analysing mRNA expression

3.1

Introduction

By and large, the study of gene expression involves the comparison of mRNA populations between two samples, treated *versus* untreated, diseased *versus* healthy, stage A of development *versus* stage B of development. This chapter is the first in the book to specifically describe methods for this purpose.

We have divided the chapter into three blocks, the first devoted to hybridisation-based methods, the second describing PCR-based methods and the third outlining methods that do not fall into the other two categories. In contrast to chapter 4, which deals almost exclusively with high-throughput applications, and which is therefore aimed more towards commercial use, the present chapter describes methods which should be within the resources of most reasonably equipped research laboratories.

Despite the modest requirements of the techniques in the present chapter, these methods are in no way inferior from a technological point of view to those in subsequent chapters. Indeed, it is likely that some of the methods described here such as the branched DNA assay or real-time RT-PCR will soon be scaled up to industrial dimensions.

Furthermore, we do not wish to suggest a hierarchy of usefulness of the methods within the chapter. To reflect this, the subsections have been arranged in alphabetical order. We hope that each contribution contains enough information on the relative advantages and disadvantages of the technique described to allow the reader to make up his own mind on which technique is best for him.

Finally, it should be remembered that many, if not most, expression studies require the use of a combination of techniques. Thus, a subtractive hybridisation step is often followed by a PCR amplification stage in order to display the cDNAs which have not been eliminated during hybridisation. Equally, differentially PCR-amplified products may be screened using hybridisation techniques. In addition, more general techniques such as restriction digestion, *in vitro* transcription, DNA sequencing or polyacrylamide-based gel electrophoresis form an integral part of nearly all expression studies.

3.2

Hybridisation-based methods

3.2.1

Branched DNA assay

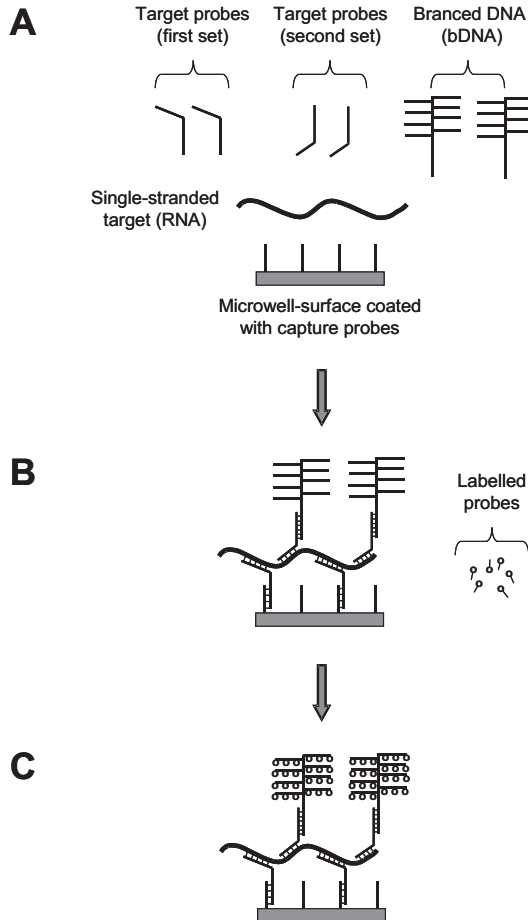
3.2.1.1 Introduction

Target amplification strategies based on polymerase chain reaction (PCR) methods are widely used for nucleic acid detection. An alternative method is signal amplification of the nucleic acid. One of such method is the branched DNA (bDNA) technology, involving capture and immobilisation of the target sequence and its detection *via* a comb-like DNA structure – the branched DNA. Since the signal amplification is based on the special structure of the bDNA, the synthesis of branched DNA molecules is an essential component of this technique (Horn & Urdea, 1989). The bDNA technology was designed to detect nucleic acid sequences *in vitro* in solution (Urdea *et al.*, 1987 and 1989). Later on, the technique was refined and commercialised by the Chiron Corporation (Emeryville, California, USA). The first and second-generation bDNA assays lacked sensitivity compared with the target amplifications systems. The changes incorporated into the third-generation assays have effectively increased the signal-to-noise ratio to such a high level that the analytical sensitivity approaches that of PCR (Nolte, 1998).

3.2.1.2 Principle and basics

Immobilisation and detection of the target nucleic acid is based on its hybridisation with two different kinds of oligonucleotides containing at one end sequences complementary to short fragments of the target. The other end is complementary to either the so called ‘capture probe’ or the bDNA. The first type of oligonucleotide (first target probe) connects the target sequence to the capture probes, which are in turn attached by one of their ends to the surface of a microwell plate. Thus, the first target probe acts like a special kind of glue by binding to the target nucleic acid with one end and to the capture probe with the overhang. The second set of oligonucleotides (second target probe) binds to different sequences of the target. Again, these probes have specific overhangs, although in this case they bind the bDNA (figure 3.1). For detection, labelled oligonucleotides (labelled probes) are used that bind to the branches of the bDNA. Since the bDNA contains a large number of binding sites for the labelled probe, like a Christmas tree decorated with many candles, the signal is greatly amplified compared to direct detection systems. Thus, the bDNA technology functions like an enzyme linked immunosorbent assay (ELISA) for nucleic acids. Alkaline phosphatase which catalyses the chemiluminescent reaction of the substrate dioxetane is often used for labelling, but other kinds of detection such as fluorescence labelling are imaginable. Refining of the method by including non-natural bases in the amplification molecules and designing a pre-amplifier molecule, has improved sensitivity and enabled the detection of fewer than 100 nucleic acid molecules per millilitre of solution (Collins *et al.*, 1997).

Figure 3.1 Principle of the branched DNA assay. **A)** The target sequence is immobilised by the capture probes *via* the first set of target probes. Branched DNAs are then attached *via* the second set of target probes. **B)** Small labelled probes are hybridised to complementary sequences in the branched DNA amplifier to build up a large number of enzyme catalytic sites (alkaline phosphatase) or fluorescence molecules. **C)** Target detection by reaction of chemiluminescence substrate (*e.g.*, dioxetane) or fluorescence.



3.2.1.3 Examples

Reported applications of the bDNA assay include the detection and quantification of the human immunodeficiency virus, or HIV (Rouet *et al.*, 2001), hepatitis viruses C, or HCV (Pawlotsky *et al.*, 1999) and hepatitis viruses B, or HBV (Chu *et al.*, 2001), cytomegalovirus (Flood *et al.*, 1997), and the plaque agent *Yersinia pestis* (Iqbal *et al.*, 1999). A further improvement to bDNA technology is the branched DNA *in situ* hybridisation (bDNA-ISH) method, which uses fluorescently labelled probes for detection. It was developed for quantification of human papillomavirus (HPV) DNA in whole cells (Roth *et al.*, 1998). By using human cervical cancer cell lines with known copies of human papillomavirus DNA, it was shown that the bDNA-ISH method is highly sensitive, detecting as few as one or two copies of human papillomavirus DNA per cell. Furthermore, it was shown that the bDNA-ISH method provides precise localisation of the target nucleic acid within the

cell. The bDNA-ISH method is therefore a powerful tool for monitoring, for example, RNA distribution in cells or tissues.

3.2.1.4 Discussion

Since the target sequence is detected directly, it is not necessary to perform reversed transcription polymerase chain reaction (RT-PCR) for the quantification of RNA. In contrast to PCR – based amplification strategies, the original amount of the target nucleic acid remains constant. Therefore, this is a true quantitative procedure with high specificity. Many groups have compared the bDNA assay with RT-PCR for diagnostic purposes, especially for detection of viral RNA (*e.g.*, hepatitis B and C virus, human papillomavirus or human immunodeficiency virus) (Debono *et al.*, 2000; Hwang *et al.*, 1999; Roth *et al.*, 1998; Won *et al.*, 2001). Based on their findings, it can be concluded that both methods provide comparable results. Branched DNA technology appears to have some advantages over PCR-based methods, including better tolerance of target sequence variability, more direct measurement of the target, less complex sample preparation, and less sample-to-sample variation (Nolte, 1998). By analysing the differential expression of cytochrome P450 in rats, it was found that the bDNA signal amplification resulted in a linear quantifiable range of RNA detection that spanned three orders of magnitude (0.1 to 100 micrograms of total RNA). The fidelity of the bDNA assay was evaluated within a single assay and between assays and was shown to be reproducible (Hartley & Klaassen, 2000). A further advantage of the bDNA assay compared to RT-PCR is that it may be used in biological samples such as seminal fluid that contain inhibitors for the PCR reaction (Debono *et al.*, 2000). Quantification of RNA in such a sample by PCR requires prior RNA purification whereas bDNA analysis can be performed directly.

Adaptation of the branched DNA signal amplification technology to flow cytometry has resulted in a quantitative nucleic acid assay with significant advantages over the microwell-based format. In this assay, microbeads, rather than microwell plates, are derivatised with nucleic acid capture probes and the derivatised beads are used to capture single nucleic acid targets, which then capture fluorescent reporter probes *via* branched DNA. The assay detects DNA or RNA targets with a current lower sensitivity limit of 500 human immunodeficiency virus RNA molecules and responds linearly to the target level from 500 to at least 50,000 molecules. Since microbeads can easily interrogate large volumes, viral lysis and genomic DNA capture can proceed in one step from comparatively large volumes, and sample preparation is greatly simplified compared to the microwell-format bDNA assay (van Cleve *et al.*, 1998). Despite these features, the sensitivity of the bDNA technology remains a limiting factor. One group, for example, compared quantitative real-time RT-PCR and bDNA assay detecting human hepatitis C virus RNA in serum of 50 patients. Hepatitis C virus RNA was detectable in all 50 samples by the use of quantitative real-time RT-PCR, but was undetectable in 14 samples by the use of a branched DNA assay (Enomoto *et al.*, 2001). For reviews on bDNA technology, see Nolte (1998) and Andras *et al.* (2001).

3.2.2

Northern blot and related techniques**3.2.2.1 Introduction**

Similar to the blotting and hybridisation technique developed by Edward Southern in 1975 for the detection of DNA, methods exist for detection of specific sequences in RNA preparations. These are Northern blotting and slot or dot blot hybridisation. The difference to Southern blotting lies mainly in the initial gel fractionation step, as RNAs are usually single strand molecules which are able to form secondary structures *via* intramolecular base pairing. For this reason, electrophoresis must be performed under denaturing conditions using either denaturing formaldehyde gels (formaldehyde reacts with the primary amino groups of the RNA bases adenine, guanine and thymine to form Schiff bases, thus preventing intramolecular base pairing; Boedtke *et al.*, 1967; Fraenkel-Conrat, 1954; Haselkorn *et al.*, 1961; Hoard, 1960) and loading buffers, glyoxal and dimethylsulfoxide treatment of the RNA or methylmercury (Brown & Mackey, 1997; Bailey & Davidson, 1976; Thomas, 1980).

For hybridisation, the complementarity between a probe molecule and a target molecule can result in the formation of a probe/target complex, also called hybrid. In case of the Northern blot technique, hybrids are formed between the single-stranded RNA target on the membrane and a labelled complementary DNA probe or between the RNA target strand and a second complementary RNA molecule.

3.2.2.2 Principles and basics

Northern blot analysis is performed in three steps. First, the RNA to be analysed is electrophoresed under denaturing conditions in an agarose/formaldehyde gel. Second, the fractionated RNA is transferred from the denaturing agarose gel to a nitrocellulose or nylon membrane by capillary or vacuum transfer. Third, the RNA sequences of interest are analysed by hybridisation using a labelled DNA or RNA probe. In dot or slot blotting only steps two and three are performed. The principle of Northern blot analysis is depicted in figure 3.2.

3.2.2.2.1 Agarose/formaldehyde gel electrophoresis

Gel electrophoresis is a technique that separates molecules largely on the basis of their size. Two main types of gel exist using either agarose or polyacrylamide. Both are solid and consist of a matrix of molecules forming sub-microscopic pores. The size of the pores can be controlled by varying the chemical composition of the gel. The gel is immersed in buffer and subjected to an electric field. Under appropriate pH and buffer conditions, the different RNA species will separate as they move in the electric field from the negative to the positive pole. Larger molecules are retarded more than smaller molecules with separation of RNA species by size. It is important to recall that size does not necessarily correspond to molecular weight.

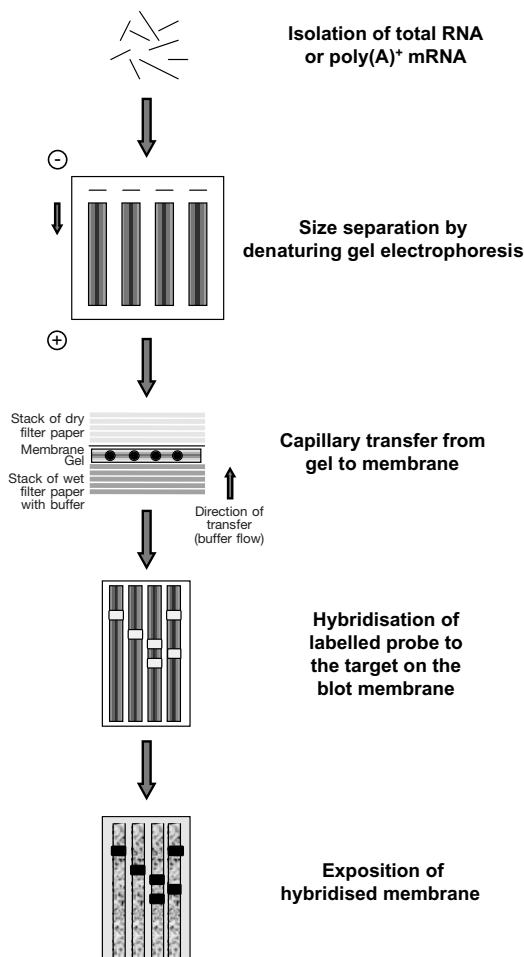


Figure 3.2 Schematic illustration of the Northern blot principle. In the first step, total RNA or poly(A)⁺ mRNA is isolated from the cell(s) or tissue(s) of interest. RNA samples are then size-fractionated using denaturing formaldehyde/agarose gels. Following electrophoresis, samples are transferred from the gel to a

membrane usually using capillary transfer. The membrane is hybridised with a labelled DNA or RNA probe complementary to the RNA of interest. After washing, the RNA of interest is detected using either auto-radiography or chemiluminescence.

For Northern blots, denaturing formaldehyde/agarose gels are usually used, although other methods have been described (Brown & Mackey, 1997; Bailey & Davidson, 1976; Thomas, 1980). To prevent formation of secondary structures, the RNA is also treated with formaldehyde. Staining is generally by means of ethidium bromide which binds to nucleic acids, and is performed to assess the quality of the RNA and the gel separation.

3.2.2.2.2 Transfer of the RNA

To facilitate hybridisation, the RNA is then transferred from the gel to a solid support. This transfer process is called blotting. Usually, a nitrocellulose or nylon membrane is used as the solid support.

Different ways of transfer exist such as electroblotting (Smith *et al.*, 1984) which takes the advantage of the molecules negative charge or capillary blotting which is often supported by a vacuum (Peferoen *et al.*, 1982). Following transfer, the RNA is often fixed to the filter by baking or by exposure to ultraviolet light (cross-linking), thus enhancing the sensitivity of the procedure ten to 40-fold due to a reduced loss of sample during washing and hybridisation (Khandjian, 1986). An indication of the efficiency of transfer of the RNA can be obtained by reversible staining the membrane with the nucleic acids intercalating reagent methylene blue (Antony *et al.*, 1987; Barabas *et al.*, 1980; Herrin *et al.*, 1988).

3.2.2.2.3 Hybridisation

Firstly, the membranes containing the blotted RNA are blocked to prevent hybridisation probes adhering to the blank areas of membrane. A blocking solution containing high concentrations of non-specific RNA is used to coat the filter and reduce background activity.

The most usual variety of probe is a double-stranded fragment of cDNA derived by restriction digestion of a plasmid containing the gene of interest. The probe is labelled using random hexamers (McCreery & Helentjaris, 1994; Rost *et al.*, 1995) as follows: the DNA is denatured by boiling, a mixture of DNA hexamers is added to the denatured probe and allowed to base-pair. DNA polymerase is then added together with dATP, dGTP, dTTP and radioactive [α - ^{32}]-dCTP. The mixture is again boiled to separate the strands and is then ready for hybridisation.

Alternatively, labelled nucleotides may be incorporated into the probe using nick translation (Kath *et al.*, 1993), cDNA synthesis (Soldo *et al.*, 1992), PCR amplification (An *et al.*, 1992; Chenal & Griffais, 1994; Emanuel, 1991; Gissing *et al.*, 1997; Holmstrom *et al.*, 1993; Lanzillo, 1990; Reischl *et al.*, 1993 and 1994), *in vitro* transcription (Höltke & Kessler, 1990), 3'-end labelling (Rosemeyer *et al.*, 1995), 3'-tailing (Schmitz *et al.*, 1991), or chemical synthesis (Bürge *et al.*, 1992).

Several alternatives have been developed to avoid the use of radioactivity for nucleic acid labelling. Non-radioactive labelling techniques use the same techniques as mentioned above to incorporate (ribo)nucleotides that are modified with a digoxigenin or biotin group. Using the latter kind of label, it is possible to detect the hybridised probe using digoxigenin-specific antibodies or streptavidin molecules that are conjugated with an enzyme such as the widely used alkaline phosphatase. Such enzyme conjugates allow non-radioactive detection using, for example, chemiluminescence substrates that emit light following enzymatic dephosphorylation.

The labelled probe is normally added to the blocked membranes in buffer and incubated for several hours. This allows the probe to hybridise to its target. When hybrids have formed between the RNA and the probe, it is necessary to remove all unbound material by extensive washing.

3.2.2.2.4 Detection of the probe/target hybrids

The sheet of membrane is now covered with spots of radiation or antibody-conjugated probe material wherever the probe molecules have hybridised to their target RNAs. If a radioactive label is used, the membrane is exposed to X-ray film for detection of hybridisation. If an antibody/enzyme conjugate is used, detection is performed by soaking the membrane in a solution of chromogenic or chemiluminescent substrate for the enzyme, which produces a colour staining or emits light in the regions of hybridisation.

3.2.2.3 Examples and typical results

Since its publication in 1977, Northern blotting has become the most widely used technique for analysing mRNA expression. Its impact in molecular biology is demonstrated by the more than 45,000 citations revealed by a PubMed search using the phrase "Northern blot".

Several examples for Northern blot analyses are shown in the different sections of this book. Nevertheless, because of the importance of this technique, two Northern blots are shown here (figures 3.3 and 3.4). Northern blotting can be used to

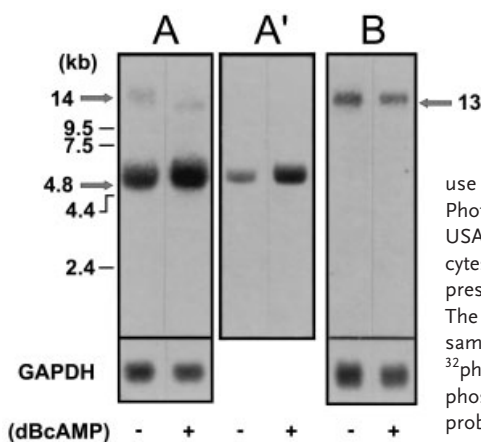
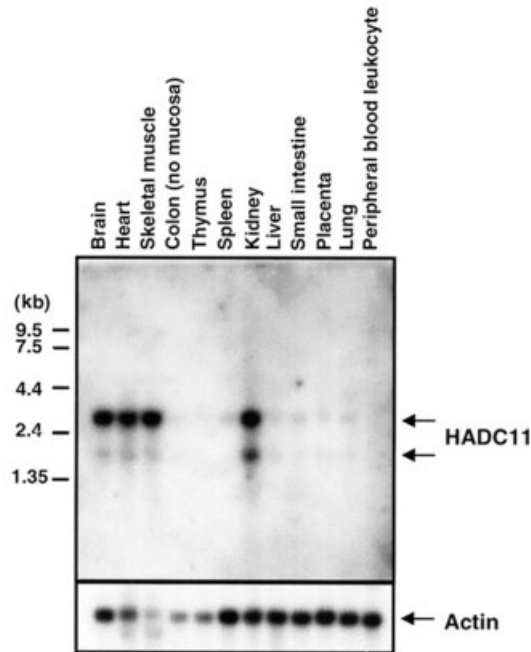


Figure 3.3 Analysis of differential mRNA expression using Northern blot analysis. The effect of dibutyl-cAMP (dBcAMP) on dystrophin, Dp71, and utrophin mRNA expression in cultured rat brain astrocytes. Five micrograms of poly(A)-enriched RNA from rat cultured astrocytes were separated on an 1.1 percent formaldehyde/agarose gel, transferred to a nylon membrane, and hybridised with ^{32} phosphorus-labelled rat dystrophin (panels A and A') PCR product (two kilobases in size) and utrophin (panel B) cDNA (two kilobases in size). The hybridisation signals (arrows) were visualised following auto-radiography for 18 hours (panels A and B) and two hours (panel A') by

use of a FUJIX BAS2000 imaging system (Fuji Photo Film USA, Inc., Elmsford, New York, USA). RNA samples were derived from astrocytes cultured for six days in the absence (-) or presence (+) of one millimolar dibutyl-cAMP. The filters used for panels A and A' were the same. The membrane was re-hybridised with a ^{32} phosphorus-labelled human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) cDNA probe to correct for the amount of RNA loaded in each lane. The sizes of the bands, 4.8 kilobases, 14 kilobases, and 13 kilobases, correspond to those reported for Dp71 isoform of dystrophin, full-length dystrophin, and utrophin mRNAs, respectively. Following culture in the presence of dibutyl-cAMP for six days, the level of the 4.8-kilobases dystrophin transcript isoform in astrocytes increased (about two-fold), whereas that of the 14-kilobases dystrophin transcript isoform was reduced to an undetectable level (panels A and A'). Expression of utrophin mRNA was also decreased by incubation with dibutyl-cAMP (panel B). The figure is reproduced from Imamura & Ozawa (1998) with permission from *Proceedings of the National Academy of Sciences of the USA*.

Figure 3.4 Analysis of mRNA expression distribution in tissues using multiple tissue Northern blots.

The tissue distribution of histone deacetylase family member 11 (HDAC11) mRNA in human samples is shown in this figure. The blot containing mRNAs from the indicated tissues was probed with a ^{32} phosphorus-labelled HDAC11 cDNA probe ((about 940 base pairs; top panel), stripped, and re-probed with a ^{32} phosphorus-labelled commercially available actin cDNA probe (BD Biosciences Clontech, Palo Alto, California, USA; bottom panel). The Northern blot analyses indicate that the overall expression level of HDAC11 mRNA is low in many human tissues and that the high expression level is limited to brain, heart, skeletal muscle, and kidney. The figure is reproduced from Gao *et al.* (2002) with permission from *Journal of Biological Chemistry*.



analyse differential expression of mRNA semiquantitatively. Usually, a second probe directed against a non-regulated housekeeping gene is used for normalisation. Using densitometry, accurate analysis is possible. Figure 3.4 shows a multiple tissue Northern blot. Such blots are used to characterise the tissue distribution of a transcript, *i.e.*, the identification of tissues or cells in which the mRNA of interest is expressed and to identify alternatively-spliced transcripts.

3.2.2.4 RNA dot blot and slot blot

This method is similar to Northern blotting with exception that the RNA is not separated on an denaturing agarose gel. Instead, 'dots' or 'slots' of non-fractionated RNA are applied to the membrane and examined as described above in order to determine the relative abundance of target mRNA sequences in the blotted samples. Dot and slot blots are less time-consuming than Northern blotting. Furthermore, partially degraded RNA can also be used with good semi-quantitative results. However, false-positive results are common and the method requires careful use of appropriate controls because it is difficult to distinguish between specific and non-specific hybridisation.

3.2.2.5 Summary and discussion

The procedure for the immobilisation of gel-fractionated RNA as an equivalent method to DNA blotting was developed only two years after the development of Southern blotting in 1975 and was called Northern blotting in deference to Edward Southern (Alwine *et al.*, 1977). Subsequently, the term Western blotting was applied to the membrane transfer of proteins. Since then, Northern hybridisation has become a standard procedure for the identification, and size analysis of RNA transcripts (including the identification of alternatively spliced transcripts). Furthermore, RNA slot blot analyses are frequently used to assess the expression profiles of tissue-specific genes (Kafatos *et al.*, 1979).

Nonetheless, critical parameters should not be overlooked. As single-stranded RNA can form secondary structures, the samples must be electrophoresed under denaturing conditions in order to ensure good separation. Most frequently, formaldehyde gels are used as they are easy to run and reasonably reliable. To remove the formaldehyde, the gels need to be rinsed prior to transfer, a minor inconvenience compared to other methods where more poisonous substances are used to denature the RNA.

Northern blotting can be applied to both total cellular RNA and poly(A)⁺ mRNA. However, total RNA can be less satisfactory because of non-specific hybridisation in the ribosomal RNA size fraction. Therefore, any hybridising signals in the vicinity of ribosomal RNA should be treated with suspicion and should be confirmed by blotting with poly(A)⁺ mRNA (Brown & Mackey, 1997).

The average effective detection limit with an overnight exposure is approximately five picograms of RNA. An mRNA is usually considered to be abundant if it constitutes more than one percent of the mRNA fraction (Brown & Mackey, 1997). In mammalian cells, the mRNA fraction makes up about 0.5 percent of total RNA, so five picograms of an abundant mRNA should be present in 100 nanograms of total RNA. If ten micrograms of total RNA are transferred, abundant mRNAs should give strong hybridisation signals, while less abundant species should be detectable following overnight exposure (Brown & Mackey, 1997). For rarer molecules, poly(A)⁺ mRNA must be used.

Although easy to perform, slot blots are a problematic technique as numbers of criteria must be satisfied if meaningful comparisons are to be made between mRNA abundance in different extracts (Brown & Mackey, 1997). Firstly, equal amounts of RNA must be loaded in each slot, which in practice is difficult to achieve, especially if RNA concentrations are estimated by absorbance spectroscopy. Errors arise if small quantities of RNA are used, and may also be caused by the presence of contaminating DNA and proteins. A second point to be considered is the choice of membrane. Nylon seems to have a greater tensile strength and allows ultraviolet cross-linking of RNA. Therefore, most transfers are now carried out using nylon rather than nitrocellulose membranes. Nylon has the added advantage that it is able to withstand the highly stringent conditions that may be required during the step of hybridisation with an RNA probe whereas nitrocellulose tends to disintegrate under these conditions (Brown & Mackey, 1997).

The analysis of hybridisation can also be influenced by different factors. Factors that influence specificity and sensitivity are incubation times, hybridisation solutions, probe length, and the mechanics of hybridisation (Brown & Mackey, 1997). Two further points need to be added. Firstly, formamide is used in almost all RNA hybridisation solutions. The primary reason for this is to permit a lower hybridisation temperature minimising RNA degradation during the incubation procedure (Brown & Mackey, 1997). The second point concerns the stability of the hybrids formed between the immobilised RNA and the probes. RNA/RNA hybrids are more stable than RNA/DNA hybrids, which implies a higher stringency during hybridisation and washing for the RNA probes in comparison to the DNA probes.

A last point to be mentioned is the re-probing of the membrane. Nitrocellulose presents problems regarding membrane integrity as well as RNA retention. Ultra-violet cross-linkage of RNA to a nylon membrane presents therefore optimal conditions for Northern blot re-probing.

Although it is possible to analyse several transcripts on a single membrane by, for example, stripping and re-probing with up to half a dozen probes or simultaneous detection of several transcripts of different size with a mixture of multiple probes, Northern blotting is inappropriate for high-throughput analyses. Nevertheless, Northern blot analysis is the standard method for analysing gene expression at the RNA level. Therefore, Northern blotting is the method of choice for confirming differential mRNA expression identified by screening techniques such as differential display RT-PCR or DNA microarrays. Northern blotting also has the advantage of assessing the size of the mRNA and of identifying alternatively-spliced transcripts.

3.2.3

Nuclear run-on assay

3.2.3.1 Introduction

The nuclear run-on transcription method (also called 'nuclear run-off' assay) was developed in 1983 to characterise the transcriptional and post-transcriptional regulation of globin gene expression in murine erythroleukemia cells (Profous-Juchelka *et al.*, 1983). The method has been used extensively to study gene regulation mediated by hormones, toxins, tumour promoters, viral infections or during embryonic development. The method measures the relative rate of transcription of a specific gene and, thus, differs from those methods which quantify the steady state levels of specific mRNAs (*i.e.*, RNase protection assays, Northern blots, real-time RT-PCR assays or DNA microarrays). The nuclear run-on assay involves the isolation of nuclei from cultured cells or tissues followed by nascent nuclear mRNA chain elongation *in vitro* performed in the presence of radiolabelled nucleotides. After an RNA isolation step, the mRNA molecules of interest are detected either by dot blot hybridisation, primer extension, RNase protection or S1 nuclease protection assays.

3.2.3.2 Principle and basics

The term 'gene expression' summarises a complex process in which the information contained within the genome is translated into the measurable phenotype of the organism. Gene expression involves several sequential steps, including relaxation of the chromatin structure, binding of transcription factors to *cis*-acting elements located within the promoter or enhancer regions of the gene, transcription initiation, unwinding of the DNA double strand, transcript elongation, transcript processing, transcript polyadenylation, mRNA export from the nucleus, mRNA transport to appropriate sites of translation, mRNA translation, protein assembly and processing, protein transport, and protein activation or inhibition (for details, see chapter 1, and Calkhoven & Ab, 1996). A common and important means of controlling gene expression occurs at the level of transcription initiation. Therefore, the steady state concentration of a particular mRNA is often used as shorthand for its rate of gene transcription. However, this is a gross simplification. Apart from transcription initiation, many other factors are known to affect the concentrations of specific mRNAs in the cell. In fact, almost any of the above mentioned steps contributing to gene expression may in principle be subject to regulatory events. One prominent example for a mode of gene expression regulation not mediated at the level of transcription initiation is provided by genes that encode differentially or alternatively spliced gene products (Smith & Valcarcel, 2000). Expression of these genes is often regulated in a complex manner at the level of transcript processing, which changes during development of the organism and/or in a tissue-specific manner. Moreover, alternative use of multiple polyadenylation sites may affect the rate of mRNA export from the nucleus or its stability in the cytoplasm. Some genes, like the haemoglobin genes, are regulated *via* modulation of the efficiency by which their mRNAs are translated. The apolipoprotein B (apoB) gene encodes two overlapping gene products, apoB-100 and apoB-48, both of which are expressed from the same gene *via* a mechanism called 'mRNA editing'.

The relative proportion of the two apolipoproteins is regulated exclusively by the activity of the apoB-100 mRNA editing enzyme (called APOBEC1) which converts a CAA glutamine codon at nucleotide position 6,666 into the premature UAA translational stop codon of the apoB-48 encoding transcript. Finally, a wide range of post-translational mechanisms of gene expression control has been identified (Calkhoven & Ab, 1996; Chianale *et al.*, 1995; van Hoof & Green, 1997).

The nuclear run-on method provides a rather straightforward and relatively simple tool to measure specifically gene expression at a level that closely tracks initiation of gene transcription *per se* (Greenberg & Bender, 1997; Marzluff & Huang, 1984; Srivastava & Schonfeld, 1994). Thus, the results are expected to reflect the primary rate of transcription as opposed to later maturation processes (splicing, cytosolic transport, degradation).

The principle of the nuclear run-on assay is illustrated in figure 3.5. Briefly, cell nuclei are isolated and nascent transcripts are elongated with radioactive [α - ^{32} P]-UTP. After an RNA isolation step, the radiolabelled RNA of each sample is hybridised to DNA probes previously blotted onto a nitrocellulose or nylon membrane (dot blot). Detected signals of each membrane are then used to calculate the rela-

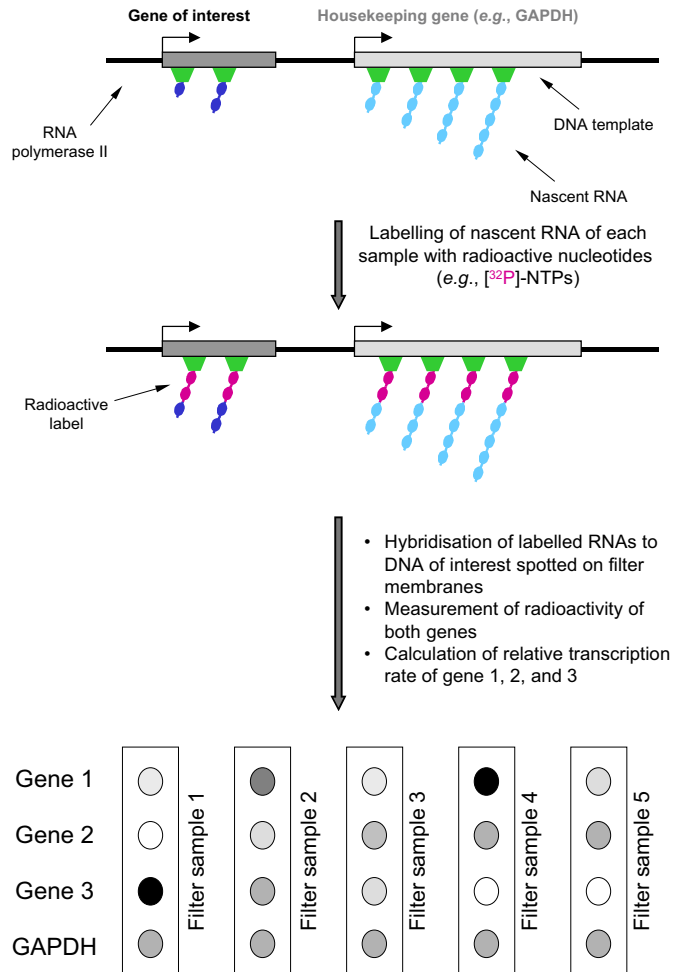


Figure 3.5 Principle of the nuclear run-off method. Cell nuclei are isolated and nascent transcripts are elongated with radioactive $[\alpha^{32}\text{P}]$ -UTP. After an RNA isolation step, the radiolabelled RNA of each sample is hybridised with cDNA probes previously blotted onto a nitrocellulose or nylon membrane (dot blot). After

detection of signals by auto-radiography, signals of each membrane are used to calculate the relative transcription rates of the genes of interest in each sample using the signals of a house keeping gene (e.g., G1 PDH) as an internal standard.

tive transcription rates of the genes of interest in each sample (compared to the transcription rate of a house-keeping gene, e.g., GA PDH) The basic steps of the nuclear run-on assay are: (i) Binding of probes (i.e., cDNA from a linearised plasmid) to the membrane. (ii) Isolation of nuclei from cells. (iii) Incubation of nuclei with $[\alpha^{32}\text{P}]$ -UTP and rNTPs in order to elongate the nascent transcripts. (iv) Purification of ^{32}P -labelled transcripts. (v) Hybridisation of ^{32}P -labelled transcripts to the immobilised cDNA probes.

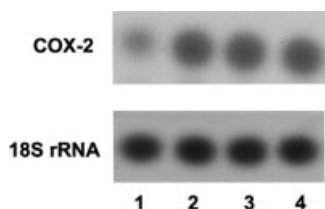


Figure 3.6 Stimulation of cyclooxygenase 2 (COX2; prostaglandin-endoperoxide synthase 2, PTGS2) gene transcription by microtubule interfering agents and cytochalasin D. Cells were treated with vehicle (lane 1) or the microtubule interfering agents taxol (ten micromolar, lane 2), cytochalasin D (ten micromolar, lane 3), or colchicine (ten micromolar, lane 4) for two hours. Nuclear run-offs were performed with 1.0×10^7 nuclei incubated in reaction buffer (ten millimolar Tris with pH 8.0, five millimolar magnesium chloride, and 0.3 molar potassium chloride) containing 100 microcuries of uridine $5' \text{-} [^3\text{P}]\text{-triphosphate}$ and one millimolar unlabeled

nucleotides. After 30 minutes, labelled nascent RNA transcripts were isolated. The human COX2 and 18S rRNA cDNAs were immobilised onto nitrocellulose and pre-hybridised overnight. Hybridisation was carried out at 42°C for 24 hours using equal counts per minute per millilitre labelled nascent RNA transcripts for each treatment group. The membranes were washed twice with $2\times$ SSC (buffer for one hour at 55°C), then treated with ten milligrams per millilitre RNase A in $2\times$ SSC (at 37°C for 30 minutes), dried, and detected by auto-radiography.

Figure 3.6 shows a result of a typical nuclear run-on experiment. Quantitative evaluation can be performed relative to the signal obtained for a (housekeeping) gene which is used as internal standard. In this case, 18S rRNA was used, but for quantification a housekeeping gene such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH) should be employed.

3.2.3.3 Discussion

Although the standard procedure works well in most instances, nuclear run-on transcription suffers from several drawbacks. One problem is the requirement for a large number of transcriptionally active nuclei, which are not always easily available. Moreover, despite the high amounts of radioactivity that are routinely employed in the assay, it normally requires several days of auto-radiography in order to obtain satisfactory signals. A further principal drawback is the necessity for the prior availability of cloned genes for analysis. Several modifications have been undertaken in order to improve the method.

One disadvantage of the ^{32}P -labelling method is the necessity to handle potentially hazardous radiolabelled RNA. To solve this problem, Merscher *et al.* (1994) developed a non-isotopic method employing digoxigenin-labelled UTP to label the chain of elongated RNA. The digoxigenin-labelled molecules are detected by chemiluminescence after hybridisation. The protocol provides results that are as sensitive and reproducible as with the ^{32}P -based method but yield lower background and require significantly reduced exposure times (one to six hours). Moreover, the digoxigenin-labelled RNA has a better long-term stability than ^{32}P -labelled RNA which degrades within days.

In 1997, Fiona G. Rolfe and William A. Sewell described a variation of the classical nuclear run-on assay by introducing RT-PCR in order to detect the overall transcript amount (Rolfe & Sewell, 1997). In this method, the run-on reaction is performed either with or without nucleotides. The over-all transcript amount is measured by RT-PCR. Transcription rates of individual genes are measured by comparing the PCR yields between samples incubated with and without nucleotides. One major difficulty users might have with this method is that nuclei contain a large amount of RNA which may differ between different nuclear preparations and may lead to inaccurate quantification.

Flow cytometric sorting of nuclei, followed by amplification of polyadenylated nuclear RNAs has been suggested as a quick and facile means for the analysis of transcripts (Macas *et al.*, 1998). Because the amplification is performed in a sequence-independent manner similar to amplified fragment length polymorphism (AFLP) (Bachem *et al.*, 1996) and differential display RT-PCR (Liang & Pardee, 1992), a wide variety of transcripts can be detected with high sensitivity. The technique requires relatively few nuclei and no radioactivity. In addition, it enables further analysis of individual transcripts through direct cloning and/or sequencing. Because the cloned transcripts provide expressed sequence tags, the method has been termed nuclear expressed sequence tag (NEST) analysis. In principle, NEST analysis is applicable to any eukaryotic organism or cell type that contains transcriptionally active nuclei. The ways in which flow cytometry and sorting can be combined with NEST analysis is discussed in the paper published by Macas *et al.* (1998).

The latest development of the nuclear run-on assay was presented by Patrone *et al.* (2000). The method tackles the problem that the chain elongated transcripts that are produced during the assay represent only a small fraction of the total nuclear RNA. In order to enrich for this RNA fraction, the *in vitro* transcription reaction is performed in the presence of biotin-16-UTP. The biotinylated RNA is then isolated by employing streptavidin-coated magnetic beads. Subsequently, the isolated RNA fraction is used for random hexamer primed reverse transcription. Finally, a semi-quantitative fluorescent PCR approach is used to determine transcript abundance in the resulting cDNA.

The combination of biotin labelling, magnetic bead capture and analysis by fluorescent-based RT-PCR offers some advantages. Since the purified RNA can be stored for long time periods, it can be used for several independent RT-PCRs. Moreover, the sensitivity of the PCR method allows the detection also of rare transcripts. Another advantage of the PCR method is that unspecific hybridisation, which represents a problem of the standard method, especially when highly homologous genes are studied, can be avoided by a careful PCR primer design. However, the semiquantitative RT-PCR method, like other end-point PCR-based quantification methods, requires a very careful correction for variations of reverse transcription and PCR efficiencies, which may be difficult and time-consuming to achieve. It appears likely, however, that new real-time quantification methods, like the TaqMan™ or LightCycler™ technology, may help to solve this problem and open up new perspectives for quantification.

3.2.4

Subtractive hybridisation

Subtractive (sometimes called competitive) hybridisation was originally developed as a method to identify the presence of families of repetitive sequences and interspersed repetitive and non-repetitive sequences between two genomes (Britten & Kohne, 1968). Using this strategy, the so-called 'tester' (or 'tracer') genome, which is the sample of interest, is hybridised with an excess amount of a control genome, the so-called 'driver', in which the target sequence is absent (Gray & Collins, 2000).

The principle of subtractive hybridisation is simple but efficient (figure 3.28). In a first step, the tester and driver populations prepared from the samples of interest, are mixed, denatured, and allowed to re-associate to form double-stranded molecules (hybrids). During hybridisation, three populations are formed: (i) tester/tester homohybrids, with both nucleic acid strands derived from the tester population; (ii) driver/tester heterohybrids, in which one strand is derived from each of the tester and driver populations; and (iii) driver/driver homohybrids, in which both strands are derived from the driver population. For a nucleic acid molecule present in equal amounts in both the driver and tester samples, the expected distribution of double-stranded hybrids in these three populations is, respectively, 25 percent, 50 percent and 25 percent. It is noteworthy that this post-hybridisation distribution is different for a nucleic acid species with greater abundance in the tester population than in the driver population. Therefore, at the end of the hybridisation, the distribution of double-stranded hybrids of such a nucleic acid will be biased towards tester homohybrids. Thus, the greater the amount of the nucleic acid of interest is in the tester compared to the driver population, the greater will be its enrichment during subtractive hybridisation.

The technique of genomic subtractive hybridisation was modified in a way such that this approach can be applied to the analysis of mRNA expression. First reported in 1987, subtractive hybridisation has been used for isolation of genes up-regulated in one cell type or tissue compared to another (Brown *et al.*, 1997; Candeliere *et al.*, 1999; Hillarby *et al.*, 1996). Modifications of the original subtractive hybridisation protocol are described in the following sections. Table 3.1 gives an overview of some other subtractive hybridisation-based approaches that are not described in detail here.

3.2.4.1 DISH – Differentiation induction subtraction hybridisation**3.2.4.1.1 Introduction**

Abnormalities in differentiation are common events in cancer (Jiang *et al.*, 1994; Leszczyniecka *et al.*, 2001; Waxman, 1996). This defect has been exploited as a therapeutic strategy, referred to as 'differentiation therapy' (Jiang *et al.*, 1994; Leszczyniecka *et al.*, 2001; Waxman, 1996). By using agents that modulate tumour cell differentiation, it has been possible to alter tumour growth and diminish cancer ag-

gressiveness (Jiang *et al.*, 1994; Leszczyniecka *et al.*, 2001; Waxman, 1996). It is hypothesised that differentiation therapy is effective because the growth and differentiation defects in cancer cells do not result from irreversible genetic changes, but rather that they reflect gene expression changes that are inducible following appropriate treatment(s) (Jiang *et al.*, 1994; Leszczyniecka *et al.*, 2001; Waxman, 1996). The process of differentiation therapy has been documented using metastatic human melanoma cells (Fisher *et al.*, 1985; Jiang *et al.*, 1993). When treated with a combination of recombinant fibroblast interferon (interferon- β , IFN- β) and the protein kinase C-activating agent mezerein, metastatic human melanoma cells lose their proliferative capacity irreversibly and terminally differentiate (Fisher *et al.*, 1985; Jiang *et al.*, 1993). This model system combined with subtraction hybridisation, referred to as differentiation induction subtraction hybridisation (DISH), is permitting the identification and cloning of genes involved in many important cellular processes, including growth, differentiation, programmed cell death (apoptosis), tumorigenicity and cellular senescence (Huang *et al.*, 1999; Jiang & Fisher, 1993; Jiang *et al.*, 2000). Using the DISH approach with high-density cDNA microarray and reverse Northern blot analyses it has been possible to identify an assortment of genes that are regulated during the process of terminal differentiation in human melanoma cells (Huang *et al.*, 1999). These included both known and novel genes (Huang *et al.*, 1999).

3.2.4.1.2 Principles and basics

By appropriate pharmacological manipulation it is possible to induce terminal differentiation in specific human cancers, such as melanoma, neuroblastoma and haematological malignancies (Jiang *et al.*, 1994; Leszczyniecka *et al.*, 2001; Waxman, 1996). This approach can be combined with molecular strategies designed to identify differentially expressed genes that associate with and contribute to the observed changes in cellular phenotype (Jiang & Fisher, 1993; Huang *et al.*, 1999). One molecular scheme that has proven particularly robust for this application is subtraction hybridisation (Jiang & Fisher, 1993). Paul B. Fisher and colleagues developed a method, called differentiation induction subtraction hybridisation (DISH), that employs sequential time point mRNAs to produce temporally-spaced subtracted (TSS) cDNA libraries enriched for genes displaying elevated expression as a function of terminal differentiation (Huang *et al.*, 1999; Jiang & Fisher, 1993). An overview of the DISH procedure is provided in figure 3.7.

Since differentiation involves coordinated changes in multiple genes that occur over defined temporal periods, understanding the kinetics of differentiation induction is paramount for using the DISH approach. Once temporal kinetics are clearly defined for induction of differentiation following treatment with a single or multiple agents, mRNAs are collected from control (untreated or solvent treated) and differentiation inducer-treated cells over the appropriate time frame for inducing differentiation. The temporal mRNAs are then converted into cDNAs, and tester (experimental) and driver (control) temporally-spaced cDNA libraries are prepared as described in detail previously (Jiang & Fisher, 1993). Subtraction hybridisation is then performed between the inducer-treated (tester) and non-treated control tem-

Table 3.1 Overview of subtractive hybridisation-based approaches for the analysis of gene expression not described in detail in this book.

Name	Abbreviation	References	Description
Chemical cross linking subtraction	CCLS	Byers <i>et al.</i> (2000) Hampson <i>et al.</i> (1992) Hampson & Hampson (1997)	cDNA/RNA hybrids are used in CCLS to simplify the subtractive hybridisation process. The method has been used to clone the p53 suppressor WAF-1 (El-Deiry <i>et al.</i> , 1993). A disadvantage of CCLS is the requirement of a great amount of starting material (5×10^8 cells or ten micrograms of poly(A) ⁺ mRNA). Otherwise, problems of handling and decay are reduced because the subtracted probe is not radiolabelled until use. In addition, several probes can be made from subtracted material and physical separation of cDNA/RNA hybrids is not necessary, improving efficiency and reducing loss of material.
Directed random oligonucleotide priming	DROP	Byers <i>et al.</i> (2000) Telenius <i>et al.</i> (1992)	In DROP, degenerate oligonucleotide-primed PCR (Telenius <i>et al.</i> , 1992) is used to synthesise cDNA so that directional overall PCR amplification of total cDNA can be achieved. Combining DROP with CCLS simplifies conventional hybridisation and overcomes the problems of cloning differentially expressed cDNAs from small cell numbers. The DROP products can be used in subtractive hybridisation in both directions and detect moderate-to-low abundance mRNAs.
Differential subtraction display	DSD	Pardinas <i>et al.</i> (1998)	DSD combines differential display RT-PCR and subtractive hybridisation of cDNAs. In DSD, standardised oligo(dT)-primed cDNA pools are used rather than total RNA as starting material for differential display RT-PCR, and a novel set of enhanced specificity anchor primers is applied. The subtractive hybridisation step prior to cloning of gel-excised and reamplified cDNA species enriches the <i>bona fide</i> cDNA species that are either exclusively present in one sample or that show variable expression between two RNA samples.

Table 3.1 (continuing)

Name	Abbreviation	References	Description
Phagemid subtractive hybridisation	–	Byers <i>et al.</i> (2000) Duguid <i>et al.</i> (1988)	This method was developed to overcome requirements for large amounts of starting material. Single-stranded phagemid vectors were used to amplify both driver and tester cDNA. In a modification of this method, Rubenstein <i>et al.</i> (1990) used directional insertion of cDNA into single-stranded phagemid vectors, together with quantification of the efficiency of subtraction by use of a bacterial transformation assay to determine the amount of tester remaining after subtraction.
Subtractive hybridisation combined with shot-gun sequencing	–	Byers <i>et al.</i> (2000) Perret <i>et al.</i> (1994)	In this method, subtracted products are amplified by PCR following initial subtractive hybridisation. After cloning, the fragments are analysed by random (shot-gun) sequencing. The method has been used in conjunction with competitive RNA hybridisation to identify genes associated with plant symbiotic functions (Perret <i>et al.</i> , 1994). However, shot-gun sequencing only identified genes with strong homologies to database entries.
Subtracted differential display	SDD	Byers <i>et al.</i> (2000) Bertioli <i>et al.</i> (1995)	Differential display RT-PCR and subtractive hybridisation have been used in combination in differential subtractive display (DSD) (Pardinas <i>et al.</i> , 1998), and in subtracted differential display (SDD). In SDD, subtractive hybridisation is followed by differential display RT-PCR, reducing the bias of more abundant mRNAs, thereby increasing the detection of less abundant mRNAs (Bertioli <i>et al.</i> , 1995). The RNA populations to be compared are used directly, without reverse transcription, in subtractive hybridisation and the enriched cDNAs subjected to differential display RT-PCR, with the advantage that fewer bands are visualised, thereby reducing the number of false-positives.

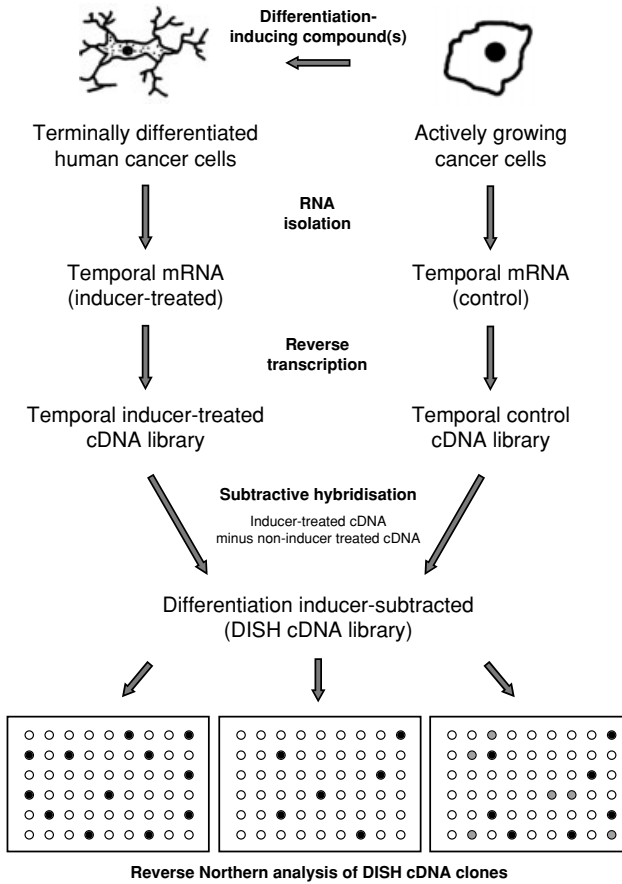


Figure 3.7 Schematic representation of the differentiation induction subtraction hybridisation (DISH) protocol for identifying genes differentially expressed in human cancer cells induced to terminally differentiate by treatment with single or multiple agents. Exposure of actively growing cancer cells to the appropriate pharmacological agent(s) results in growth inhibition, suppression of oncogenic potential and induction of reversible or terminal differentiation. Temporally-spaced mRNAs are isolated that encompass the time period required for induction of differentiation. These mRNAs are independently pooled (inducer-treated and control) and used to construct cDNA libraries, *i.e.*, temporal inducer-treated cDNA libraries (IND⁺ cDNA library; tester) and temporal control cDNA library (IND⁻ cDNA library; driver). Subtraction hybridisation is then performed, inducer-treated cDNA libraries minus control cDNA library, resulting in the construction of a

temporally-spaced subtracted (TSS) DISH cDNA library enriched for genes displaying elevated expression during the process of growth inhibition, suppression of oncogenic potential and induction of reversible or terminal differentiation. Random cDNA clones are isolated and PCR amplified inserts (smaller than 500 base pairs) are spotted on nylon membranes for reverse Northern screening. The custom-fabricated microarrays are screened using poly(A)⁺ mRNAs by the reverse Northern blot procedure to identify cDNAs up-regulated during different cellular processes, *i.e.*, terminal and reversible differentiation, growth arrest, apoptosis, *etc.* This approach results in the identification of pathway specific cDNAs and cDNAs with overlapping specificities (overlapping pathway screening (OPS) approach). The figure is modified from Huang *et al.* (1999a) with permission from *Gene*.

porally-spaced cDNA libraries resulting in a temporally-spaced subtracted cDNA library enriched for genes up-regulated in inducer treated cells, *i.e.*, a DISH cDNA library (Huang *et al.*, 1999a). By reversing the subtraction approach, *i.e.*, the non-treated control temporally-spaced cDNA library is the tester and the inducer-treated temporally-spaced cDNA library is the driver, it is possible to identify and clone genes displaying decreased expression as a consequence of induction of differentiation. The DISH library can then be screened using various approaches, including random colony isolation and probing Northern blots, colony isolation combined with formatting of cDNAs on nylon membranes and hybridisation with reverse transcribed and labelled first strand cDNAs (reverse Northern blot) (figure 3.8) and/or formatting large numbers of subtracted cDNAs (larger than 500 base pairs in size) on slides or other substrates (high-density DNA microarrays) and screening with reverse transcribed cDNAs (Huang *et al.*, 1999; Jiang & Fisher, 1993). This technique results in the identification of temporally regulated genes whose expression varies as a consequence of induction of terminal differentiation, including genes associated with growth control, expression of specific differentiation phenotypes, loss of tumourigenic potential, altered cell survival and other physiologically important processes. Differentiation induction subtraction hybridisation|

The types of originally novel genes identified using the DISH approach in human melanoma cells treated with interferon- β plus mezerein includes the cyclin-dependent kinase (CDK) inhibitor p21 (also called CDK-interacting protein 1, *Cip1*; wild-type p53-activated fragment 1, *Waf1*; *Sdi1* or melanoma differentia-

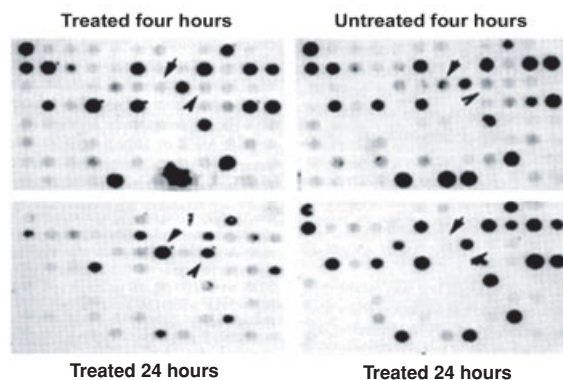


Figure 3.8 Analysis by reverse Northern blotting of random differentiation induction subtraction hybridisation (DISH) cDNA clones identified from HO-1 human melanoma cells induced to terminally differentiate and irreversibly lose growth potential by treatment with interferon- β plus mezerein. Random DISH cDNA clones were microarrayed on nylon membranes and probed with labelled cDNAs derived from four and 24 hours untreated con-

trol HO-1 cells and HO-1 cells treated with interferon- β plus mezerein (2000 units per millilitre and ten nanograms per millilitre). The arrows indicate two representative cDNAs displaying elevated expression as a function of treatment with the differentiation agents at 24 hours, but not at four hours. The figure is reproduced from Huang *et al.* (1999a) with permission from *Gene*.

tion associated gene 6, *mda-6*) (Jiang & Fisher, 1993; Jiang *et al.*, 1995a), a ubiquitous cancer growth suppressing and apoptosis inducing gene, interleukin-24 (IL-24; also termed *mda-7*) (Ekmekcioglu *et al.*, 2001; Huang *et al.*, 2001; Jiang *et al.*, 1995b and 1996; Lebedeva *et al.*, 2002; Madireddi *et al.*, 2000; Mhashilkar *et al.*, 2001; Pestka *et al.*, 2002; Saeki *et al.*, 2000; Su *et al.*, 1998 and 2001), a differentiation-associated gene encoding syntenin-1 (or *mda-9*), an adapter-like molecule that binds to the cytoplasmic domains of all four vertebrate syndecans and associates with the cytoplasmic tail of the interleukin-5 receptor (IL-5R α) and mediates activation of the transcription factor *Sox4* (Geijsen *et al.*, 2001; Grootjans *et al.*, 1997; Lin *et al.*, 1996 and 1998) and a novel interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties termed *mda-5* (Kang *et al.*, 2002). Moreover, microarray-based cDNA analysis of 1,000 DISH cDNA clones (about ten percent of the DISH library) resulted in the identification and cloning of 26 known and eleven additional novel cDNAs of potential relevance to growth control and terminal differentiation in human melanoma cells (Huang *et al.*, 1999b). The known genes included interferon-inducible genes, genes associated with the immune system, genes regulated during differentiation and growth (including cyclin-dependent kinase inhibitor p21 (*mda-6*) and *mda-7*), genes associated with transcription regulation and signal transduction, genes related to the cytoskeleton, genes associated with mitochondria and genes without clear relevance to terminal differentiation (Huang *et al.*, 1999b).

A limitation of this high-throughput screening approach is the requirement for cDNAs larger than 500 base pairs. In contrast, DISH combined with reverse Northern blot analysis, which can use cDNAs smaller than 500 base pairs, has yielded 30 known and 26 novel cDNAs after random screening of 400 DISH cDNA clones (Huang *et al.*, 1999a). A drawback to both the high-density DNA microarray and reverse Northern blot approaches is redundant clone isolation, since differential expression is determined before sequencing (Huang *et al.*, 1999). By pre-screening cDNAs prior to or after formatting arrays with cDNAs anticipated to be overexpressed or previously identified in the original DISH library it will be possible to significantly reduce redundant clonal isolation. In addition, by creating DISH libraries for specific time points in the differentiation process it will be possible to define specific cDNAs of interest that display very restricted expression profiles and which may not be detected because of dilution of the signal resulting from combining multiple time spaced mRNAs to produce temporally-spaced cDNA libraries.

Temporally-spaced cDNA libraries represent valuable reagents to define gene changes occurring as a consequence of complex physiological changes in cells (Huang *et al.*, 1999; Jiang & Fisher, 1993; Leszczyniecka *et al.*, 2001). This strategy will prove useful for analysing cellular changes that occur over specific time frames in cells, as well as changes which are not cell-cycle related. The DISH approach offers particular promise for defining the complete repertoire of gene changes associated with and causing cellular differentiation.

3.2.4.1.3 Discussion

Terminal cell differentiation involves sequential changes in the expression of diverse sets of genes (Huang *et al.*, 1999; Jiang & Fisher, 1993; Leszczyniecka *et al.*, 2001; Waxman, 1996). The DISH approach represents an experimental strategy for defining the repertoire of genes underlying growth control and differentiation. The DISH method incorporates two important components that increase its utility for defining gene changes involved in growth control, induction of terminal differentiation and regulation of cancerous properties. First, it uses temporally-spaced mRNAs for the construction of both tester and driver cDNA libraries. This aspect of DISH limits the cloning of genes that display normal fluctuations as a consequence of changes in cell-cycle. Second, by using reverse Northern hybridisation of arrayed temporally-spaced subtracted cDNAs multiple or overlapping screens are possible to identify genes modulated during specific treatment protocols. In these contexts, the DISH approach provides a unique way of identifying and cloning genes regulated by and controlling important physiological changes in cells. Moreover, by using small molecules that can induce terminal differentiation in cancer cells, the DISH approach will help in elucidating important genes and pathways involved in cancer growth control. These genes can then serve as targets for developing effective therapies for cancer and other proliferative diseases.

3.2.4.1.4 Outlook

The DISH library has also provided the basis for a new method, termed the overlapping pathway screening (OPS) approach, which permits the identification and ultimate cloning of genes that display changes as a function of induction of common cellular changes induced during complex biological processes. By probing the DISH library with cDNAs derived from cells undergoing defined changes in cellular physiology it is possible to examine which program of gene expression changes are shared by these processes. Potential applications of OPS include defining genes displaying overlapping expression as a function of terminal differentiation and cellular senescence, growth suppression and DNA damage induction and conversion of totipotent stem cells into specific differentiated cell types. The OPS approach will open up new vistas for defining complex circuitry in cells that could provide new insights into individual and groups of genes associated with and regulating apparently distinct but nevertheless converging pathway responses in cells.

3.2.4.2 DSC – Differential subtraction chain

3.2.4.2.1 Introduction

Methods for enriching DNA fragments of interest are some of the most frequently used procedures in modern biomedical research. This procedure of PCR, for example, allows exponential enrichment of a DNA and to rapidly isolate a specific DNA sequence from a myriad of DNA sequences. However, use of PCR requires prior knowledge of the DNA sequences that flank the region of interest. Several methods had therefore been developed to identify and enrich sequences present in one sam-

ple but not in another where the sequences in question are not known in advance (Duguid & Dinauer, 1990; Hara *et al.*, 1991; Kunkel *et al.*, 1985; Nussbaum *et al.*, 1987; Sargent & Dawid, 1983; Straus & Ausubel, 1990). Earlier techniques used excessively dephosphorylated driver DNA to hybridise with testers in order to prevent ligation of testers to a phage vector (Kunkel *et al.*, 1985; Nussbaum *et al.*, 1987). Later subtraction techniques involved using excessive drivers that were tagged with chemical reagents to enable them to hybridise with testers, with subsequent elimination of the tagged-driver/tester hybrid by means of affinity chromatography (Duguid & Dinauer, 1990; Hara *et al.*, 1991; Sargent & Dawid, 1983; Straus & Ausubel, 1990). Messenger RNA differential display (Liang & Pardee, 1992) and RNA arbitrarily primed PCR fingerprinting (Welsh *et al.*, 1992) represent potentially faster and easier techniques to identify differential expression of genes. In 1983, a PCR-based technique, called representational difference analysis (RDA) was also developed (Lisitsyn *et al.*, 1993). This technique employed a representational sampling approach by cutting the DNA into fragments based on its restriction enzyme cutting pattern, and then attaching these restriction fragments to a PCR-adaptor for PCR amplification. A differential enrichment approach is then used to identify and enrich the differences between the tested DNA samples without physical separation (Hubank & Schatz, 1994; Lisitsyn *et al.*, 1993). Linker capture subtraction (LCS) was later developed to simplify the process of RDA by eliminating the procedure of adapter digestion and re-ligation (Yang & Sytkowski, 1996). However, the enrichment process remains complex because it requires amplification, repeated dilution of testers and addition of a large amount of drivers in each cycle of the subtraction. To overcome these problems, and to improve on the simplicity and subtraction efficiency of the procedure, differential subtraction chain (DSC) was developed (Luo *et al.*, 1999).

3.2.4.2.2 Principles and basics

Before differential subtraction chain can be applied to the analysis of genomic DNA, the complexity of the genome sequences must be reduced in order to improve hybridisation efficiency. In order to reduce the complexity of genomic sequences in hybridisation, the concept of genomic representational sampling proposed by Lisitsyn *et al.* (1993) was adopted, in which short fragments of endonuclease-digested genomic DNAs are ligated with adapters and amplified to produce 'amplicons'. In theory, genomic representation by amplicons should reduce the complexity of sequences by 80 to 100-fold and dramatically reduce the time needed to perform hybridisation. However, once the amplicons were generated, the differential enrichment approach was abandoned. Instead, a negative amplification strategy to eliminate homologous sequences present in two different DNA pools was developed. The principle of DSC is based on the assumption that maximum efficiency of a subtractive process depends on the ratio of the DNA that is used to subtract (the driver or non-adapter-tagged amplicon) to the DNA being subtracted (the tester or adapter-tagged amplicon). In order to achieve high driver to tester ratios, a strategy is applied that converts testers to homologous drivers after each subtractive round. This 'negative enrichment' process is accomplished by removing the prim-

ers from both ends of the testers by a nuclease specific for single-stranded DNA after they cross-anneal to drivers (figure 3.9). In the subsequent cycles of the same procedure, the testers without primers will become drivers, and serve to convert more testers to drivers from the tester population. The result is a double exponential decrease in the copy number(s) of tester species that have homologous counterparts in the driver population and a corresponding increase in the copy number(s) of the driver species (see table 3.2 for mathematics evaluation). This process is termed 'negative amplification'. If 10^9 copies of a target sequence present in one population must be eliminated ($Y_e < 1$, see table 3.2 for definition) in order to achieve subtraction in the presence of a 99-fold ($\gamma = 10^{11}$, $A = 1/100$) excess of driver, then three ($n = 2.46$) rounds of negative amplification will be required. This negative amplification process achieves two goals: (i) By complete removal of the primers from the unwanted sequences, it eliminates any possible amplification while preserving the rare DNA sequences. This will improve the subtraction sensitivity and efficiency, and produce cleaner products (fewer false-positive results). (ii) By bypassing the need for positive enrichment processes between hybridisations, it dramatically simplifies and speeds up the subtractive process.

3.2.4.2.3 Mathematical consideration

In the experiment of DSC, many factors affect the enrichment efficiency. These include, but are not limited to, the completeness of hybridisation, the digestion specificity and efficiency of a nuclease specific for single-stranded DNA, and the equality of tester hybridisation affinity with tester or driver. In order to reduce these variables in mathematical consideration, it was assumed in theory that in a DSC reaction, a complete hybridisation of tester and driver occurs, annealing of tester to tester and tester to driver has the same efficiency, and the nuclease specific for single-stranded DNA digests all single-stranded DNA but not double-stranded DNA. As demonstrated in table 3.2, the mathematical model of DSC, if the starting tester/driver ratio is 1, the survival number(s) of tester after each round of DSC follows the probability of tester to self-anneal in the preceding round, which is $\frac{1}{2}$ before round one of DSC, $\frac{1}{4}$ after round one, $\frac{1}{16}$ after round two, $\frac{1}{256}$ after round three, $\frac{1}{65536}$ after four rounds, ..., and $1/(2^2)^n$ after n rounds. Therefore, the survival copy(s) of tester can be predicted by the equation

Table 3.2 Mathematics evaluation of differential subtraction chain (DSC) procedure ($Y = Y_e + Y_{or}$).

DSC round(s) <i>n</i>	Tester copies Y_e	Driver copies Y_{or}	Testers to total ratios <i>A</i>
0	1,000,000	1,000,000	$\frac{1}{2}$
1	500,000	1,500,000	$\frac{1}{4}$
2	125,000	1,875,000	$\frac{1}{16}$
3	7,812.5	1,992,187.5	$\frac{1}{256}$
4	30.5	1,999,969.5	$\frac{1}{65,536}$
5	~ 0	$\sim 2,000,000$	$\frac{1}{4,294,967,296}$

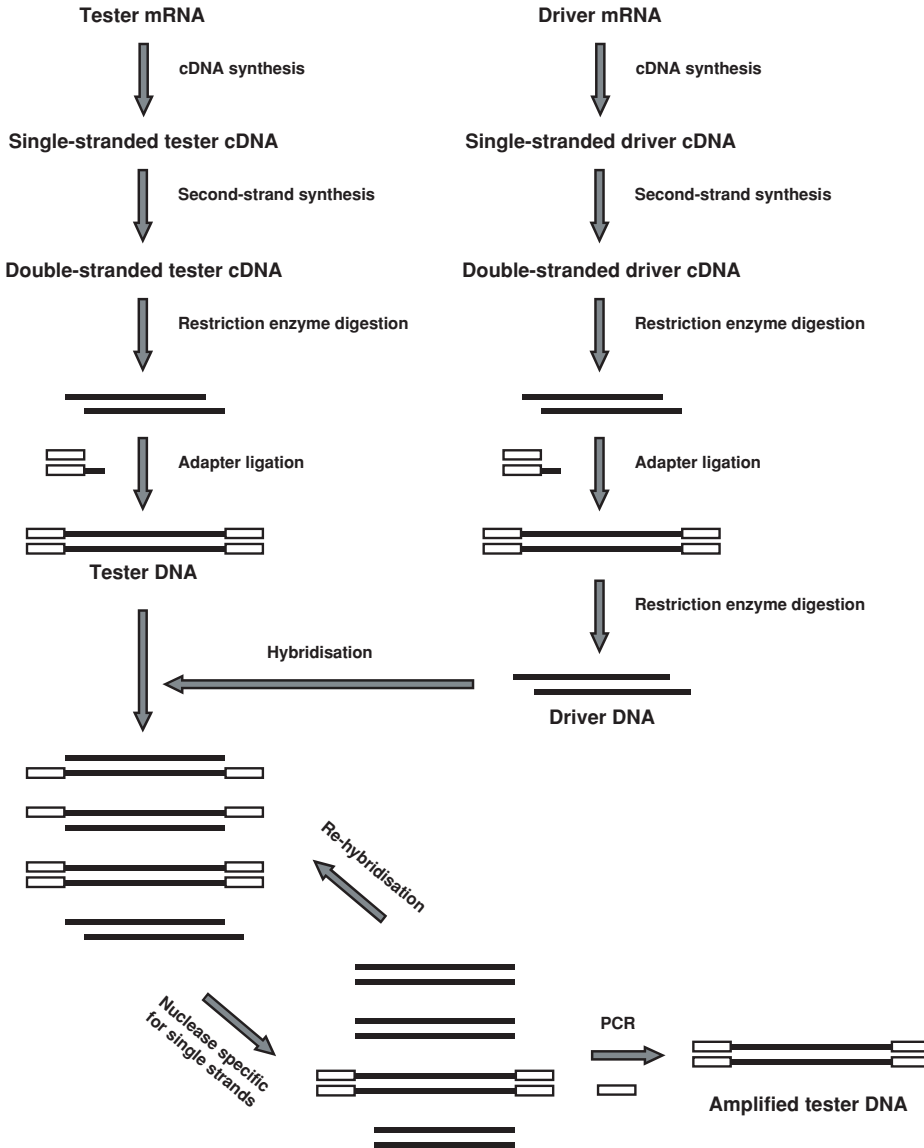


Figure 3.9 Scheme of differential subtraction chain (DSC) principle. Pool A cDNA is used as testers and pool B cDNA as drivers. Pool A and B represent restriction enzyme-digested cDNA fragments. These fragments are ligated with separate sets of adapter primers. Subsequently, the ligation products are amplified with primers specific to adapter primers to generate amplicons (adapter-tagged restriction DNA fragments). For pool B cDNA, amplicons are digested with restriction enzyme to re-

move the adapters to become drivers before hybridisation. In the DSC reaction, mixtures of testers and drivers are denatured and hybridised. Subsequently, the hybridisation products are treated with mung bean nuclease to remove the single-stranded primers. After mung bean nuclease is removed, DNA is subjected to denaturing and rehybridisation. An aliquot (five to ten percent) of DSC reaction stock is taken for PCR after each round of DSC.

$$Y_e = Y \cdot (A^2)^n,$$

where Y_e is the probable copy number of survived tester after n round(s) of DSC, Y is the total copy number of a tester species and its relevant driver, and A is the probability for tester(s) to self-anneal before the first round of DSC. Similarly, the probability of the conversion of tester to driver in each round of DSC is proportional to the probability that a tester will anneal with a driver. This is $1/2$ before round one of DSC, $3/4$ after round one of DSC, $15/16$ after round two, and effectively $1 \cdot (1 - 1/2 \cdot 2^n)$ after all subsequent rounds. The copies of driver that survive each round of DSC can be predicted by equation

$$Y_{or} = Y \cdot (1 - Y \cdot (A^2)^n),$$

where Y_{or} denotes total copy number(s) of driver after n round(s) of DSC. The probability for complete elimination of a tester species can be assessed as $1 - Y \cdot (A^2)^n$.

In DSC, with the double exponential kinetics of tester to driver conversion, the conversion of high abundance tester to driver is faster than the conversion of low abundance testers. This normalisation (equalisation) of the rate of tester to driver conversion in DSC can be predicted mathematically, since the conversion rate ($Y_{convert}$) of a tester to driver is expressed in a linear regression relationship with the total abundance of that specific tester Y_e in a given round n :

$$Y_{convert} = Y_e \cdot (A - (A^2)^n) + Y_d \cdot (A - (A^2)^n),$$

where $Y_{convert}$ is the probable copy number of tester conversion to driver after n round(s) of DSC, Y_e is the total copy number(s) of a tester species, Y_d is total copy number(s) of relevant driver and A is the probability for tester(s) to self-anneal before the first round of DSC. In other words, testers with copy numbers ranging from 10^7 to 10^{14} will be completely converted to driver in the same round of DSC if 100-fold copies of driver are applied, *i. e.* testers with a copy number within this range will survive the second round of DSC but not the third round. This feature may make DSC quite useful in mRNA analysis where the abundance of individual species varies widely.

3.2.4.2.4 Examples

A 560 base pair fragment of *Hind* III-digested λ DNA was ligated with an adapter primer, and was amplified to generate a tester (λ amplicon, or λa). More than seven billion (7.7×10^9 or five nanograms) copies of this amplified fragment were used in a DSC subtraction in the presence of ten or 100-fold of corresponding DNA fragment of λ DNA (driver). As shown in figure 3.10, complete conversion of tester to driver was achieved after four rounds of DSC when a ten-fold excess of driver was applied, while three rounds of DSC were required to reach the similar result when a 100-fold excess of driver was used. These results fit well with the model shown in table 3.2. When ten times more driver than tester is present, about 395 copies of

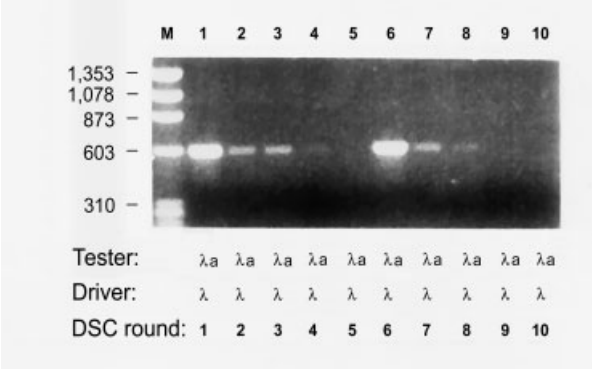


Figure 3.10 Subtraction efficiency of differential subtraction chain (DSC) on λ amplicon. Five nanograms of 610 base pair λ amplicon (λa) were subtracted by 3.9 micrograms (lanes 1 to 5) or 39 micrograms (lanes 6 to 10) of total λ DNA (*Hind* III restricted). The DSC products were amplified by PCR after zero (lanes 1 and 6), one (lanes 2 and 7), two (lanes 3 and 8), three (lanes 4 and 9), and four (lanes 5 and 10) rounds of DSC.

tester survive after three rounds of DSC. The minimum number of tester sequences that can survive four rounds of DSC is 4.18×10^{15} (when $Y_e = Y \cdot (A^2)^n = 1$, and $Y_0 = Y \cdot A$, where Y_0 denotes tester copy number before first round of DSC, then is $Y_0 = 1 / Y \cdot (A^2)^n - 1$). When a 100-fold excess of driver is used, the minimum copies of tester that are needed to survive three rounds of DSC is 1.07×10^{14} . Clearly, complete subtraction of λ amplicon should be achieved within four rounds of DSC if a ten-fold excess of driver is used, and within three rounds if a 100-fold excess is applied.

In a recent approach, differential subtraction chain was used to identify genes whose expression is activated or inactivated in high grade prostate cancer (Yu *et al.*, 2001). In this study, a novel gene, designated as SAPC (= suppressed in aggressive prostate cancer), was found that is expressed in normal prostate acinar cells. Its expression is dramatically down-regulated in high grade prostate cancers but is unaltered in low grade prostate cancers. It encodes a 7.7 kilodalton protein. Its sequence shares some homology with the cysteine-rich domain of 2-5A-dependent RNase L, which is a critical component of the interferon-induced apoptosis cascade.

3.2.4.2.5 Discussion

Differential subtraction chain is an method for identifying unique DNA/RNA sequences and an alternative to representational difference analysis (RDA) and linker capture subtraction (LCS). Compared to these methods, DSC requires fewer steps, and is therefore less labour-intensive and potentially faster. Although there is some similarity between DSC and RDA or LCS, the enrichment kinetics and the operation mechanism to enrich the unique sequences by DSC are quite different from the other two methods. In theory, DSC offers greater enrichment speed, and toler-

ates lower starting driver/tester ratios because of its double exponential feature. However, the actual enrichment efficiency has yet to be directly compared among the three methods.

Similar to RNA fingerprinting and differential display RT-PCR, DSC is designed to identify mRNA amplicons that represent the differences of two comparing RNA samples. Differential display RT-PCR identifies the differences between the two samples based on visual differences of bands of comparable sizes in a sequencing gel, while DSC sets out to further simplify the pool of amplicons based on the sequence homology between testers and drivers. The time and effort required for both of these methods would appear to be about the same.

Despite high efficiency of enrichment of unique tester sequences in the tester/driver mixes by DSC, the DSC reaction is not suitable for quantitative analysis. This is because DSC subtracts most of the testers in two to three rounds. In addition, in order to estimate the original quantity of a tester template that has been amplified twice in PCR is quite complicated. Thus, when quantitative analysis for RNA level is needed, methods such as serial analysis of gene expression (SAGE) or DNA microarray analysis may be preferable.

DSC works at both the genomic and the RNA level. It is particularly useful in organisms where knowledge of the genome sequences is patchy. One useful application is the discovery of microbial sequences in samples. However, several hurdles have to be overcome before it can be widely used as a novel sequence discovering tool. First, the sensitivity of DSC needs improvement. When infectious agents are scarce in the samples, DSC may have difficulty detecting them. Several factors contribute to this problem, including efficiency of restriction enzyme cutting and religation of infectious agents with adapter sequences. In addition, testers of interest may be underrepresented among the PCR amplicons when dominant species are present. Hybridisation of single-stranded testers in each cycle of the DSC reaction may be incomplete, tester sequences may also be damaged due to repeat hybridisation in high temperature and exposure to mung bean nuclease, and loss of tester sequences in the purification processes. Techniques that speed up hybridisation and procedures that prevent DNA damage or repair damaged DNA may be required in order to preserve the minute amount of testers present in repeat cycles of DSC. Although limited cycles of PCR can be applied to amplify testers after each cycle of DSC, results indicate that the tester-preserving effect of such procedures appeared minimal, probably reflecting substantial loss or unrecoverable damage of tester sequences during hybridisation and exposure to mung bean nuclease. Linear amplification of amplicons using *in vitro* transcription by a T7 promoter, instead of exponential amplification by PCR, has been shown to closely reflect the pre-amplified quantity of templates, and may offer a means to ensure proper representation of amplicons of interest in the samples. The specificity of DSC also needs to be improved. At the present time, the specificity of DSC relies on a complete digestion of single-stranded primer attached to tester sequences. The condition of mung bean nuclease digestion may need to be optimised to achieve this goal while avoiding degradation of double-stranded testers or drivers.

3.2.4.3 EDS – Enzymatic degrading subtraction

3.2.4.3.1 Introduction

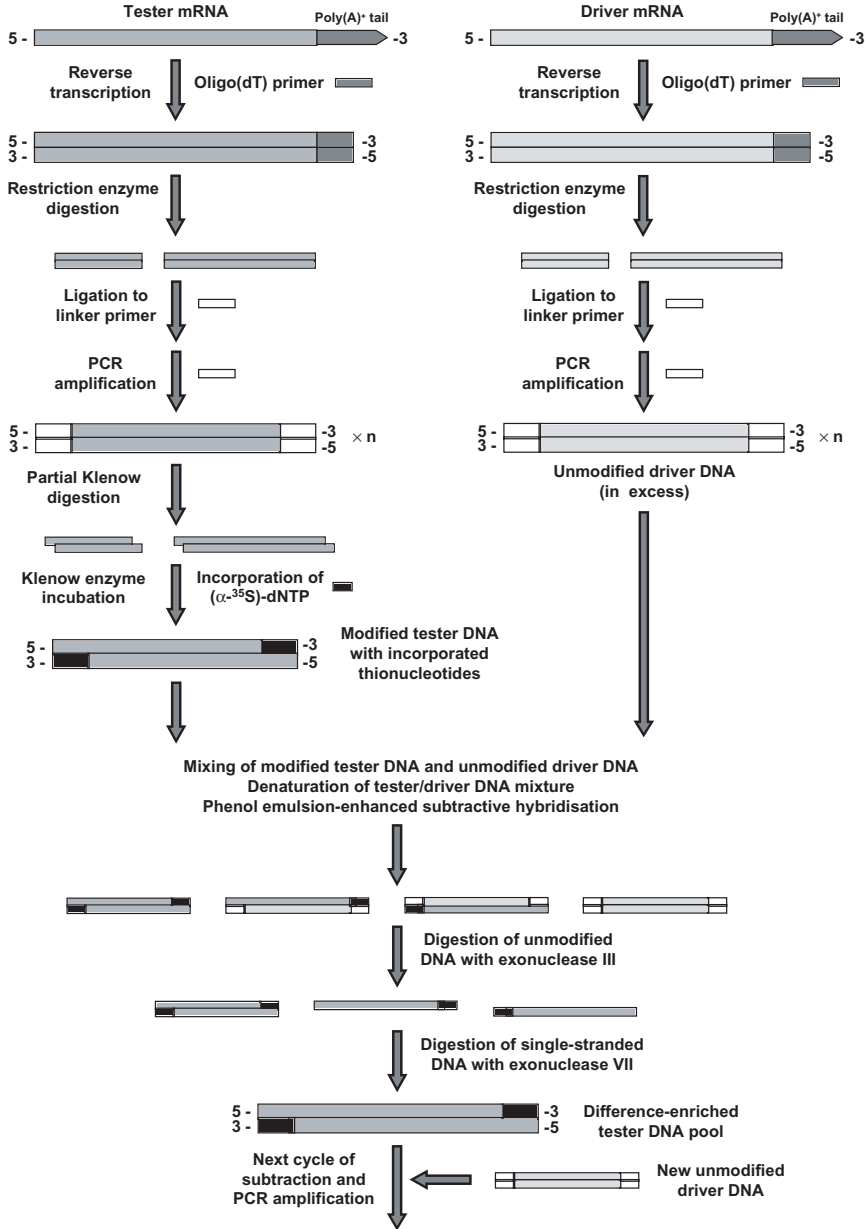
Subtractive cDNA library construction is a popular method for cloning target sequences that are present in one DNA population, the 'tester', but almost completely absent in another, the 'driver'. The power of subtractive cDNA library construction has been substantially increased by the inclusion of several modifications such as PCR amplification. However, difficulties in the methodology for constructing subtractive cDNA libraries remain. The modifications of the subtractive hybridisation technique that are described in this section of the book have been developed to circumvent these problems.

Enzymatic degrading subtraction (EDS) is a subtractive hybridisation method which was developed by Zeng *et al.* (1994) for identifying differentially expressed genes in adult but not in embryonic rat brains. EDS was designed (i) to increase the inefficiency with which restriction enzyme digestion of the lone and single-stranded linker primer of the double-stranded driver/driver DNA homohybrids is blocked, (ii) to circumvent problems with the biotinylation reaction that is performed in subtractive hybridisation methods, and (iii) to overcome the cumbersome separation of modified and unmodified DNA molecules in most subtractive hybridisation protocols (Zeng *et al.*, 1994).

To circumvent these drawbacks, the method of EDS incorporates several novel features: First, the tester DNA is protected from restriction enzymatic digestion by incorporation of thionucleotides. This avoids the need to label the driver cDNA by biotinylation, thus allowing the biotinylated molecules to be removed prior to enzymatic digestion, as is also done in the selective amplification *via* biotin and restriction-mediated enrichment method (SABRE). Second, the driver cDNA and tester/driver hybrid molecules are removed by digestion with exonuclease III, which digests double-stranded and unmodified DNA but not double-stranded and thionucleotide-modified DNA, and by digestion with exonuclease VII, which is specific for single-stranded DNA. Thus, only double-stranded tester DNA fragments that are subsequently ligated to linker primers remain in solution. PCR amplification as described by Wang & Brown (1991) is used to enrich tester-specific cDNAs. Third, it is possible to accelerate the rate of hybridisation by using the phe-

Figure 3.11 Schematic presentation of the enzymatic degrading subtraction (EDS) procedure. First, double-stranded cDNA is synthesised by reverse transcription from two isolated RNA pools, which is then amplified by PCR and digested into short fragments using restriction enzymes. The tester DNA is modified by incorporation of (α -³⁵S)-dNTPs using Klenow enzyme after partial digestion with this enzyme. Because of the excess of driver cDNA hybridisation of tester DNA and driver DNA will yield mostly tester/driver DNA heterohy-

brids for genes that are expressed equally in both cell populations and also driver/driver DNA homohybrids. By contrast, differentially expressed genes yield tester/tester or driver/driver DNA homohybrids. Unmodified driver DNA is digested by exonuclease III and single-stranded tester DNA is digested by exonuclease VII. The remaining double-stranded tester DNA is subjected to further cycles of subtraction until the enrichment of different genes is sufficient for further analysis. ►



nol emulsion reassociation technique (PERT) so that the amount of cDNA required is smaller than with other methods for subtractive library construction.

3.2.4.3.2 Principles and basics

Total RNA is isolated from the so-called 'tester' and 'driver' cell populations and poly(A)⁺ mRNAs are selected for reverse transcription to double-stranded cDNA (figure 3.11). Separate aliquots of the tester and driver cDNA populations are digested to completion into short fragments of less than one kilobase with the restriction enzymes *Alu* I and *Alu* I plus *Rsa* I, respectively (Zeng *et al.*, 1994; Wang & Brown, 1991). The cDNA fragments are then ligated at their blunt end to linker primers with one flush end and one four base 3'-protruding end (Duguid & Dinauer, 1990) to allow PCR amplification.

The tester DNA is then treated with Klenow enzyme, which has 3'-5' exonuclease activity in the absence of any nucleotide, in order to partially digest the cDNA. Single-strand ends are filled with (α -³⁵S)-dNTPs using the 5'-3' polymerase activity of the Klenow enzyme. As a result of these modifications, tester DNA is protected against digestion by exonuclease III. The modified tester DNA is hybridised with an excess of unmodified driver DNA. The speed and sensitivity of the selection procedure is improved by the use of protocols that accelerate DNA hybridisation, such as the phenol emulsion reassociation technique (PERT) developed by Kohne *et al.* (1977) which can increase the speed of hybridisation by up to 100-fold (Miller & Riblet, 1995; Kohne *et al.*, 1997).

The emulsion can be maintained by continuous vortexing but remains a problem for lower concentrations of DNA (Kohne *et al.*, 1977). In such cases, thermal cycling may be used (Miller & Riblet, 1995). Because of the excess of driver DNA in the hybridisation reaction almost all tester DNA fragments with genes which match the driver DNA will yield tester/driver DNA heterohybrids, with only a small number of tester/tester DNA homohybrids. The tester DNA containing cDNA almost unique to the tester sample will yield mainly tester/tester DNA homohybrids, with only a small number of tester/driver DNA heterohybrids. Driver DNA containing almost uniquely expressed sequences will form mainly driver/driver DNA homohybrids, with only a small number of tester/driver DNA heterohybrids. After hybridisation, all unmodified driver DNA strands are degraded by exonuclease III which is inactive towards thionucleotide-containing DNA, leaving only tester/tester DNA homohybrids and single-stranded tester DNA in solution. The remaining single-stranded tester DNA is degraded by exonuclease VII which is specific for single DNA strands, whereas double-stranded and thionucleotide-modified tester/tester DNA homohybrid molecule remain unaffected. Repeated cycles of subtraction enriches sequences present only in the tester DNA library by accumulation of tester/tester DNA homohybrids.

3.2.4.3.3 Examples and presentation of typical results

EDS was used to construct a subtractive library enriched for cDNAs expressed in adult but not in embryonic rat brains (Zeng *et al.*, 1994). In this study, the original non-subtracted cDNAs from adult rat brain moved as a smear between 0.15 to 0.4

Figure 3.12 Electrophoretic analysis of tester cDNAs after multiple rounds of subtraction and amplification. *Lane 1*: Starting tester cDNA preparation. *Lane 2 to 4*: cDNA fragments remaining following the first (*lane 2*), second (*lane 3*) and third (*lane 4*) round of subtraction and amplification. *Lane M*: 100 base pair ladder as a molecular weight marker. Each lane contains 0.8 micrograms of PCR-amplified cDNA fragments.

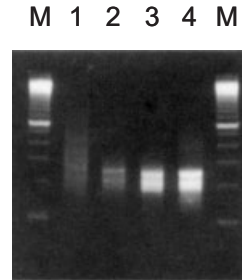


Table 3.3 Summary of five cDNA fragments expressed selectively in adult brain which were identified using the enzymatic degrading subtraction method (adapted from Zeng *et al.*, 1994).

cDNA fragment no.	Size of cDNA insert (kilobases)	Size of mRNA (kilobases)	Number of identified clones ¹	Difference in expression
a1	0.26	2.5	1	> 20-fold
a4	0.32	0.9	1	> 20-fold
a6	0.22	1.2	2	2.5-fold
a8	0.29	0.5	5	> 20-fold
a13	0.30	0.7	2	> 20-fold

¹ Number of identified clones is the number of clones representing a unique gene in 16 randomly selected clones.

kilobases in an agarose gel. After the first cycle of subtraction, several distinct bands of between 0.2 to 0.4 kilobases were detected. The intensity and sharpness of these bands increased progressively with successive subtractions as expected for the continuing removal of common sequences (figure 3.12). Preparation of a plasmid library revealed eleven cDNA fragments corresponding to mRNAs specifically or more abundantly expressed in adult brain (table 3.3). These selectively expressed clones represented five distinct mRNA sequences. In addition, the authors demonstrated the ability of enzymatically degraded subtraction to select cDNA species with different abundance by detecting mRNAs that differed in abundance by a factor of ten. Differences in gene expression were verified by Northern blot analysis using two representative clones (a1 and a4) as probes (figure 3.13). Up to now, this study is the only recorded use of the EDS method.

3.2.4.3.4 Discussion

Despite its rare use, EDS is an economical method for the construction of subtractive cDNA libraries and, compared to other methods of subtractive hybridisation, EDS avoids problems associated with biotinylation and with the cumbersome separation of modified and unmodified DNA molecules. Because the incorporation of thionucleotides into the tester DNA strand reduces the amount of tester cDNA required, and because repeated labelling reactions are not required, this method is more economical than other procedures using biotinylation of the driver

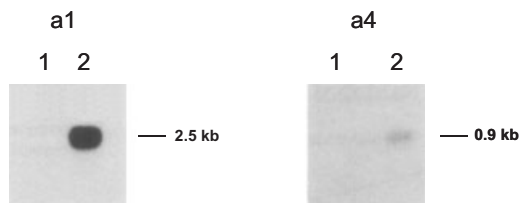


Figure 3.13 Northern blot analysis of two representative clones in embryonic and adult rat brains. Five micrograms of total RNA were size-fractionated by denaturing formaldehyde agarose gel electrophoresis and hybridised to a

random primer-synthesised and radioactively labelled probe obtained by PCR amplification of the clone inserts a1 and a4 which were identified using the enzymatic degrading subtraction method (for information see table 3.3).

cDNA. Another advantage of EDS is the possibility of using the phenol emulsion reassociation technique which allows hybridisation with only a few micrograms of cDNA. Methods employing biotinylated driver cDNAs normally cannot use the phenol emulsion reassociation technique because the biotinylated DNA is partitioned into the organic phenol phase where it forms an insoluble precipitate (Zeng *et al.*, 1994). More recently, this limitation was circumvented by the development of a thermal cycler-based adaptation of the phenol emulsion reassociation technique (Lavery *et al.*, 1997; Miller & Riblet, 1995).

In addition to the economical advantages and the small amounts of required DNA, the most important benefit of the EDS method is the enzymatic removal of tester/driver DNA heterohybrid and driver/driver DNA homohybrid molecules.

EDS also has some disadvantages. Since it is based on PCR-based cloning, EDS is subject to errors due to non-random amplification. In addition, comparison of more than two cDNA populations by EDS is difficult and time-consuming as it is with all subtractive hybridisation based methods for analysing gene expression. Finally, in order to identify down-regulated mRNAs, the whole procedure must be repeated in reverse.

3.2.4.4 GES – Gene expression screen

3.2.4.4.1 Introduction

The gene expression screen (GES) method was developed in 1991 to identify mRNAs that differ in abundance between two mRNA mixtures (Wang & Brown, 1991). This PCR-based subtractive hybridisation method differs from other subtraction methods in that it provides an estimation of the number genes that are up and down-regulated. Since its development in 1991, the gene expression screen has been used to identify thyroid hormone-response genes in *Xenopus laevis* (Wang & Brown, 1991) and androgen-response genes in the normal prostate or prostate cancer cells (Geck *et al.*, 1999; Gubbay *et al.*, 1998; Wang *et al.*, 1997), as well as genes involved in the development of *Drosophila* (Casal & Leptin, 1996). It is the basis of one of the subtraction methods presented in 'Current Protocols in Molecular Biology' (Patel & Sive, 1996).

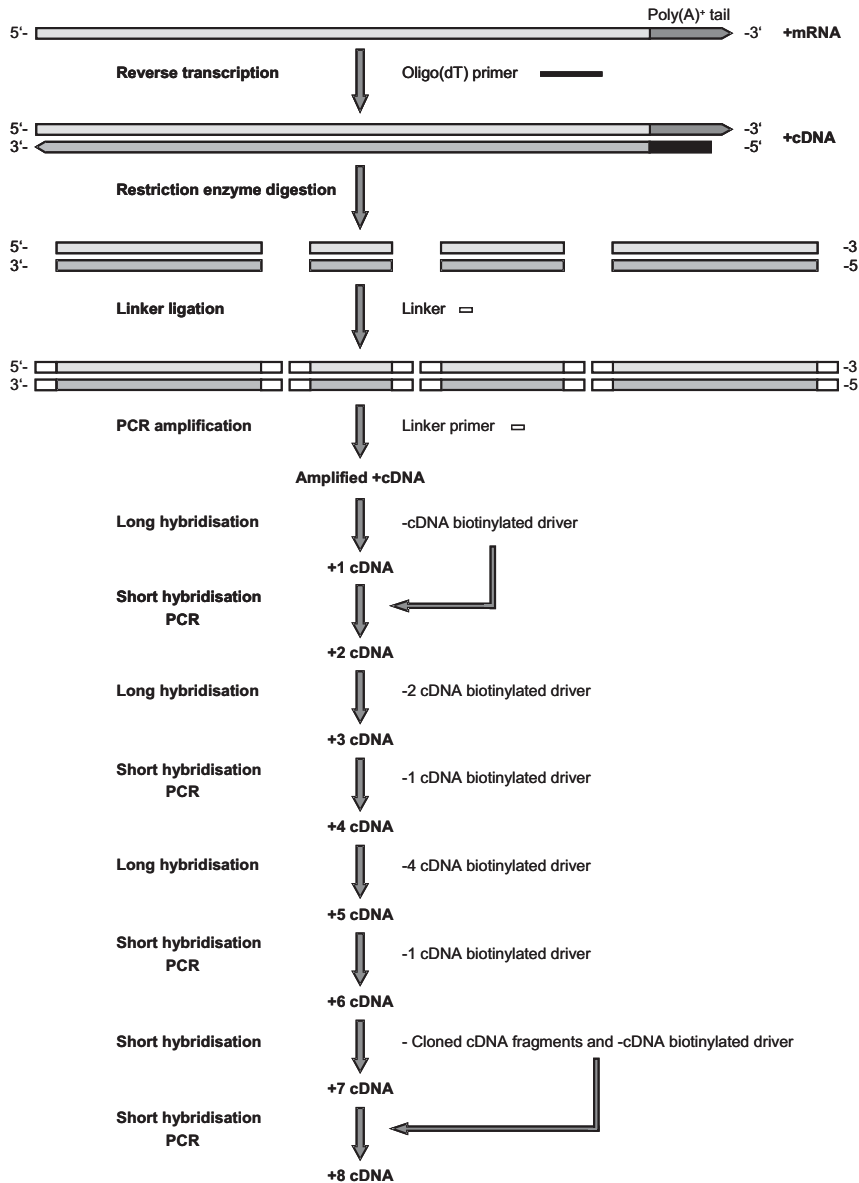


Figure 3.14 Flow diagram for the isolation of **up-regulated genes**. A plus sign (+) refers to the mRNA isolated from tadpole tails treated with thyroid hormone for 24 hours, as well as cDNAs derived from this plus mRNA; a minus sign (-) refers to mRNA and cDNAs from untreated tadpoles. The number n (1,2,3,...,8) fol-

lowing the plus or minus sign corresponds to the n^{th} cycle of hybridisation of +cDNA and -cDNA. The exact opposite protocol is carried out simultaneously to obtain the down-regulated genes. Adopted from Wang & Brown (1991) with permission from *Proceedings of the National Academy of Sciences of the USA*.

3.2.4.4.2 Principles and basics

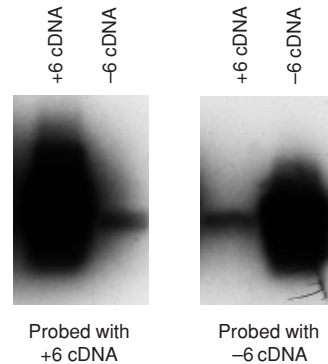
The principle of the gene expression screen is illustrated in figure 3.14. It involves multiple rounds of long subtractive hybridisation, short subtractive hybridisation, and PCR amplification (Wang & Brown, 1991; Wang *et al.* 1997). The long hybridisation removes low abundance common cDNAs, while the short hybridisation depletes high abundance common cDNAs effectively. PCR then amplifies the enriched differentially expressed sequences. Equal suppression of commonly expressed cDNAs in the two parallel reactions is critical, which requires accurate measurement of cDNA concentration and efficient DNA purification. To minimise unequal suppression of abundant but constantly expressed cDNAs, a small amount of the original (unenriched) biotinylated driver cDNA may be included in the driver cDNA during each subtractive hybridisation step (Wang *et al.*, 1997).

The double-stranded cDNA is cleaved with four-base cutter restriction enzymes prior to linker ligation and PCR amplification (figure 3.14), yielding cDNA fragments that are more suitable for PCR amplification than the undigested cDNA (Wang & Brown, 1991). The restriction enzyme digestion of cDNA prior to PCR amplification makes it possible to isolate different cDNA fragments that are derived from the same cDNA. The frequency with which one finds multiple cDNA fragments derived independently from the same mRNA provides an estimation of the complexity of all differentially regulated genes (Wang & Brown, 1991). This conclusion is supported by the following considerations. First, the vast majority of genes can generate more than one fragment within the size range that can be amplified efficiently by PCR, which is between 0.15 and 0.8 kb in this procedure. A computer-based analysis of 45 randomly selected *Xenopus laevis* cDNAs showed that two restriction enzymes digests (*Alu* I alone and *Alu* I plus *Rsa* I) generated multiple fragments within the size range for PCR amplification (0.15 to 0.8 kb) for 43 of the 45 cDNAs. Second, the probability that any individual gene will generate PCR-amplifiable cDNA fragments after restriction digestion is assumed to be the same as that of every other gene. Third, PCR amplification of the same cDNA fragment is identical in the two parallel PCR-based subtractions, particularly when the same linker is used for preparations of both cDNA populations. Thus, each cDNA can be divided into multiple fragments that are analogous to different alleles in a genetic screen. Finding multiple cDNA fragments from the same mRNA is analogous to identifying multiple alleles of a gene in a genetic screen. Using the same probability calculations, the total number of the differentially expressed genes can be estimated.

3.2.4.4.3 Presentation of typical results

The gene expression screen is effective in suppressing both low abundance and high abundance cDNA fragments that are common while enriching cDNA fragments that are derived from differentially expressed mRNA (Wang & Brown, 1991; Wang *et al.* 1997). The subtractive hybridisation is considered complete when two cDNA populations no longer cross hybridise to each other (see figure 3.15). The enriched cDNA sequences are cloned into a plasmid vector, such as pBluescript from Stratagene (La Jolla, California, USA), for sequencing. The en-

Figure 3.15 A comparison of enriched cDNA for up-regulated genes (+6 cDNA) and enriched cDNAs for down-regulated genes (–6 cDNA) by PCR Southern blot. See figure 3.14 for the definition of +6 cDNA and –6 cDNA. Typically, 1.5 microgram of PCR-amplified +6 cDNA or –6 cDNA was loaded in each lane of an agarose gel. After Southern transfer, the filter was hybridised with radioactively labelled +6 cDNA (left panel). After stripping off the hybridised probes, the same filter was probed with –6 cDNA (right panel).



riched cDNAs can be used as probes to identify plasmids containing cDNA fragments that are enriched by subtractive hybridisation. Northern blot analyses or other independent methods are used to verify whether a cDNA has been derived from a gene that is truly differentially expressed or not.

3.2.4.4.4 Examples

The gene expression screen was developed initially to identify genes that are involved in tadpole tail resorption during amphibian metamorphosis, a thyroid hormone-controlled process (Wang & Brown, 1991). Total RNA samples were isolated from pre-metamorphic tadpoles treated with thyroid hormone or vehicle for 24 hours. The synthesis of cDNA, restriction enzyme digestion, PCR amplification, and subtractive hybridisation was performed as described in figure 3.14. After three cycles of subtractive hybridisation, cDNA populations enriched either for the genes that are up-regulated or genes that are down-regulated by thyroid hormone were amplified and subsequently cloned into a plasmid for transformation into competent *Escherichia coli* cells. The transformed bacteria colonies were screened using probes prepared from the same enriched cDNA used in cloning. About one third of the colonies contained cDNA fragments that were enriched by subtraction and hybridised to the probes. In contrast, the negative colonies that did not hybridise with the probes contained very low abundance cDNA fragments. These cDNA fragments were derived mainly from the common mRNAs and were suppressed but not eliminated by the subtraction. All of the positive colonies contained DNA fragments from differentially regulated genes (Wang & Brown, 1991).

A significant percentage of positive clones in the enriched cDNA libraries were derived from a small number of abundant cDNA fragments. These cDNAs can be removed from the tracer DNA by short hybridisation using these cDNA fragments mixed with the original PCR-amplified cDNA as driver (figure 3.14). When this is done, additional differentially regulated cDNA fragments will be enriched and identified (Wang & Brown, 1991).

Screening of the enriched cDNA libraries identified a total of 30 non-cross-hybridising cDNA fragments, five from the genes that are down-regulated and 25 from the genes that are up-regulated by thyroid hormone in tadpole tail of *Xenopus laevis* (Wang

Table 3.4 Summary of thyroid hormone up-regulated genes in the tail of *Xenopus laevis* (Wang & Brown, 1991). The identified genes are up-regulated by thyroid hormone at least six-fold and are expressed at the level of at least ten copies mRNA per cell. By applying Poisson distribution analysis, it is possible to predict that there are about a total of 30 up-regulated genes.

<i>Number of genes</i>	<i>Number of isolated cDNA fragments</i>
10	1
4	2
1	3
1	4

& Brown, 1991). Multiple cDNA fragments mapped to six different up-regulated genes and one down-regulated gene (table 3.4). By applying a Poisson distribution analysis to the frequency with which multiple cDNA fragments were found for any gene, it was estimated that there were about a total of 30 up-regulated genes that are induced at least six-fold by thyroid hormone having the abundance of ten copies of mRNA per cell. A total of 16 out of the estimated 30 up-regulated genes were identified. The identification of thyroid hormone-response genes provides a basis for elucidating the molecular and cellular mechanisms of amphibian metamorphosis.

Another example of the application of the gene expression screen has been the identification of androgen-response genes in the rat ventral prostate (Wang *et al.*, 1997). Twenty-five out of the estimated 56 genes that are up-regulated by androgens and four of the less than ten genes that are down-regulated by androgen in the rat ventral prostate were isolated. Identification of these androgen-response genes made it possible to characterise the downstream events triggered by androgens leading to differentiation, proliferation, and apoptosis of prostatic cells.

Use of the gene expression screen in other studies led to the identification of genes involved in limb development and intestinal remodelling during amphibian metamorphosis and of genes associated with *Drosophila* embryogenesis (Buckbiner & Brown, 1992; Casal & Leptin, 1996; Shi & Brown, 1993).

3.2.4.4.5 Discussion

The advantages of gene expression screen include its ability to identify low abundance differentially expressed genes, the dramatic suppression of commonly expressed sequences to the extent that enriched cDNA populations do not cross hybridise to each other, and isolation of multiple cDNA fragments from the same mRNA (Wang & Brown, 1991). The use of restriction enzymes in the digestion of cDNA is critically important for the PCR-based subtraction, which is also adopted in other PCR-based cDNA subtraction methods, such as the linker capture subtraction (Yang & Sytkowski, 1996). In our experience, there were virtually no false-positives in subtractive enriched cDNA clones. Furthermore, gene expression screen also provides an estimation for the total number of genes that are differentially expressed between two closely related mRNA populations.

The gene expression screen consists of multiple steps and can be technically challenging. In comparison, DNA microarrays are now commercially available

and are very powerful in identifying differentially expressed genes. However, DNA microarrays require collections of a large number of cDNA sequences or expressed sequence tags, which are presently only available in a limited number of species including human, mouse, rat, and a few other organisms. Thus, DNA microarrays are not applicable in organisms without the availability of a large number of cDNA sequences or expressed sequence tags. In contrast, the gene expression screen and other subtractive hybridisation methods can be applied to identify differentially expressed genes related to developmental, physiological, or pharmacological events in any organism.

It is important to point out that differential regulation of any gene is suggestive but not proof for its functional involvement (Wang & Brown, 1991). One of the most challenging tasks in studying a biological process initiated by identification of differentially regulated genes is to select amongst many candidate genes for functionally significant genes. Separate functional assays are clearly necessary to define the significance of a gene in the biological events under study. Nevertheless, identification of the majority of up and down-regulated genes associated with some biological event provides a solid foundation to further explore the mechanism underlining the biological change.

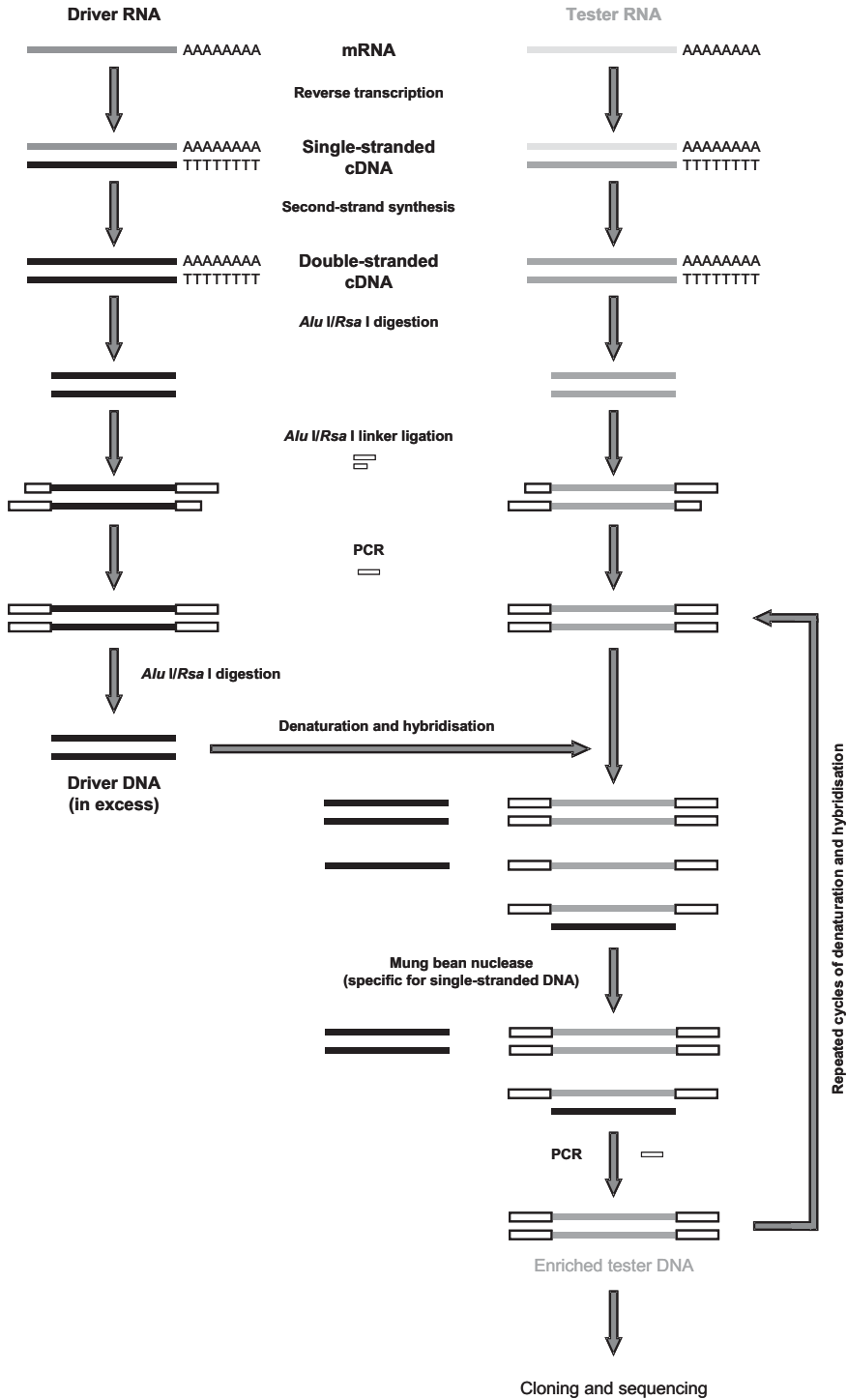
3.2.4.5 LCS – Linker capture subtraction

3.2.4.5.1 Introduction

Linker capture subtraction (LCS) was initially developed by Yang & Sytkowski (1996) to identify and clone genes that are differentially expressed between two prostate cancer cell lines, LNCaP and PC-3, which respond differently to androgen stimulation and possess different tumourigenic and metastatic potentials (Lim *et al.*, 1993; Pretlow *et al.*, 1991). LCS is a simple and efficient method that enriches target cDNA fragments representing genes that are expressed at a higher level in one cell type than in another. It was developed primarily to isolate genes, however, it might also prove to be an effective method to detect transient expression of a pool of genes under specified conditions in the same type of cells or tissues. LCS can be modified further or combined with other molecular techniques (such as DNA microarray technology) to enhance its capability and versatility and to reduce the total time of the experiment.

3.2.4.5.2 Principles and basics

Numerous methods have been developed to isolate DNA sequences representing genes that are either up or down-regulated in two populations of cells, tissues of different types, or cells of the same type but under different physiological or pathological characteristics. These methods include a variety of approaches that either use PCR to selectively amplify target sequences or to eliminate common sequences by subtractive hybridisation or a combination of both. Each of these methods has advantages and limitations.



◀ **Figure 3.16 Schematic illustration of the linker capture subtraction (LCS) method.** Driver and tester mRNAs are converted into double-stranded cDNA pools using reverse transcriptase. After restriction with *Alu* I and *Rsa* I, linkers are attached (solid bars). After PCR amplification of these both pools, the linkers are removed from the driver DNA. Driver and tester cDNAs are hybridised (with driver in

excess) followed by digestion of single-stranded DNA using mung bean nuclease. Three types of double-stranded species result, but only the tester-specific homohybrids still contain the linker oligonucleotide and, therefore, are able to be amplified by PCR. Hybridisation with excess driver DNA is then repeated for further enrichment of tester species.

LCS applies both subtractive hybridisation and PCR amplification to enrich those sequences that are more abundant in one sample than the other, and, thus, does not rely on a kinetic mechanisms of enrichment (for a detailed protocol, see Yang & Sytkowski, 1998).

Briefly, LCS starts with conventional methods of purifying mRNA from two samples of interest and proceeds to the synthesis of double-stranded cDNA (see figure 3.16). The two samples to be subtracted are thus referred to as 'tester' and 'driver' as is usual in subtractive hybridisation methods. The tester cDNA represents the experimental sample that contains transcripts that are either absent from or reduced in driver cDNA. Both tester and driver cDNA are cleaved by restriction enzymes that recognise four base pairs of nucleotide sequence to reduce their average sizes, leaving blunt ends. The resulting cDNA fragments are then selected by purifying from an agarose gel and are ligated to a synthetic DNA linker that contains sites for the restriction enzymes *Alu* I and *Sac* I (or other sites). The linkers can be used as primers for PCR reaction. These cDNA fragments with linkers attached to both ends are purified and amplified by PCR. The driver cDNA is digested successively with restriction enzymes *Alu* I and *Sac* I to remove the linker sequences. Hybridisation is performed in the presence of excess driver DNA to remove sequences from the tester DNA that represent genes with similar levels of expression. The DNA fragments that are present in both tester and driver will be eliminated by single strand DNA-specific nuclease digestion using mung bean nuclease, and the sequences highly represented in the tester DNA will be amplified and enriched by PCR. The subtraction product is used as tester DNA in two more rounds of subtraction to enrich target sequences further. The final captured cDNA sequences are cloned into a plasmid and propagated as a DNA library in *Escherichia coli*. The candidate DNA clones in the subtractive library that represent differentially expressed genes are either screened by direct sequencing or used as probes in a Northern analysis to confirm their differential expression.

3.2.4.5.3 Results

Restriction digestion of the cDNA yields fragments of various sizes that resolve on the agarose gel as a smear although distinct bands are observed after successive subtraction (figure 3.17). These DNA fragments represent specific sequences that become more and more abundant in the subtraction product over repeated rounds of subtraction. The candidate sequences are finally captured by insertion into any conventional plasmid cloning vector. The identities of these genes are de-

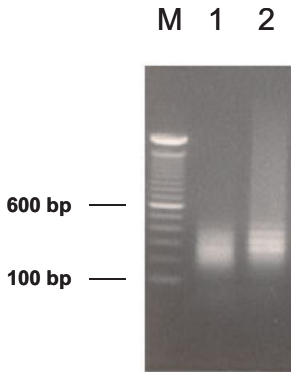


Figure 3.17 Enrichment of specific cDNA sequences using the linker capture subtraction (LCS) method. LCS enriches specific cDNA sequences from two prostate cancer cell lines. Subtraction product DNA was amplified by PCR and resolved on a three percent agarose gel by electrophoresis. *Lane M*: 100 base pairs (bp) DNA marker. *Lane 1*: product from the second subtraction. *Lane 2*: product from the third subtraction.

terminated by direct DNA sequencing and their differential expression is verified by a Northern analysis or other molecular methods.

3.2.4.5.4 Examples of successful application of LCS

Using LCS, numerous genes were identified that are differentially expressed between prostate cancer cell lines LNCaP and PC-3 (Yang *et al.*, 1998). Figure 3.18 shows a Northern blot result of a candidate gene. It is expressed at a significantly higher level in LNCaP cells than PC-3 cells. Its differential expression has been further confirmed at protein level as well (unpublished data by Yuxun Wang and Arthur J. Sytkowski. Currently, functional studies are performed to identify its role in the progression of prostate cancer.

Other investigators have reported successful uses of LCS. For example, Guan *et al.* (2001) have used LCS to identify genes with altered expression in kainic acid-induced spontaneous recurrent seizures and succeeded in isolating a rat epilepsy-related gene (ERG) called ERG1. It is differentially expressed in brain cells. Dave N. T. Aryee and co-workers applied LCS to gene expression patterns in type

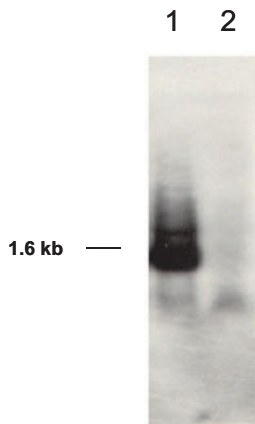


Figure 3.18 Differential expression of a gene identified using the linker capture subtraction (LCS) method. LCS identified a gene that was differentially expressed between prostate cancer cell lines LNCaP (L) and PC-3 (P). Northern blot analysis verified the differential expression of this gene. Total RNA from either cell was electrophoresed and probed with radiolabelled cDNA that was isolated by three rounds of LCS.

1 versus type 2 Ewing's sarcoma family tumours cell lines (Aryee *et al.*, 2000). Pools of cDNA samples from many cell lines of the same types were prepared as tester and driver DNA to identify genes that show differential expression between two types of cell lines instead of between two individual cell lines. In addition, candidate genes have been isolated by LCS studies of primary ovarian cancer specimens from patients exhibiting different clinical responses to treatment (unpublished data by Dimitrios Spentzos and Arthur J. Sytkowski).

3.2.4.5.5 Discussion

A frequent objective of current biological and medical investigations is the determination of specific spatial and temporal expression of genes of interest. With the recent introduction of DNA microarray technology, it is possible now to probe comprehensively the expression of large numbers of genes. Also, many new protocols have been reported for identifying and isolating differentially expressed genes combining DNA microarray technology with other molecular methods (Beck *et al.*, 2001). However, DNA microarray technology has its limitations. It can be used only when the genome sequence or collection of cDNA of the species under study is known. Subtraction-based methods such as LCS are still valuable tools for analysing gene expression and cloning differentially expressed genes. The major advantageous features of LCS in comparison to other subtraction-based methods are its efficiency, minimum cost, and its use of relatively simple molecular biological techniques. Therefore, it does not require extensive experience in order to obtain reliable results.

LCS is very effective for capturing differentially expressed sequences as demonstrated by the examples described above. It does not require expensive equipment or computer software. The investigator has the flexibility to use alternative reagents and to make modifications for specific needs.

However, clonal redundancy is one of the drawbacks of LCS and is the result of relative abundance of mRNA in the original samples. However, it can be reduced to a minimum by adjusting hybridisation condition. It is also possible to isolate rare transcripts using LCS.

LCS can be linked to DNA microarray assay to provide fast high-throughput screening of many samples. Enriched subtraction product DNA fragments can be directly labelled by PCR in the presence of fluorescent nucleotide analogues and subsequently applied in DNA microarray assays as has been done using other subtractive hybridisation-based approaches. Those genes that are up-regulated will produce stronger hybridisation signals on the array and will be readily identified. This would greatly reduce the abundance of common sequences in the probe that are not differentially expressed and not the focus of study. Therefore, the unchanged hybridisation signals in the conventional microarray assay will be suppressed. This facilitates post-hybridisation data analysis and interpretation.

3.2.4.6 RaSH – rapid subtraction hybridisation

3.2.4.6.1 Introduction

The identification and temporal expression profiles of differentially expressed genes are providing important information on the molecular determinants of normal cellular physiology and relevant insights into various disease processes, including cancer, cardiovascular disease and neurodegeneration (de la Monte *et al.*, 1995; Heller *et al.*, 1997; Leszczyniecka *et al.*, 2001; Metcalfe *et al.*, 2001; Zhang *et al.*, 1997). A number of approaches have been applied to accomplish this objective, such as differential display RT-PCR (DDRT-PCR) (Liang & Pardee, 1992; Liang *et al.*, 1995), subtraction hybridisation (Jiang & Fisher, 1993; Sagerström *et al.*, 1997), representational difference analysis (RDA) (Hubank & Schatz, 1994), RNA fingerprinting by arbitrarily primed PCR (RAP-PCR) (McClelland & Welsh, 1994), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995), reciprocal subtraction differential RNA display (RSDD) (Kang *et al.*, 1998), electronic subtraction (Wan *et al.*, 1996), and combinatorial matrix gene analysis (Schena *et al.*, 1995). Of these approaches, subtraction hybridisation represents a particularly useful methodology for identifying and cloning target specific differentially expressed genes, without *a priori* information about their functional characteristics or biochemical properties (Jiang & Fisher, 1993; Sagerström *et al.*, 1997; Sambrook *et al.*, 1989). A drawback to this approach involves the technical complexities of successfully performing this method, *i.e.* the ability to obtain efficient subtraction and to identify both rare and abundant transcripts that differ in copy number in cells (Sagerström *et al.*, 1997; Sambrook *et al.*, 1989; Wan & Erlander, 1997). The conventional subtraction hybridisation approach involves hybridisation of first strand cDNAs produced from tester mRNA populations with mRNAs obtained from driver nucleic acids. Hydroxyapatite column chromatography or biotin/avidin extraction is used to remove single-stranded non-hybridised cDNAs and provide the necessary subtracted templates for second strand cDNA synthesis. Limitations in this scheme include an ability to analyse only a fraction of the numerous changes occurring in gene expression, the need for relatively large amounts of mRNA (ten micrograms poly(A)⁺ mRNA) and the laborious nature of this scheme (Sagerström *et al.*, 1997; Sambrook *et al.*, 1989). Some relief has been obtained by first generating cDNA libraries in phage plasmid vectors and using these as both testers and drivers to generate subtracted cDNA libraries (Jiang & Fisher, 1993). However, constructing high-quality and representative cDNA libraries is not a trivial task and success is not always realised. The use of PCR-based cDNA subtraction can accelerate the procedures for cDNA library preparation, but this approach also contains a number of tedious steps during or after hybridisation (Diatchenko *et al.*, 1996), and efficiency of subtraction may vary and not be consistent for all applications.

Rapid subtraction hybridisation (RaSH) simplifies the process of cDNA subtraction compared to other subtraction hybridisation methodologies (Jiang *et al.*, 2000). Moreover, RaSH provides a high level of subtraction, which can be experimentally manipulated, and is very cost efficient. In RaSH, cDNA libraries are prepared from double-stranded cDNAs that are digested into small fragments using specific re-

striction endonucleases, ligated to adapters, and PCR-amplified followed by incubation of tester and driver PCR fragments. The RaSH method has been applied to a number of experimental systems resulting in the identification and cloning of genes associated with human melanoma differentiation (Jiang *et al.*, 2000; Kang *et al.*, 2001), resistance and sensitivity to human immunodeficiency virus type-1 (HIV-1) infection (Simm *et al.*, 2001), and to the genes that are up-regulated and down-regulated in human fetal astrocytes as a function of HIV-1 infection or treatment with HIV-1 envelope glycoprotein (*gp120*) (Su *et al.*, 2002). These findings document that RaSH will prove amenable to numerous experimental systems and will facilitate the identification of relevant genes associated with and potentially causative of significant biological phenomena.

3.2.4.6.2 Principles and basics

RaSH consists of three key steps: (i) preparation of tester and driver cDNAs for subtractive hybridisation, (ii) performance of subtractive hybridisation, followed by cloning of subtractive cDNA fragments into vectors, and (iii) screening differentially expressed message by reverse Northern and Northern blot hybridisation (figure 3.19) (Jiang *et al.*, 2000). Duplex tester and driver cDNAs are digested with four-base (*Dpn* II, GATC) or four and a half-base (*Eco* RII, CCWGG) recognising restriction enzymes and ligated to adapters containing an additional restriction enzyme site (*Xho* I) allowing cloning into vector. Fragmented cDNAs are amplified by 25 cycles of PCR and only tester cDNAs are digested with the cloning restriction enzyme *Xho* I. Subtractive hybridisation is performed by incubation of restriction enzyme-digested tester cDNA and driver cDNA in molar excess. Hybridisation of common cDNAs between tester and driver masks *Xho* I restriction sites and prevents the hybridised cDNAs from being cloned into vector. In this manner, unique cDNAs in the tester containing exposed *Xho* I sites are cloned into plasmid and propagated in bacteria without further manipulation. Randomly selected clones from the RaSH library are PCR-amplified and evaluated by reverse Northern hybridisation, in which cDNA fragments immobilised on a membrane are hybridised with first strand cDNA (figure 3.20) (Huang *et al.*, 1999a; Kang *et al.*, 1998). Northern blot hybridisation is used to confirm the differential expression of positive clones identified using reverse Northern analysis.

RaSH uniquely combines various aspects of previously described subtractive cloning methods including cDNA fragmentation by restriction enzyme digestion, PCR, selection of subtracted cDNA species by matching restriction sites and plasmid library generation, thereby reducing cost and decreasing the labour normally involved in constructing conventional subtractive cDNA libraries (Jiang *et al.*, 2000; Sagerström *et al.*, 1997). In RaSH, subtraction hybridisation involves digestion of the cDNA with a frequent cutting restriction enzyme (four-base recognition site yielding 256 fragments on average). Cleavage of cDNA with a four-base cutter and adapter ligation not only promotes non-biased and specific subtractive hybridisation, but it also prevents differential amplification by template size and preferential primer binding during PCR. However, since fragmentation of cDNA inevitably results in redundancy owing to gene size, a four and a half-base recognising restric-

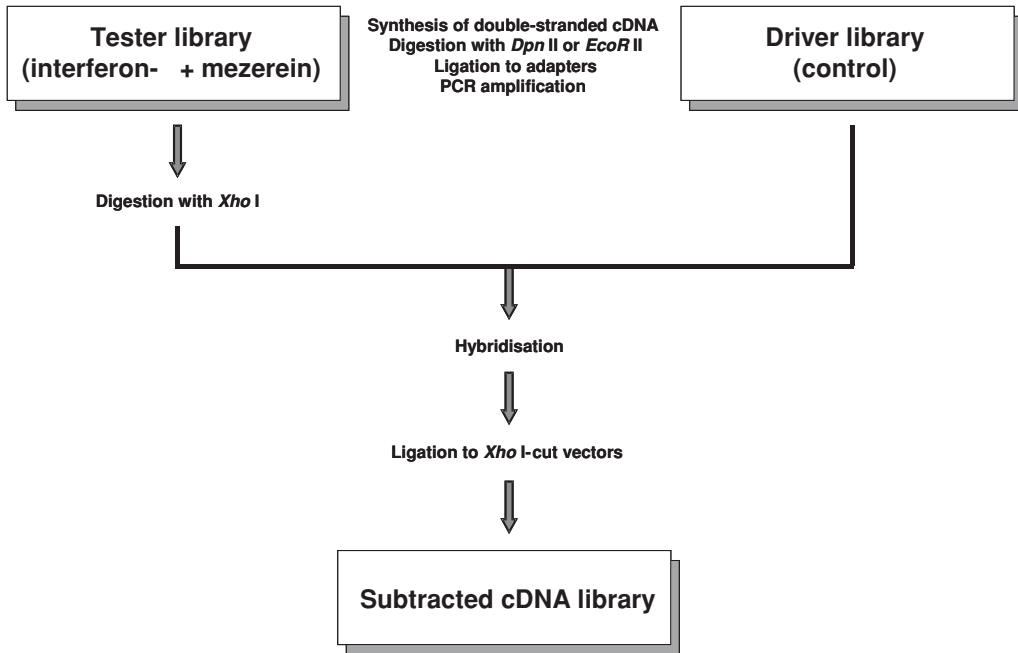


Figure 3.19 Schematic outline of the rapid subtraction hybridisation (RaSH) protocol. This scheme involves construction of tester (interferon- β and mezerein-treated) and driver (non-treated) HO-1 human melanoma cell libraries followed by digestion of only the tester cDNA with *Xho* I. Following hybridisation, differentially expressed sequences are cloned into *Xho* I-digested vectors resulting in a subtracted cDNA library enriched for melanoma differentiation-

associated genes displaying elevated expression. By using the control HO-1 human melanoma cell library as the tester and the interferon- β and mezerein-treated human melanoma cell library as the driver, RaSH can also be used to produce a subtracted cDNA library enriched for genes down-regulated during terminal differentiation. The figure is reproduced from Jiang *et al.* (2000) with permission from *Proceedings of the National Academy of Sciences of the USA*.

tion enzyme (*Eco* RII) that cleaves DNA two-fold less frequently on average than a four-base cutter is also used to digest cDNA to compare the effect of fragment size on redundancy of gene identification. The restricted cDNA fragments are ligated to adapter and PCR-amplified with primers based on adapter sequence. PCR is employed only once to amplify cDNAs in order to reduce the amount of starting material required for subtraction. Since extensive cycles of PCR can normalise the differences in gene expression levels, appropriate numbers of PCR cycling (25 cycles in this case) are used to maintain the original proportion of gene expression levels by comparing post-PCR and pre-PCR gene expression levels.

The most labour-saving feature of RaSH is positive selection of non-hybridised unique tester cDNA fragments during the cloning step, which has been successfully applied for subtracted cDNA library construction (Jiang & Fisher, 1993; Su *et al.*, 1997). By hybridisation, the cloning restriction site (*Xho* I) in driver/tester hybrid or single-stranded tester and driver cDNA is not exposed for subsequent

cloning into plasmid. The only fragment that derives from the tester contains intact cohesive ends, which can therefore be cloned into appropriately digested vector. Thus, the unique cDNA in tester is selected by matching restriction sites in the cDNA fragment and in the cloning vector. Subtractive hybridisation is based on a mass-driven reaction using excess amounts of driver cDNA and the reaction volume is maintained at a minimal level to facilitate hybridisation. Finally, with recent advances in bacterial transformation techniques, the subtracted cDNA library can be constructed as a plasmid library instead of a phage library, which significantly reduces cost and labour. In order to ensure the incorporation of the cDNA insert, the digested vector is treated with calf intestinal alkaline phosphatase and blue/white screening is used for detecting positive clones. In addition, an excess amount of vector is used to clone every subtracted cDNA fragment present in the subtracted tester library.

A PCR-based subtracted cDNA library technique called suppression subtractive hybridisation (SSH) also employs hybridisation with short cDNA fragments digested with a four-base recognising restriction enzyme (*Rsa* I) (Diatchenko *et al.*, 1996). Exponential PCR amplification of commonly expressing cDNA fragments is selectively blocked by the suppressive effect of a panhandle structure, which is formed by ligation of a specifically designed adapter. Thus, the selective exponential amplification of tester-specific cDNA in PCR is the basis of subtraction in SSH, which is fundamentally different from RaSH which utilises mass-driven subtractive hybridisation and selective cloning into vector (Jiang *et al.*, 2000).

The differential expression of RaSH-enriched cDNAs is confirmed by reverse Northern hybridisation followed by Northern blot hybridisation (Huang *et al.*, 1999a; Kang *et al.*, 1998). Randomly selected bacterial clones are amplified by PCR with the T7/T3 promoter primers flanking the cDNA inserts in the cloning vector. The PCR-amplified fragments can be immobilised on a nylon membrane either by dot blotter or by transferring directly from agarose gels. Immobilised cDNAs are hybridised with first strand cDNA reverse transcribed from source RNA samples. Control cDNAs such as *GAPDH* and β -*actin* should be included to monitor the hybridisation efficiency. Pre-screening for differential expression of RaSH-identified cDNA fragments by reverse Northern hybridisation can greatly reduce the cost and effort involved in verifying genuine differential expression by Northern blot hybridisation.

3.2.4.6.3 Presentation of typical results

The first application and proof of principle for the RaSH approach involved the identification of differentially expressed genes regulated during induction of terminal differentiation of human melanoma cells following combination treatment with interferon- β and the protein kinase C activator mezerein (Jiang *et al.*, 2000). Combination treatment of HO-1 human melanoma cells induces irreversible growth arrest, loss of tumourigenic properties and terminal cell differentiation (Fisher *et al.*, 1985 and 1986; Jiang *et al.*, 1993). These profound changes in cellular phenotype are regulated by temporal changes in gene expression profiles (Huang *et al.*, 1999; Jiang & Fisher, 1993; Jiang *et al.*, 1994). Previously, a temporally-spaced

subtracted cDNA library had been constructed by subtraction of untreated control HO-1 human melanoma cell cDNAs derived from interferon- β and mezerein treated HO-1 cell cDNAs and analysed for differential expression by random clone selection, high-density cDNA microarrays and differential screening by reverse Northern (Huang *et al.*, 1999; Jiang & Fisher, 1993). Twenty-three, 112 and 65 differentially expressing clones potentially associated with melanoma differentiation were identified by random cloning of 70 phage colonies, high-density cDNA microarray of 1,000 cDNAs (probes larger than 500 base pairs) and differential screening of 400 cDNAs by reverse Northern (probes smaller than 400 base pairs), respectively (Huang *et al.*, 1999; Jiang & Fisher, 1993). Based on the extensive characterisation of this differentiation model system, the application of RaSH provided a means of verification of the efficacy of this new method and its capacity to identify similar and additional differentially expressed cDNAs.

In order to encompass the complete spectrum of gene expression changes occurring during melanoma differentiation induced by interferon- β and mezerein treatment, RNA samples were collected at two, four, eight, 16 and 24 hours after treatment, pooled and subjected to RaSH. Initial screening of 32 colonies from each RaSH library constructed by *Dpn* II (*Dpn*-sLib library) or *Eco* RII (*Eco*R-sLib library) digestion by reverse Northern demonstrated that about 50 percent of the colonies contained cDNAs representing differentially expressed genes (figure 3.20) (Jiang *et al.*, 2000). Subsequent analysis by Northern blot hybridisation confirmed differential expression, in most cases with a concordance rate of about 89 percent, as demonstrated by reverse Northern hybridisation. Thus, about 45 percent of RaSH colonies appear to harbour cDNAs representing differentially expressed genes associated with melanoma differentiation. The about 45 percent positive concordance rate of RaSH is significantly higher than subtracted cDNA library analysis using the same model system (an average of about 20 percent concordance rate) (Huang *et al.*, 1999; Jiang & Fisher, 1993).

The enrichment factor for differentially expressed genes by the RaSH protocol is calculated by comparing Southern blot signal intensity of a specific gene in the pre-

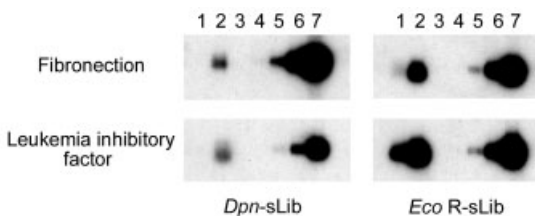


Figure 3.20 Reverse Northern blot analysis of differentially expressed sequence tags identified by rapid subtraction hybridisation (RaSH). PCR-amplified products from bacterial clones of both RaSH-derived subtracted libraries, *Dpn*-sLib and *Eco* R-sLib, were dot-blotted onto nylon membranes and were probed

with ^{32} phosphorus-labelled cDNA reverse transcribed from RNA samples of control or interferon- β and mezerein-treated treated HO-1 human melanoma cells. The figure is reproduced from Jiang *et al.* (2000) with permission from *Molecular and Cellular Differentiation*.

subtracted cDNA library and the subtracted cDNA library (figure 3.21) (Jiang *et al.*, 2000). The enrichment factors for the leukemia inhibitory factor (LIF) cDNA are 544-fold in *Dpn*-sLib library and 96-fold in *Eco* R-sLib library, while the enrichment factors for fibronectin are 75-fold in *Dpn*-sLib library and six-fold in *Eco* R-sLib library. The enrichment factors for fibronectin and leukemia inhibitory factor directly demonstrate not only the efficiency of RaSH, but also the enhanced subtraction hybridisation efficiency of the *Dpn* II-cDNA libraries.

Since RaSH employs frequent cutting restriction enzymes, the incidence of a specific gene in a subtracted library may increase with the size of the gene. As anticipated, both fibronectin (eight kilobases) and leukemia inhibitory factor (7.6 kilobases) are identified multiple times in both *Dpn*-sLib library (five and two times, respectively) and *Eco* R-sLib library (seven and three times, respectively) (Jiang *et al.*, 2000). To assess the effect of cDNA fragment size on the redundancy of identification of specific genes in the subtracted libraries, *Eco* R-sLib library (predicted 512 base pairs fragments) was compared with *Dpn*-sLib library (predicted 256 base pairs fragments). Theoretically, there should be half as many fragments generated by *Eco* RII digestion. In this context, redundancy in *Eco* R-sLib library should be lower than in *Dpn*-sLib library. An unexpected observation was that the *Eco* R-sLib library exhibited consistently higher redundancy at both an individual gene level and on a whole library level (about 61 percent uniqueness) than that of *Dpn*-sLib library (about 69 percent uniqueness) (Jiang *et al.*, 2000). In fact, although not identified in *Eco* R-sLib library, short but strongly expressing messages such as 1-8U (0.8 kilobases) were identified three times in the *Dpn*-sLib library. Thus, redundancy in RaSH is ascribed to a combination of effects resulting from gene size and expression level, rather than simply being a reflection of the fragment size generated by restriction digestion. In addition, restriction digestion in the

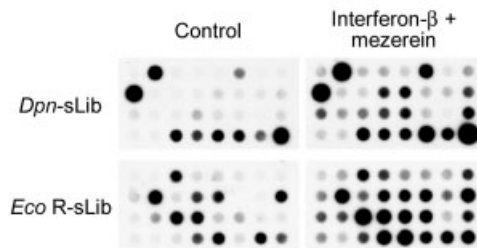


Figure 3.21 Determination of the amounts of fibronectin and LIF fragments in PCR-cDNA libraries. Relative amount of fibronectin and leukemia inhibitory factor fragments in PCR libraries was determined by comparison of the signal intensity of fibronectin and leukemia inhibitory factor in specified cDNA libraries (*Dpn*-sLib and *Eco* R-sLib) with the signal intensity of defined amounts of the cDNA fragments by Southern blot hybridisation. Lane 1: 100 nanograms PCR-cDNA library from untreated cells;

lane 2: 100 nanograms PCR-cDNA library from interferon- β and mezerein-treated cells; lane 3 to 7: increasing amounts of cDNA fragments of fibronectin and leukemia inhibitory factor (0.001, 0.01, 0.1, one and ten nanograms, respectively). PhosphorImager™ (Amersham Biosciences, Inc., Piscataway, New Jersey, USA) scanning determined hybridised signal intensity. The figure is reproduced from Jiang *et al.* (2000) with permission from *Proceedings of the National Academy of Sciences of the USA*.

RaSH approach does not significantly increase redundancy in comparison with the uniqueness of a temporally-spaced-subtracted cDNA library from the same system (on average about 71 percent) (Huang *et al.*, 1999; Jiang & Fisher, 1993).

Analysis of about ten percent of the RaSH libraries identified 25 genes from *Dpn*-sLib library and 17 genes from *Eco* R-sLib library. RaSH recognised genes previously cloned by screening temporally-spaced interferon- β and mezerein subtracted HO-1 cDNA libraries, including interferon-inducible genes (major histocompatibility complex, HLA, genes; viperin, *cig5*; 1-8U; guanylate-binding protein 1, *GBP1*), mezerein-inducible genes (prolactin receptor-associated protein), a differentiation factor (leukemia inhibitory factor, LIF), genes involved in growth inhibition or apoptosis (cyclin-dependent kinase inhibitor p21, *mda-6*; growth arrest and dna damage-inducible gene 34, *GADD34*) and cytoskeleton and extracellular matrix genes (fibronectin, integrin $\alpha 5$) (Huang *et al.*, 1999; Jiang & Fisher, 1993; Jiang *et al.*, 2000). Although both RaSH libraries and conventional subtracted cDNA libraries have not been exhaustively analysed, some genes are uniquely identified by each method. Seven cDNAs identified using RaSH are without representation in current DNA databases. Of the 25 distinct genes identified from the *Dpn*-sLib library and 17 distinct genes cloned from the *Eco* R-sLib library, only three genes were common to both libraries, partly because of incomplete analysis of complete libraries. RaSH was also applied to cloning differentially expressed genes in an HIV-1-resistant T cell clone (Simm *et al.*, 2001). RaSH with *Dpn* II identified eleven up-regulated and six down-regulated genes in comparison between HIV-1 resistant and sensitive T-cells. These results clearly demonstrate the utility of RaSH in cloning differentially expressed genes of functional relevance in diverse biological systems. In addition, RaSH has recently been applied to define the spectrum of gene expression changes occurring in normal human fetal astrocytes following infection with HIV-1 or treatment with HIV-1 envelope glycoprotein (*gp120*) (Su *et al.*, 2002). This approach resulted in the identification of 15 astrocyte-elevated genes, representing both previously identified as well as novel genes not reported in current databases (Su *et al.*, 2002). Specific genes modulated by HIV-1 in astrocytes identified using RaSH may be important in regulating astrocyte physiology, potentially contributing to HIV-1-associated dementia (HAD).

3.2.4.6.4 Discussion

Generation of subtractive cDNA library is one of the most widely used methods to clone differentially expressed genes in order to analyse molecular determinants that are fundamental to various physiological changes (Jiang & Fisher, 1993; Sagerström *et al.*, 1997; Su *et al.*, 1997). However, the downside of the subtractive cDNA library technique is the complicated nature of this procedure, which requires significant cost, intensive labour and technological proficiency (Jiang & Fisher, 1993; Sagerström *et al.*, 1997; Sambrook *et al.*, 1989). RaSH, by its unique subtraction cDNA library protocol, significantly reduces the cost and labour burden of the conventional subtractive cDNA library technique. Application of RaSH in defining molecular determinants of differentiating melanoma cells (Jiang *et al.*, 2000; Kang *et al.*, 2001), HIV-1 resistant T-cells (Simm *et al.*, 2001) and genes up-regulated by

HIV-1 infection and HIV-1 envelope glycoprotein treatment in normal human fetal astrocytes (Su *et al.*, 2002) attests to the robustness of this approach in identifying biologically relevant genes displaying altered expression in diverse cellular contexts. In the case of human melanoma cells, initial screening of about ten percent of the *Dpn*-sLib library from differentiating melanoma cells identified 25 unique cDNA species representing differentially expressed genes. Direct extrapolation of this result would suggest that about 250 distinct genes would be identified that display altered expression if the complete library were evaluated (Jiang *et al.*, 2000). If it is assumed that in many contexts the frequency of differentially expressed cDNAs identified by RaSH will be about 200 to 300, and given the fact that about two percent changes in gene expression are projected between closely related cell types (T *versus* B-cells) out of about 30,000 distinct genes (Sagerström *et al.*, 1997), RaSH libraries should cover more than half of all changes. Furthermore, considering that RaSH is applied to physiological changes within a specific cell type by treatment with differentiating reagents, the RaSH library in this system may identify the vast majority of differentially expressed genes. In these contexts, RaSH is unique not only in its simplicity of performance but also in its potency in differential gene identification.

Two restriction enzymes (*Dpn* II and *Eco* RII) that produce different sized cDNA fragments have been used in RaSH library construction to measure the effect of cDNA digestion on redundancy of gene identification. Although *Eco* RII digests cDNA two-fold less frequently than *Dpn* II, for currently unknown reasons the redundancy of *Eco* R-sLib library is somewhat higher than *Dpn*-sLib library (Jiang *et al.*, 2000). Furthermore, the enriching effect of the *Dpn*-sLib library is significantly higher than that of the *Eco*R-sLib library (seven to sixteen-fold higher with respect to the specific genes examined). In addition, a short but abundant message like 1-8U is identified three times in *Dpn*-sLib, but not in *Eco* R-sLib library. Incomplete analysis of *Eco*R-sLib library could be the reason for the failure to identify this gene using the *Eco* R-sLib library. A second possibility is that the spacer recognition site of this gene in *Eco* R-sLib library is less representative than that in *Dpn*-sLib library (Jiang *et al.*, 2000). Considering these results, *Dpn*-sLib library is preferred to *Eco* R-sLib library for use with the RaSH approach.

Use of RaSH in its present form highlighted two potential problems that can readily be addressed by minor modifications in the protocol. Although the majority of RaSH clones contained single inserts, some clones contained more than one insert that were ligated in tandem (Jiang *et al.*, 2000). Multiple inserts can obscure differential expression in screening procedures, such as reverse Northern hybridisation. Moreover, if the gene has not been reported previously, the hybrid molecule can inappropriately serve as a basis for attempting to clone the spurious molecule. Careful consideration, especially with respect to the presence internally of the restriction site used in library construction (*Dpn* II) in the RaSH clone could be used to circumvent this problem. Digestion of a cDNA with a frequent cutter could increase additional redundancy due to cloning different parts of the same gene. Immobilisation of 3'-ends by using biotinylated primers in reverse transcription of mRNA with a cloning site (*e.g.*, *Xho* I) and ligation of adapter with another

cloning site (*Bam* HI) may prove useful in ameliorating this problem of redundant clone isolation and the isolation of clones containing multiple inserts. This modification in the original protocol will also enhance the cloning efficiency of the differentially expressed insert into vector.

3.2.4.7 cDNA-RDA – Representational difference analysis of cDNA

3.2.4.7.1 Introduction

Representational difference analysis of cDNA (cDNA-RDA) is a subtractive hybridisation technique that combines kinetic enrichment with PCR amplification to purify cDNA sequences present in one population, but not in another (Bowler *et al.*, 1999 and 2001; Hubank & Schatz, 1994 and 1999). cDNA-RDA is a flexible and highly effective technique in which target cDNA fragments are sequentially enriched by favourable hybridisation kinetics and subsequently amplified by PCR. The positive selection of differences and removal of sequences common to both groups simplifies the interpretation of results and facilitates identification of the differentially expressed genes. Furthermore, the exponential degree of enrichment achieved by the use of PCR in cDNA-RDA enables the detection of rare transcripts and allows application of the technique to low amounts of starting material

3.2.4.7.2 Principles and basics

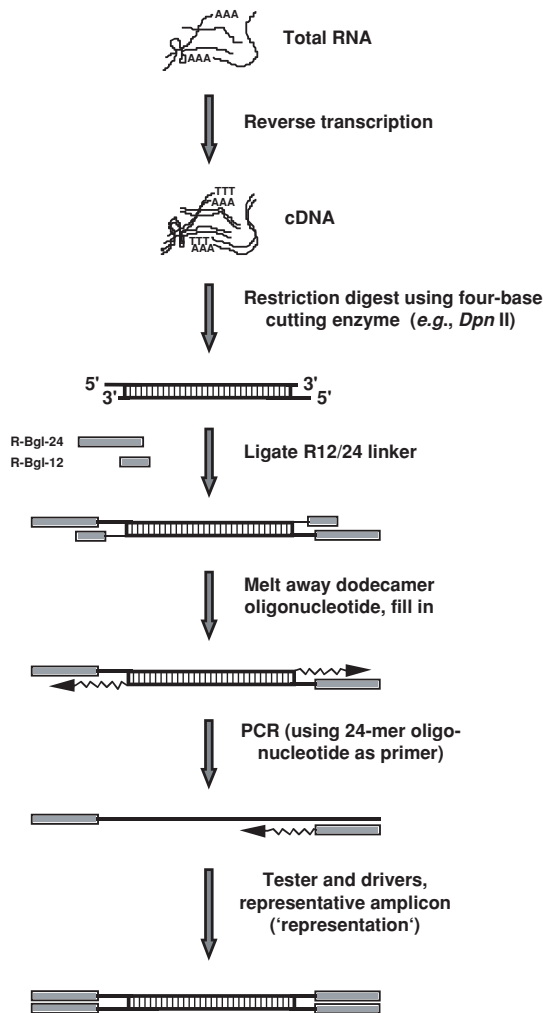
cDNA-RDA is based on the representational difference analysis (RDA) technique previously described by Lisitsyn *et al.* (1993). RDA is a method for the identification of differences between two complex genomes, and was a significant advance in the field of subtractive cloning, bringing together the advantages afforded by both subtractive hybridisation and the PCR-based amplification used in differential display-type techniques. In contrast to both of these methods, cDNA-RDA uses RNA rather than genomic DNA as its starting material and therefore targets those genes that are expressed at the time of RNA isolation.

The cDNA-RDA technique comprises three main stages: (i) The generation of PCR products, representative of the original RNAs from each given population, also known as the representation. The aim of this step is to provide large amounts of representative material for the subsequent steps of the analysis. (ii) The PCR-coupled subtractive hybridisation of the different representations. This results in the depletion or removal of common DNA sequences and the amplification of fragments from the differentially expressed genes. (iii) The cloning and screening of the resulting PCR products.

Examination of differential gene expression using cDNA-RDA requires the sampling of a population of cells grown under the condition(s) of interest, and a population grown under conditions that differ only by those of interest. For example, in a study to identify iron-regulated genes in bacteria, bacteria would be grown under iron-limitation (to provide the so-called 'tester' material), and another population would be grown under identical conditions, except that iron would be made freely available (to provide the 'driver' component). RNA is then extracted from

both populations and used as a template for cDNA synthesis. With eukaryotic material, mRNA may be isolated directly and an oligo(dT) primer is used for priming cDNA first strand synthesis. However, with prokaryotic RNA, given that bacterial mRNAs possess short, or no poly(A) tails, random hexamers must be used for priming cDNA synthesis (Bowler *et al.*, 1999). The derived cDNAs are then cut with a restriction enzyme, and oligonucleotide linker/adapters are ligated onto the fragments. PCR using primers specific to these linkers is then used to amplify each DNA fragment population. After amplification, the linkers are removed with the same enzyme, and the amplified cDNA fragments (the 'representations') are purified (see figure 3.22). For the hybridisation and amplification stages, new oligonucleotide linkers are ligated to the tester DNAs only. Tester and driver DNAs (the latter in excess) are then combined, denatured and allowed to hybridise.

Figure 3.22 Schematic representation of generation of the initial representations in representational difference analysis of cDNA (cDNA-RDA). cDNA is synthesised by oligo (dT) priming – either of total RNA or purified mRNA (for eukaryotic material), or random-primed reverse transcription of total RNA (for bacteria). This cDNA is then digested with a restriction enzyme (e.g., *Dpn* II, recognition site 5'-GATC-3'). The oligonucleotide linker pair, R-Bgl-12 and R-Bgl-24 (R12/24), is then ligated to the resulting cDNA fragments (sequences of the primers are as follows: R-Bgl-12, 5'-GATCTGCGGTGA-3'; R-Bgl-24, 5'-AGCACTCTCCAGCTCTCACCGCA-3'). Since the R-Bgl-12 dodecamer oligonucleotide is not phosphorylated, it does not become covalently attached, and accordingly dissociates at 72 °C. The *Taq* DNA polymerase is then added, and 'fills in' the sequence complementary to the R-Bgl-24 24-mer oligonucleotide. This creates a binding site for the R-Bgl-24 oligonucleotide used to prime DNA synthesis in the subsequent PCR. The figure is adapted from Bowler *et al.* (2001) with permission from Humana Press.



PCR using primers specific to the new oligonucleotide linker extensions results in enrichment of sequences unique to the tester DNA population. By repeating these steps, the degree of enrichment is increased. The result is a number of PCR products (difference products) that represent the messages unique to the tester population. These can then be cloned and characterised. The principle of cDNA-RDA is illustrated schematically in figure 3.23.

In most studies to date, the restriction endonuclease *Dpn* II has been the enzyme of choice for cutting the derived cDNAs, prior to linker addition. Mathematical modelling predicts that a *Dpn* II-derived representation will typically include at least one potentially amplifiable fragment from over 86 percent of expressed mammalian genes, and this can be in principle increased to 94 percent if the primer used for cDNA first strand synthesis is designed to contain a *Dpn* II site (Hubank & Schatz, 1999). Other enzymes may be substituted for *Dpn* II in subsequent experiments, which will maximise the number of differentially expressed genes potentially detectable.

cDNA-RDA is most efficient in detecting transcripts when the differences between the level of expression in two populations are great. However, the methodology can be modified to facilitate the identification of transcripts with lower levels of differential expression using a technique known as 'melt depletion'. This process effectively depletes representations of low copy sequences. Melt depletion of linker-ligated cDNA (the representation) is performed by denaturing (melting) unmixed driver samples at 98 °C, then allowing them to reanneal for one hour, a duration determined empirically to allow much more efficient reannealing for more abundant species than rarer species. Subsequent amplification of the annealed product by PCR generates a population with considerable bias against low abundance sequences, as only annealed fragments are capable of amplification (Hubank & Schatz, 1994).

3.2.4.7.3 Results obtained using cDNA-RDA

Many studies have been published using the cDNA-RDA technique, although to date this has been predominantly with eukaryotic systems. Work using cDNA-RDA to study bacterial gene expression has been more limited (Becker *et al.*, 2001; Bowler *et al.*, 1999, Westbrook-Wadman *et al.*, 1999), although a number of papers are in preparation. Bowler *et al.* (1999) used a model system comprising a strain of *Neisseria meningitidis* grown under iron-replete (to provide the driver component) and iron-limited conditions (the tester), in a study demonstrating the successful adaptation of the cDNA-RDA technique to the examination of differential gene expression in bacteria. This work was not intended as a thorough analysis of iron-regulated gene expression, but did identify both absolute and relative (up-regulated) differences in gene expression when the bacteria were grown under iron limitation. Southern hybridisation and blotting of the initial representations (tester and driver) and RT-PCR of the initial total RNA preparations were used to confirm that the differences in gene expression detected were genuine. Figure 3.24 shows a typical non-denaturing agarose gel of cDNA-RDA difference products.

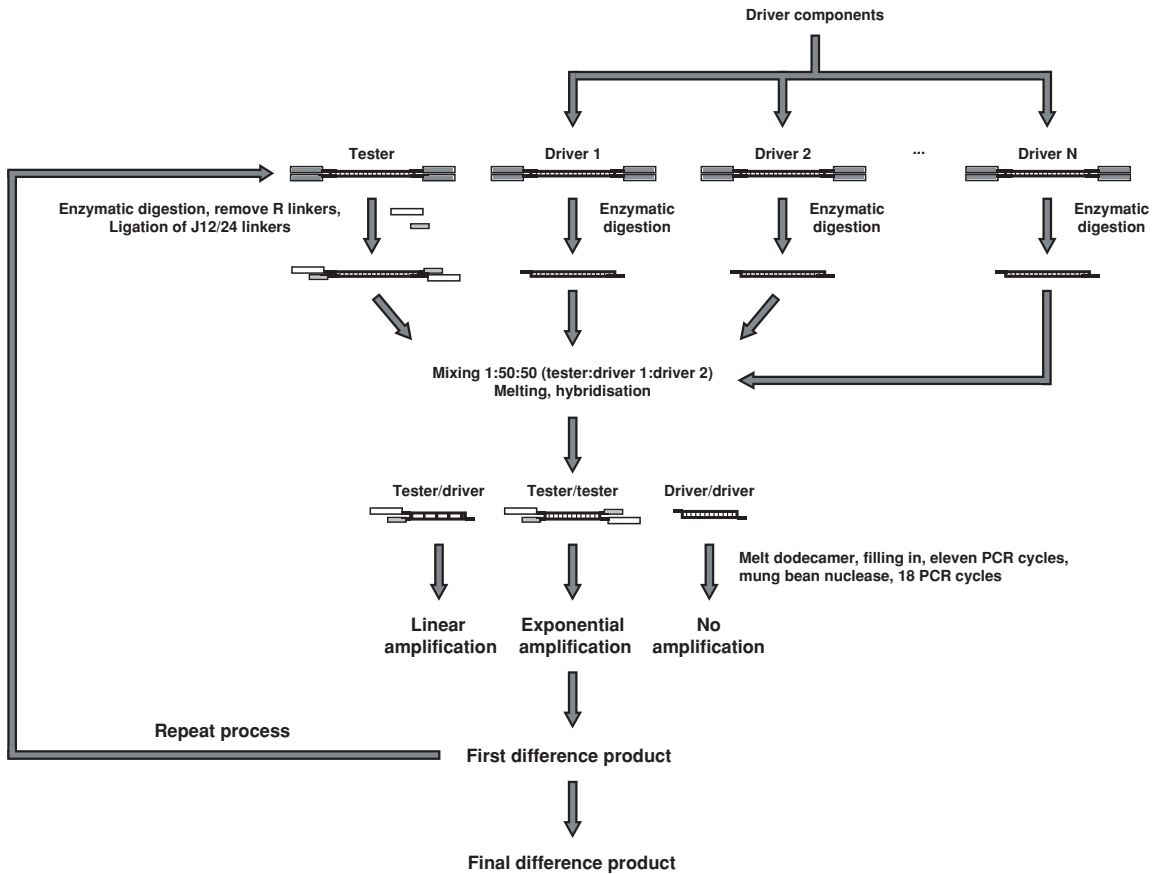
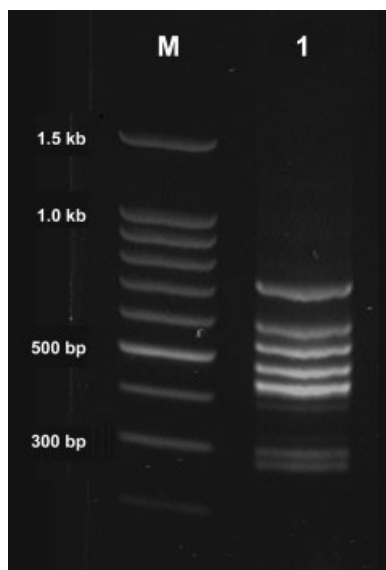


Figure 3.23 Flowchart illustrating principle of representational difference analysis of cDNA (cDNA-RDA) procedure. Where random priming of total RNA preparations has been used for cDNA synthesis, increased selection against abundant rRNA sequences can be usefully applied through the addition of extra driver components (drivers 2 to N) derived from the relevant cloned rRNA genes. Isolation of other high abundance sequences can similarly be suppressed if desired. The R12/24 adapters are digested from the tester and driver representations. A new set of linkers (J12/24) is ligated to the tester only. The tester-linker DNA is then mixed with excess driver DNA, denatured to obtain single-stranded DNA, and allowed to hybridise. Tester sequences present in both the tester and driver pools will form heteroduplexes with driver sequences, whereas unique tester sequences ('targets') can only hybridise with

their complementary tester-originating sequences. The subsequent 'fill in' will create molecules with the 24-mer oligonucleotide sequence and its complement at both the 5' and 3'-ends of the molecule for target sequences only. In the following PCR, target sequences are amplified exponentially whereas non-target sequences are either linearly 'amplified' or not amplified at all. Further enrichment is achieved by degradation of single-stranded DNA (including single-stranded overhangs of double-stranded molecules) with mung bean nuclease, and further PCR amplification. This process is usually repeated. If target sequences are not sufficiently enriched in the second difference product, the procedure can be repeated, although this increases the chance that particularly rare sequences may be lost. The figure is adapted from Bowler *et al.* (2001) with permission from Humana Press.

Figure 3.24 Analysis of difference products generated during an analysis of differential gene expression in *Streptococcus uberis* by representational difference analysis of cDNA (cDNA-RDA).

The non-denaturing agarose gel shows a typical result from a cDNA-RDA experiment designed to identify genes differentially expressed in *Streptococcus uberis* during growth in fresh cows milk versus growth in rich media (Todd Hewitt broth). Lane M: molecular mass markers, sizes as given. Lane 1: difference products obtained following two rounds of the PCR-coupled subtractive hybridisation illustrated schematically in figure 3.23. Each of the eight major bands visible potentially represent genes whose expression is up-regulated during growth in milk, and would be cloned for further characterisation. The individual bands shown are often heterogenous, *i.e.* each may comprise more than one sequence species. The kinetics of the PCR under the conditions used in cDNA-RDA preferentially enrich for products in the size range between 200 and 1,000 base pairs. The figure was kindly provided by Dr. Deborah L. Taylor (Trafford Centre for Medical Research, University of Sussex, Brighton, United Kingdom).



3.2.4.7.4 Discussion

Whilst comparatively demanding technically, cDNA-RDA has a number of significant advantages over similar differential screening techniques. In contrast to the differential display RT-PCR technique (Liang & Pardee, 1992), and the conceptually similar technique, RNA fingerprinting by arbitrarily primed PCR (Welsh *et al.*, 1992), sequences common to both groups are eliminated, greatly simplifying the interpretation of results and identification of the differentially expressed genes. Furthermore, the exponential degree of enrichment achieved by the use of PCR in cDNA-RDA enables the use of only small amounts of RNA and allows the detection of very rare transcripts.

The cDNA-RDA approach is also relatively inexpensive, particularly when compared to DNA microarray-based approaches, as no specialised equipment is necessary beyond a refrigerated microcentrifuge and a good PCR machine.

Nevertheless, as with other PCR-based approaches, there can be problems with background noise (false-positives), and general sensitivity to DNA contamination. However, these difficulties can in large part be minimised by good technique and careful screening of identified differences (results should always be confirmed by Southern hybridisation against the initial representations, and by Northern blotting or RT-PCR against individual RNA preparations).

Finally, it has recently been shown by Kim *et al.* (2001a), that the use of cDNA-RDA in combination with DNA microarrays may provide the most effective way to identify differentially expressed genes. The advantage of this hybrid approach is

likely to be even greater with prokaryotes, particularly for the examination of host-pathogen interactions or where starting material is very limited, with an amplification and enrichment stage such as provided by cDNA-RDA likely to prove essential.

3.2.4.8 RSDD – Reciprocal subtraction differential RNA display

3.2.4.8.1 Introduction

Differential gene expression is a key element in normal cellular physiology, contributing to and regulating the processes of cell growth, development, differentiation, cell death and senescence (Han & Hilsenbeck, 2001; Lee *et al.*, 1999; Leszczyniecka *et al.*, 2001; Watson & Margulies, 1993; Winkles, 1998). Moreover, abnormalities in cellular gene expression are primary factors leading to pathophysiology, including developmental abnormalities, autoimmune diseases and cancer (Antonarakis *et al.*, 2001; Leszczyniecka *et al.*, 2001; Yamanishi *et al.*, 2002; Zhang *et al.*, 1997b). Subtraction hybridisation has proven to be very efficient for identifying and cloning differentially expressed genes (Jiang & Fisher, 1993; Jiang *et al.*, 2000; Sagerström *et al.*, 1997; Su *et al.*, 1997). However, the originally described approaches require relatively large amounts of RNA (ten micrograms of poly(A)⁺ mRNA). They are also quite labour-intensive, technically demanding and do not result in isolation of the full spectrum of genes displaying altered expression (Sagerström *et al.*, 1997; Sambrook *et al.*, 1989; Wan *et al.*, 1996). PCR-based approaches require smaller quantities of RNA (about one to two micrograms of poly(A)⁺ mRNA), have accelerated the procedures for cDNA isolation and provide a new direction for subtraction, although this approach also involves several tedious steps during or after hybridisation (Diatchenko *et al.*, 1996). A very popular method for analysing and cloning differentially expressed genes employs the approach developed by Liang & Pardee (1992), RNA differential display RT-PCR (DDRT-PCR). This scheme has several advantages over standard subtraction hybridisation, including the ability to use relatively small amounts of total RNA (about five micrograms), ease of performance and the ability to identify and isolate both up and down-regulated differentially expressed genes in the same reaction. Some drawbacks to DDRT-PCR include a relatively high incidence of false-positives and redundant gene identification, poor reproducibility, a lack of functional information about the cloned cDNA and biased gene display (Averboukh *et al.*, 1996; Wan *et al.*, 1996).

Based on the positive as well as negative attributes of subtraction hybridisation and DDRT-PCR, Paul B. Fisher and colleagues focused on strategies for enhancing the identification and cloning of differentially expressed genes (Huang *et al.*, 1999; Jiang *et al.*, 2000; Kang *et al.*, 1998). An approach is outlined in this section, called reciprocal subtraction differential RNA display (RSDD), that sequentially employs subtraction hybridisation to construct subtracted cDNA libraries and DDRT-PCR to display these gene products for isolation and cloning (Kang *et al.*, 1998). The combination of subtractive hybridisation and DDRT-PCR in RSDD should augment the benefits of each strategy when used independently, including a reduction in band complexity on a display gel, resulting in fewer false-positive reactions and an ability

to employ fewer primer sets for PCR reactions to analyse a large repertoire of differentially expressed genes. RSDD can be employed to screen subtractive libraries for efficiently identifying differentially expressed genes. The positive qualities of RSDD have been applied to study the process of cancer progression, resulting in the identification of novel differentially expressed genes previously unidentified using either subtraction hybridisation or DDRT-PCR approaches individually (Kang *et al.*, 1998).

3.2.4.8.2 Principles and basics

DDRT-PCR is a powerful methodology permitting the analysis of a vast array of mRNA species (more than 20,000, if no redundancy occurs) with only a small quantity of RNA (about five micrograms of total RNA) (Liang & Pardee, 1992; Liang *et al.*, 1995). In DDRT-PCR, the unique combinations of primers in reverse transcription and subsequent PCR (for example, twelve two-base anchored oligo(dT) primers and 20 random primers) are designed not only to yield a detectable size and number of bands, but also to display nearly the complete repertoire of mRNA species. By comparison of displayed gene expression patterns, DDRT-PCR allows the simultaneous identification of both up and down-regulated genes, detection of temporal gene expression changes and the effect of multiple reagents or experimental protocols on gene expression in a single assay. However, frequent problems encountered when using DDRT-PCR include complexity of the displayed band patterns and poor separation between bands which results in false-positives requiring a significant expenditure of effort to separate and analyse the isolated expressed sequence tags (ESTs).

Subtractive hybridisation, in which hybridisation between tester and driver is followed by selective removal of common gene products, enriches for unique gene products in the tester cDNA population and reduces the abundance of common cDNAs (Sagerström *et al.*, 1997). Subtraction of driver from tester can be achieved basically in two ways, subtraction between cDNAs and subtraction between cDNA libraries, both of which have proven effective. A subtracted cDNA library can be analysed to identify and clone differentially expressed genes by randomly picking colonies or by differential screening (Jiang & Fisher, 1993; Rangnekar *et al.*, 1992). Although subtractive hybridisation has been successfully used to clone a number of differentially expressed genes (Jiang & Fisher, 1993; Su *et al.*, 1997), the approach is labour-intensive and does not result in the isolation of the complete spectrum of genes displaying altered expression (Sagerström *et al.*, 1997; Wan & Erlander, 1997; Wan *et al.*, 1996).

RSDD combines three methods to effectively identify and clone differentially expressed sequences: (i) reciprocal subtraction of cDNA libraries, (ii) DDRT-PCR and (iii) reverse Northern blot analysis (figure 3.25) (Kang *et al.*, 1998). RSDD merges the identification and separation power of DDRT-PCR (Liang *et al.*, 1995) with the selective unique gene enrichment properties of subtractive hybridisation (Jiang & Fisher, 1993; Sagerström *et al.*, 1997). In principle, since subtractive hybridisation should result in the enrichment of unique sequences and a reduction or elimination of common sequences and DDRT-PCR performed with subtracted RNA or

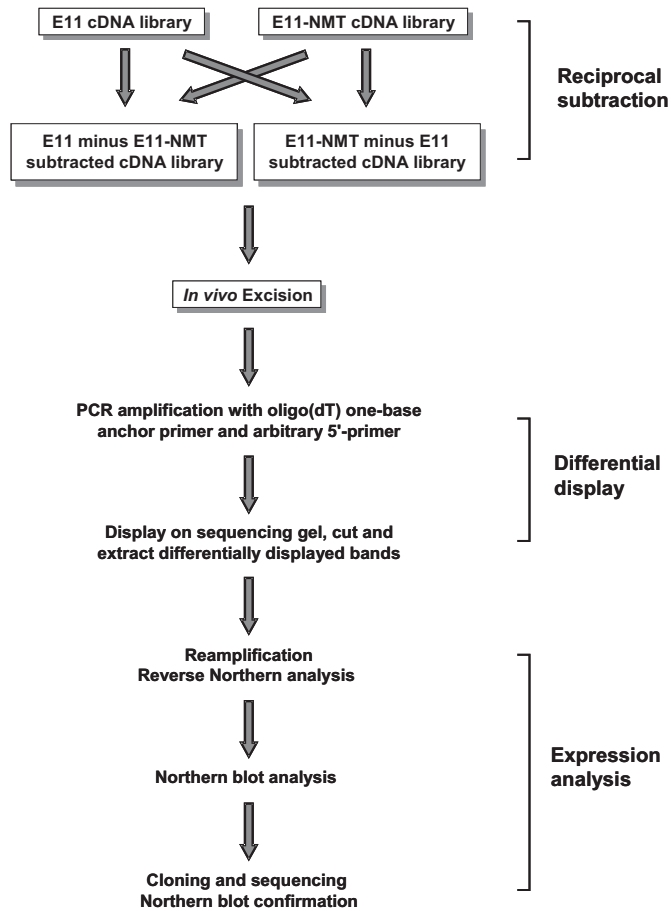


Figure 3.25 Schematic outline of the reciprocal subtraction differential RNA display (RSDD) protocol. This scheme incorporates three steps: (i) reciprocal subtraction of cDNA libraries, (ii) differential display of cDNAs excised *in vivo*, and (iii) expression analysis by reverse Northern and standard Northern blotting. For the present application of RSDD, reciprocal subtraction hybridisation was performed using libraries constructed from E11 and E11-NMT

cells (see text for details), *i.e.*, E11 minus E11-NMT and E11-NMT minus E11. Differentially expressed cDNAs identified on gels using differential RNA display were isolated, reamplified and analysed for expression by reverse Northern blotting. To confirm differential expression, cDNAs were analysed using Northern blotting. The figure is reproduced from Kang *et al.* (1998) with permission from *Proceedings of the National Academy of Sciences of the USA*.

cDNA samples should result in a consistent reduction in band complexity on a display gel, RSDD should permit a clearer separation of cDNAs resulting in fewer false-positive reactions. In order to facilitate the screening process, cDNA fragments isolated by differential display are immobilised on a nylon membrane and analysed by reverse Northern blot (Kang *et al.*, 1998; Zhang *et al.*, 1997a;

Zhao *et al.*, 1996). The differential expression of selected cDNA fragments is conclusively confirmed by Northern blot hybridisation.

In order to take full advantage of DDRT-PCR, in which reciprocal gene regulation can be detected by displaying two samples side by side, two subtracted cDNA libraries are constructed and reciprocally subtracted as illustrated in figure 3.25 (Jiang & Fisher, 1993). Construction of cDNA libraries from source RNAs to perform reciprocal subtraction results in a continuous supply of materials for DDRT-PCR. Double-stranded tester cDNAs digested with defined restriction enzymes (*Eco* RI and *Xho* I) are hybridised with biotinylated single-stranded phagemid driver DNA, and cDNAs commonly present are removed by streptavidin-binding followed by phenol/chloroform extraction (Jiang & Fisher, 1993). A subtracted cDNA library is constructed by cloning unique cDNAs in the tester into the *Eco* RI/*Xho* I site of the phage. A reverse subtracted cDNA library is also constructed in the same way by simply switching tester and driver. This library-based subtraction approach has been found to efficiently enrich unique cDNAs and has been successfully employed to clone differentially expressed genes in various systems (Jiang & Fisher, 1993; Su *et al.*, 1997).

The subtraction protocol reduces the complexity of the cDNA population by about 100-fold (10^4 total plaque forming units of subtracted library *versus* about 10^6 total plaque forming units in the primary non-amplified library) (Jiang & Fisher, 1993). In addition, cDNAs generated by the library-based subtraction approach generally retain poly(A) tails unless an internal *Xho* I site is present downstream of the *Eco* RI site. In this way, PCR display is directly applied to the libraries of plasmids that are excised *in vivo* from the reciprocally subtracted cDNA libraries.

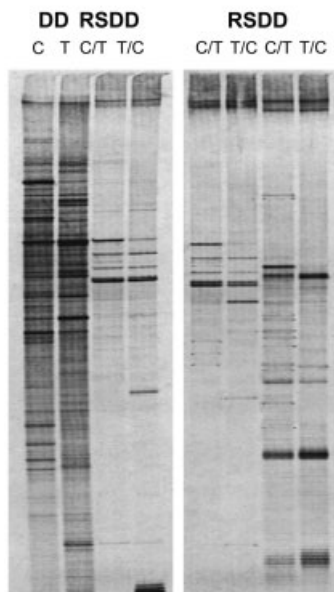


Figure 3.26 Identification of differentially expressed sequence tags using reciprocal subtraction differential RNA display (RSDD). *Left panel:* Differential RNA display pattern of conventional differential display RT-PCR with RNA derived from E11 (C) and E11-NMT (T) cells and an RSDD analysis of reciprocally subtracted E11 minus E11-NMT (C/T) and E11-NMT minus E11 (T/C) cDNA libraries. *Right panel:* Representative RSDD patterns using different sets of primers. The figure is reproduced from Kang *et al.* (1998) with permission from *Proceedings of the National Academy of Sciences of the USA*.

Considering the reduced complexity in the subtracted library, single anchor oligo(dT) primers (dT₁₃V; V is either adenine, cytosine or guanine but not thymine) and 18 random decamer primers are used for differential display RT-PCR (Kang *et al.*, 1998). Assuming no redundancy in the subtracted library, an average of about 200 bands will be displayed by RSDD with the 10⁴ plaque forming units cDNA library using these primer combinations. If the average redundancy factor is two, 100 bands will be displayed by a given set of primers. Thus, by differential display of subtracted cDNAs, a decipherable display pattern can be obtained with fewer primer sets for PCR reactions resulting in analysis of the full spectrum of differentially expressed genes.

In order to streamline the confirmatory process of differential expression, the cDNAs identified by RSDD are subjected to reverse Northern hybridisation (Huang *et al.*, 1999a; Kang *et al.*, 1998; Zhang *et al.*, 1997a; Zhao *et al.*, 1996). Bands identified in RSDD are recovered from a display gel and reamplified with the same primer set that was used for the initial amplification. The reamplified cDNA fragments are immobilised on nylon membranes and hybridised with labelled first strand cDNA, reverse transcribed from source RNA samples. Either two immobilised membranes can be hybridised in parallel or an immobilised membrane can be hybridised sequentially. Various control cDNAs, such as GAPDH and β -actin and genes with confirmed differential expression patterns should be included to monitor the hybridisation efficiency. When a single membrane is hybridised with two or more reverse transcription products sequentially, it is important to determine baseline signal after each hybridisation strip-off cycle. Pre-screening for differential expression of RSDD-identified cDNA fragments by reverse Northern blot hybridisation can greatly reduce the cost and effort necessary for confirming differential expression by Northern blot hybridisation.

3.2.4.8.3 Presentation of typical results

Proof of principle for the RSDD approach comes from its application for identifying genes differentially expressed during cancer progression (Kang *et al.*, 1998). The model used for RSDD is an adenovirus-transformed rat embryo cell line, termed E11, which acquires an aggressive oncogenic progression phenotype when re-established in cell culture after tumour formation in an athymic nude mouse (called E11-NMT) (Babiss *et al.*, 1985; Reddy *et al.*, 1993; Su *et al.*, 1997). Injection of E11 cells into nude mice results in tumours in 100 percent of animals with a tumour latency time of 35 to 40 days, whereas E11-NMT cells form tumours in 100 percent of nude mice with a tumour latency time of 15 to 20 days. Additionally, E11 cells form colonies in agar with an efficiency of about three percent, whereas E11-NMT cells display an agar cloning efficiency of more than 30 percent. The increased tumourigenicity and enhanced anchorage independence phenotypes are key indicators of tumour progression in the E11/E11-NMT model system (Babiss *et al.*, 1985; Reddy *et al.*, 1993; Su *et al.*, 1997).

RSDD has been used to analyse genes differentially expressed during transformation progression (Kang *et al.*, 1998). The differential RNA display patterns of E11 and E11-NMT cells using standard differential display RT-PCR (DDRT-PCR)

and RSDD are shown in figure 3.26 (left panel). As predicted, the differential display RNA pattern of RSDD is simplified as compared with DDRT-PCR. Most bands common to both cDNA samples are eliminated using RSDD. These experiments demonstrate that RSDD is effective in reducing display patterns and subtractive hybridisation prior to differential display RT-PCR permits the efficient identification of differentially expressed cDNAs. Since RSDD significantly reduces the number of bands displayed, single anchor oligo(dT) primers that can increase band numbers were successfully used in subsequent applications of the RSDD approach (figure 3.26, right panel). Using RSDD, 235 differentially displayed cDNAs were isolated using the E11/E11-NMT tumour progression model system (Kang *et al.*, 1998).

cDNA fragments extracted from gel were reamplified, dot-blotted on nylon membranes and sequentially probed with reverse transcribed ^{32} phosphorus-labelled cDNA derived from E11 or E11-NMT RNAs (figure 3.27). Signals were detected in 181 reamplified bands out of 235 (77 percent). The signal intensities of the various cDNAs in reverse Northern analysis were quantified with a PhosphorImagerTM (Amersham Biosciences, Inc., Piscataway, New Jersey, USA) and normalised against that of GAPDH, which remained unchanged in E11 and E11-NMT cells. The progression elevated gene-3 (*PEG-3*) (Su *et al.*, 1997) was used as an additional control, to verify increased expression in E11-NMT *versus* E11 cells. In the reverse Northern blot analyses, *PEG-3* levels were four-fold higher in E11-NMT than in E11 cells, which coincided with Northern blotting results, thereby demonstrating the concordance of reverse Northern and Northern assays. A 1.8-fold differential cut-off (after normalisation for GAPDH expression) was used to identify and isolate cDNA bands displaying modified expression in E11 *versus* E11-NMT cells. This resulted in the identification of seven cDNAs with higher expression in E11 *versus* E11-NMT cells and 65 cDNAs with elevated expression in E11-NMT *versus* E11 cells, with a positive concordance ratio of about 40 percent (72 of 181 cDNAs). Thirty-eight cDNA species out of 72 have been analysed by Northern blot hybridisation and 31 (82 percent) display differential expression in E11 *versus* E11-NMT cells. Sequence analysis of the cloned cDNA fragments revealed 16 unique genes at an average redundancy rate of 1.93 (Kang *et al.*, 1998).

The average positive concordance rate of an HO-1 human melanoma cell differentiation subtracted cDNA library constructed in an identical manner was about 17.6 percent, which assuming that 200 genes are differentially expressed out of 20,000 genes, suggests approximately a 17-fold enrichment from the initial RNA

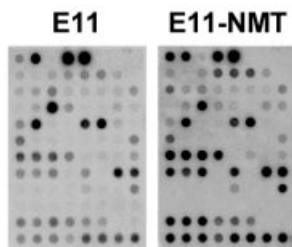


Figure 3.27 Reverse Northern blot analysis of differentially expressed sequence tags identified by reciprocal subtraction differential RNA display (RSDD). Differentially expressed sequence tags obtained from RSDD were dot-blotted onto nylon membranes and probed with ^{32} phosphorus-labelled cDNA reverse transcribed from RNA samples derived from E11 and E11-NMT cells. The figure is reproduced from Kang *et al.* (1998) with permission from *Proceedings of the National Academy of Sciences of the USA*.

population (Huang *et al.*, 1999; Jiang & Fisher, 1993). RSDD in the E11/E11-NMT model showed 33 percent positive concordance rate (about 40 percent positive in reverse Northern blot and an 82 percent concurrence between Northern and reverse Northern blot assays). In this context, RSDD increases the positive concordance rate of the starting library by two-fold, emphasising the efficiency of the protocol *versus* application of only subtraction hybridisation.

3.2.4.8.4 Discussion

DDRT-PCR is a frequently used approach for identifying differentially expressed genes in diverse biological systems and experimental contexts. Despite the conceptual elegance and simplicity of DDRT-PCR, this strategy is not problem-free. Difficulties encountered using DDRT-PCR include a high incidence of false-positives, which is partly ascribed to the complexity of the displayed band pattern resulting in poor EST separation (Averboukh *et al.*, 1996). Furthermore, poor distinction between expressed sequences can mask differentially expressed genes of low abundance, which are hidden by the intense signals generated by common genes with strong expression. RSDD decreases the complexity of the displayed band pattern by enriching for unique cDNAs and removing common ones. PCR display of reciprocally subtracted cDNA libraries significantly diminishes band complexity and permits the use of a smaller combination of primer sets. Subsequent analysis of RSDD-identified genes by reverse Northern and Northern blot hybridisation demonstrates that RSDD can be successfully applied to clone differentially expressed genes of biological relevance (Kang *et al.*, 1998).

A technical limitation of RSDD in its current format is the difficulty encountered in constructing high-quality phage-based subtracted cDNA libraries. Although a cDNA library obviates the necessity of a continuous supply of RNA, cDNA library construction requires significant amount of starting material, is labour-intensive and can be methodologically challenging. Two groups attempted to combine subtractive hybridisation and DDRT-PCR (Fuchs *et al.*, 2000a; Hakvoort *et al.*, 1994). A subtraction-enhanced display technique developed by Hakvoort *et al.* (1994) employed multiple rounds of subtractive hybridisation of cDNA fragments generated by four-base recognising restriction enzymes ligated to defined adapters. This subtraction-enhanced display technique effectively enriched for differentially expressed cDNAs, but did not significantly reduce the complexity of the display pattern. The unexpected complexity of the display profile may be ascribed to PCR artefacts due to repeated PCR of cDNAs for multiple rounds of subtraction. Recently, a subtracted RNA fingerprinting (SuRF) technique that significantly simplifies subtraction has been reported (Fuchs *et al.*, 2000a). In SuRF, differentially expressed genes are displayed by employing DDRT-PCR of unbound RNA to magnetic bead-trapped driver cDNA. Not only does SuRF decrease complexity in the displayed band pattern, but the majority of SuRF-identified cDNAs are reported to represent genuine differentially expressed genes (41 out of 44). However, since the subtraction ratio between tester and driver (1:1, 1:3 and 1:4) does not influence the display pattern, it is not clear whether the high positive concordance rate is due to the subtraction protocol or a result of the specific modified primers (double restriction site prim-

ers) used in SuRF. In addition, an average of about 1.8 cDNA species are present in each displayed band, probably resulting from the poor resolution of the sequencing gel or the specific primer set used for SuRF-PCR. However, most of the bands appear to represent differentially expressed sequences. A few of the RSDD-identified bands contain multiple cDNA species (Kang *et al.*, 1998), but this occurs at a significantly lower frequency than when using the SuRF method. Mixed cDNA species in a band prohibit applying a pre-screening method such as reverse Northern blotting and necessitate separation of cDNAs prior to analysis by resolution of reamplified bands or by cloning. Thus, all three methods have positive attributes and drawbacks.

Although SuRF-type subtraction is apparently effective and is technically simple, the subtraction ratio is limited by the amount of RNA available (Fuchs *et al.*, 2000a). Moreover, co-migrating bands also pose impediments in subsequent screening of the SuRF-identified cDNA fragments. In principle, RSDD can be improved by employing a SuRF-type subtraction to avoid the necessity of producing subtracted cDNA libraries. With limited amounts of starting material, linear amplification of driver cDNA with T7 RNA polymerase represents an alternative scheme to produce driver amounts that would permit the use of multiple subtraction ratios (Kacharina *et al.*, 1999). With the subtracted RNA, conventional DDRT-PCR followed by pre-screening with reverse Northern blotting can be performed to identify differentially expressed genes. With adequate amounts of driver cDNA, this modification can be applied to analysing gene alterations in a series of samples simultaneously. With respect to the lengthy and labour-intensive screening procedures necessitated when using cDNA libraries, RSDD can also prove amenable for efficiently screening cDNA libraries for differentially expressed genes.

Of note, the expression of most of the RSDD-identified clones is altered about 2.4-fold, which represent sequences that would not be readily detected by conventional DDRT-PCR. In addition, RSDD has identified eleven of 16 genes that were previously unknown (Kang *et al.*, 1998). Both the subtraction-enhanced display technique and subtracted RNA fingerprinting also identified genes displaying small differences in expression (Fuchs *et al.*, 2000a; Wu *et al.*, 1996). In addition, a significant portion of identified genes in both subtraction-display approaches (44 and 58 percent, respectively) are unknown, which suggests that many potentially physiologically important genes and gene changes remain undetected because of their low expression levels. Rare gene products that are masked by strong common gene products should be displayed by using subtraction hybridisation in combination with DDRT-PCR. It is also worth emphasising that the displayed bands in RSDD are different from those detected using conventional DDRT-PCR, which may result from alterations in PCR preference of selected gene populations *versus* whole RNA populations (Fuchs *et al.*, 2000a; Kang *et al.*, 1998). DDRT-PCR following subtractive hybridisation can identify genes displaying small differences in expression, rarely expressed genes, genes displaying large changes in expression and abundantly expressed genes. RSDD can also be employed to clone both rare and abundantly expressed genes. In addition, the DDRT-PCR approach with subtractive libraries could also prove valuable

for efficiently screening subtracted cDNA libraries for differentially expressed genes.

3.2.4.9 **SABRE – Selective amplification *via* biotin and restriction-mediated enrichment**

3.2.4.9.1 **Introduction**

Gene regulation at the level of mRNA abundance has been shown to be an important component of many biological systems. Identification of genes whose mRNAs are differentially expressed can provide insight into the molecular mechanisms underlying this regulation, and can reveal novel regulatory pathways. Unfortunately, no single methodology has been found to be completely satisfactory for detection of differentially expressed genes in all experimental systems. For instance, recent advances in robotics and genome sequencing have led to vast improvements in microarray technologies (Greenberg, 2001); nonetheless, these approaches still have drawbacks, including issues of cost, sensitivity, and species specificity. As an alternative, other methods are available which, rather than measuring mRNA accumulation directly, enrich specifically the abundance of cDNAs showing differential expression. This in turn facilitates subsequent identification and analysis of these species. Described in this section is one such approach, selective amplification *via* biotin and restriction-mediated enrichment, or SABRE (Lavery *et al.*, 1997 and 2000).

3.2.4.9.2 **Principles and basics**

3.2.4.9.2.1 **Competitive hybridisation**

To enrich differentially expressed species, SABRE uses an approach termed competitive hybridisation, which is also found in protocols such as representational difference analysis, or RDA (Lisitsyn & Wigler, 1993). Competitive hybridisation relies on the random re-assortment of two different double-stranded cDNA populations, each of which has been tagged so as to identify the starting population from which each DNA strand was derived. Historically, the sample in which overexpressed species are sought is referred to as the tester population (yellow in figure 3.28), while the control sample is referred to as the driver population (green in figure 3.28). After the tester and driver populations have been tagged, they are mixed together, denatured, and allowed to re-associate to form double-stranded molecules. These molecules will be distributed into three populations: (i) tester homohybrids, with both cDNA strands derived from the yellow tester population; (ii) driver/tester heterohybrids, in which one strand is derived from each of the tester and driver populations; and (iii) driver homohybrids, in which both strands are derived from the green driver population (figure 3.28). For a cDNA expressed equally in both the driver and tester samples, such as the cDNA encoding actin in figure 3.28, the expected distribution of double-stranded actin cDNAs in these three populations would be, respectively, 25 percent : 50 percent : 25 percent.

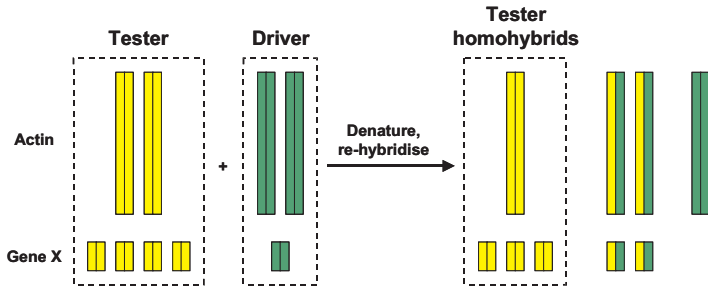


Figure 3.28 Principle of enrichment by competitive hybridisation. In two double-stranded cDNA populations, tester (yellow) and driver (green), the cDNA encoding actin is expressed equally, while the cDNA for a gene X of interest is overexpressed in the tester population relative to the driver. When the two populations are mixed, denatured, and allowed to re-hybridise,

three classes of molecules are formed: tester homohybrids (two yellow strands), driver/tester heterohybrids (one yellow, one green), and driver homohybrids (two green strands). In the tester homohybrid population, the ratio of gene X to actin cDNAs is over-represented compared with the starting tester population.

In contrast, this post-hybridisation distribution is different for a species with greater abundance in the tester population than in the driver population, represented by the gene X cDNA in figure 3.28. When a single strand of the gene X cDNA finds a hybridisation partner to form a double-stranded molecule, it will be more likely to find a DNA strand derived from the tester population than from the driver population. Therefore, at the end of the hybridisation, the distribution of double-stranded gene X cDNAs will be biased towards tester homohybrids. If the gene X cDNA is ten-fold overexpressed in the tester cDNA compared to the driver cDNA, more than 80 percent of double-stranded gene X cDNAs will be tester homohybrids. Even if it is only two-fold overexpressed, it will still have over 44 percent, rather than 25 percent, of its double-stranded molecules in tester homohybrids.

Thus, differentially expressed species will be over-represented in the tester homohybrids. If, as in figure 3.28, the tester homohybrids are purified away from contaminating driver homohybrids and driver/tester heterohybrids, the ratio of gene X cDNAs to actin cDNAs will increase from 4:2 to 3:1. While this enrichment may not be dramatic in this case, it can be influenced by two factors. First, the greater the overexpression of gene X in the tester *versus* driver populations, the greater will be its enrichment. Second, the use of an excess of driver DNA over tester DNA in the hybridisation will augment the enrichment. This is because an excess of driver DNA will drive proportionately more actin cDNA molecules than gene X cDNA molecules into driver/tester heterohybrids, at the expense of tester homohybrids. In addition, if the selection procedure is repeated, this should lead to further, near-exponential enrichment of the differentially expressed species. Thus, in two successive selection rounds with ten-fold excess driver DNA in the hybridisation, if the gene X cDNA is ten-fold overexpressed in the original tester and driver samples, this will be increased to greater than 500 to 1 in the tester homohybrids of the second selection (Lavery *et al.*, 2000).

3.2.4.9.2.2 SABRE selection

The enrichment of differentially expressed cDNAs by competitive hybridisation depends upon the efficient separation of tester homohybrid cDNA molecules from the unwanted driver/tester heterohybrids and driver homohybrids. This is key, as the tester homohybrids represent only a fraction of the double-stranded molecules in the hybridisation mix. For instance, with a ratio of driver DNA to tester DNA of ten to one, the tester homohybrids represent only 0.8 percent of all double-stranded species.

In the SABRE procedure, PCR is used to tag differentially the tester and driver cDNA samples and, thus, to permit the purification of the tester homohybrids (Lavery *et al.*, 1997). First, the two cDNA samples to be compared are digested with a restriction enzyme to generate fragments of an average size of 250 to 300 base pairs, which are more efficiently amplified by PCR than larger fragments. These are then ligated to oligonucleotide adapters containing sequences for PCR amplification primers, and amplified by PCR. The PCR primer used to generate tester material contains a 5'-covalently linked biotin moiety, and the sequence recognised by the restriction enzyme *Bam* HI (recognition site: 5'-GGATCC-3'). For PCR reactions to generate driver DNA, a similar oligonucleotide primer is used, except that it contains no biotin group, and two bases within the *Bam* HI site have been mutated to 5'-CCATCC-3', which is no longer a functional site for this enzyme.

After amplification of the tester and driver DNA samples, the two are mixed, denatured by heating, and allowed to re-hybridise. Tester homohybrids are then isolated by first capturing all biotin-containing molecules by incubation with streptavidin-coated magnetic beads, and concentration of the beads using a magnet (figure 3.29). Given the high affinity of streptavidin for biotin (dissociation constant k_D of about 1×10^{-15}), the bead/DNA complex can be stringently washed to minimise contamination from driver homohybrids.

To separate the tester homohybrids from the driver/tester heterohybrids, the captured DNA molecules are digested on the magnetic beads with the restriction enzyme *Bam* HI. As the driver/tester heterohybrids contain one strand of the tester DNA with the *Bam* HI site and one of the driver DNA with the mutated site, their ends do not contain a functional *Bam* HI site (figure 3.29). Therefore, only tester homohybrid molecules are digested by the enzyme and released from the magnetic beads. They can then be collected free of the bead-bound driver/tester heterohybrids by immobilising the beads with a magnet.

The supernatant containing the tester homohybrids is next used in a PCR reaction to amplify the selected species. This generates microgram quantities of selected material, which can be used immediately to start a further round of SABRE selection, or can be immediately analysed to identify species enriched during the selection procedure. Enriched fragments can be directly isolated by cutting out of a sequencing gel, followed by PCR amplification and direct cloning into plasmid vectors (Lavery *et al.*, 1997). Conversely, the SABRE selection products can be cloned directly, without gel selection, into viral or plasmid vectors and screened by differential library screening on duplicate filter membranes in order to identify

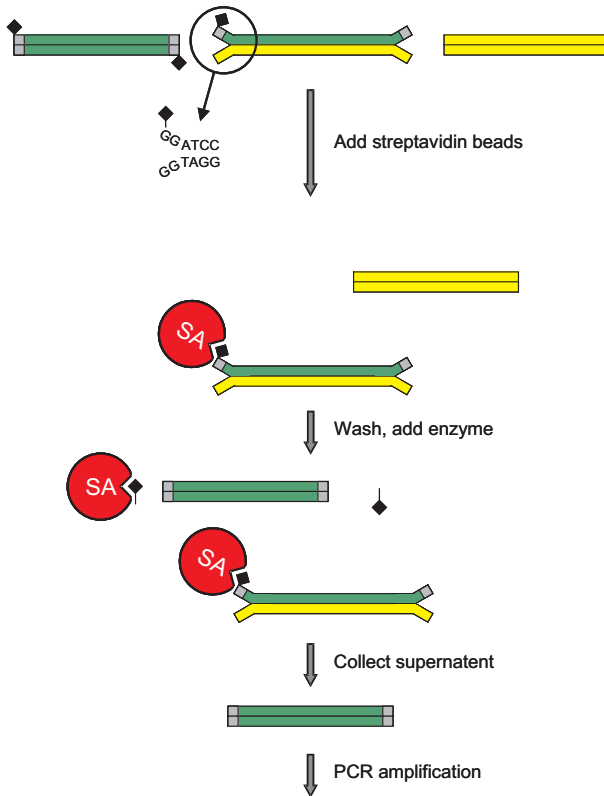


Figure 3.29 Selective amplification via biotin and restriction-mediated enrichment (SABRE) procedure. Purification of tester homohybrids away from other hybrids is performed by first incubating the hybridisation mix with streptavidin-coated magnetic beads (red circle), which captures molecules containing a biotin group (black diamond). Driver homohybrids are removed by stringent washing of the bead/DNA complexes, using a magnet to concentrate the magnetic beads. Tester homohybrids are recov-

ered specifically by incubating the bead/DNA complexes with the restriction enzyme recognising the duplex DNA sequence in grey; this sequence is not found in the driver/tester heterohybrids (green/yellow), due to mismatches in the sequence on the driver-derived strand (see enlargement at right, top panel). Cleavage at this site releases the tester homohybrid into the supernatant, which is recovered and amplified by PCR for subsequent analysis or another round of SABRE selection.

species which might not be apparent when displayed on a sequencing gel. The cloned cDNA fragments are used as probes in Southern hybridisations of the starting libraries, and by detection of their RNAs in the RNA samples (or duplicate samples) from which the cDNAs were synthesised (Lavery *et al.*, 2000). This is to confirm that the enrichment of the isolated cDNA fragment was not an artefact of the selection process, and did not represent a contaminating band from the subcloning of the species enriched by SABRE. However, when the proper control hybridisations are performed, the level of false-positives arising from the SABRE se-

lection itself is quite low, usually no more than ten percent of all clones identified (Lavery *et al.*, 2000).

3.2.4.9.3 Presentation of typical results

To illustrate one application of SABRE selection, differential gene expression studies between areas 17 and 18 of the macaque developing visual cortex are presented (Kennedy & Dehay, 1993). SABRE cDNA libraries were generated from RNA isolated from the two cortical areas, 17 and 18. Area 17 cDNA was used as tester DNA, and area 18 cDNA was used as driver DNA. After four rounds of SABRE selection, several cDNA fragments were enriched in abundance in the tester population. These fragments were isolated, cloned, and sequenced. Results for one such clone, C4-7, are shown in figure 3.30. Clone C4-7 demonstrates greater expression in the area 17 library than the area 18 library; this difference is enriched in the selection products after each SABRE selection round, until by round three it represented a major species in the selected population. Further experiments are in progress to confirm the differential expression of this cDNA fragment in areas of the macaque visual cortex by *in situ* hybridisation analysis.

3.2.4.9.4 Examples

SABRE selection has been used to identify differentially expressed genes within the mouse liver showing circadian expression patterns (Lavery *et al.*, 1997 and 1999). It has also been used to identify genes dysregulated in cells derived from knockout mice lacking expression of the metal-responsive transcription factor, MTF-1 (Lichten *et al.*, 2001). SABRE selection has also been used to identify novel genes regulated by starvation conditions (Kersten *et al.*, 2000).

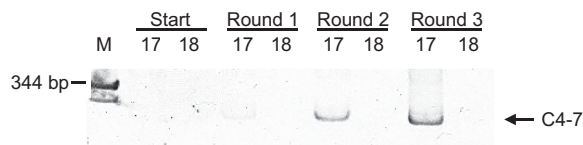


Figure 3.30 Enrichment of a cDNA restriction fragment overexpressed in developing monkey cortex area 17 by selective amplification via biotin and restriction-mediated enrichment (SABRE). SABRE cDNA libraries were generated from areas 17 and 18 of the developing macaque visual cortex, and enriched for species with greater expression in area 17 compared with 18, in three rounds of SABRE selection. One such species, C4-7, was isolated and cloned, and used as a probe in Southern hybridisation experiments to determine its accumulation in

starting libraries and after each round of SABRE selection. A C4-7 probe was hybridised to a nylon filter containing equal amounts of PCR products of the SABRE cDNA area 17 or area 18 libraries before selection ('Start') or after selection rounds as indicated. The 295 base pairs fragment corresponding to the C4-7 species is overexpressed but barely detectable in the starting area 17 library; after SABRE round three, it represents one of the major species of the selected population. The following abbreviation is used in this figure: bp, base pair(s).

3.2.4.9.5 Discussion

The SABRE protocol has many attributes that make it attractive as a tool in gene expression profiling. First, it can be performed relatively inexpensively using materials commonly found in a molecular biology laboratory, with the exception of some materials such as a magnetic selector and streptavidin-coated magnetic beads. The selection procedure itself is relatively rapid, with one selection round requiring only about two days. As SABRE uses PCR amplification to generate the tester and driver samples, small amounts of starting material can be used as the source for the cDNA. Also, SABRE is capable of enriching initial tester-to-driver ratios as small as two-fold or less. Finally, this analysis can be performed with no prior knowledge of the sequences to be identified. This is an advantage for studies of organisms for which little sequence data is known. This can also be an advantage even when analysing better characterised genomes, where limited sequence data for some cDNA species may not be completely satisfactory for designing specific microarray probes.

Disadvantages of the SABRE approach include that the approach may not be as global in the genes it enriches as might be wished. One drawback is that enough restriction sites for the enzyme used to digest the cDNA must be present in a cDNA molecule in order to generate at least one restriction fragment of a size amplified by PCR. If not, that species will be lost during the selection, either during library creation or PCR amplification.

A further drawback is that, unlike DNA microarray analysis, species enriched by SABRE analysis must be cloned and sequenced in order to identify them. With a DNA microarray, the identity of each species is known from the beginning; with SABRE, the correct DNA fragment must be amplified, purified, cloned, and sequenced, making the downstream analysis more laborious than the selection itself. Automation of cDNA fragment analysis might aid this step, but it will always remain a bottleneck in SABRE analysis.

Furthermore, the amplification of cDNA fragments by PCR is itself a double-edged sword. While this aids analysis by permitting the use of sub-microgram amounts of starting cDNA, the PCR reaction itself will inevitably introduce biases in the population through several rounds of PCR based SABRE selection. In a complex PCR reaction where thousands of different templates are competing for amplification by the same primers, enzymes, and other components, it is impossible to eliminate this bias.

Finally, while the SABRE procedure employs solution hybridisation conditions designed to accelerate the process, it is possible that rare species will not be hybridised to a significant degree under these conditions. As SABRE selection will work only for molecules in duplex DNA, these species would be lost during the procedure. This might be avoided by using as starting material cDNA samples that have already been enriched for rarer species, such as in suppression subtraction.

3.2.4.10 Subtractive hybridisation of transcriptionally active genomic DNA fragments

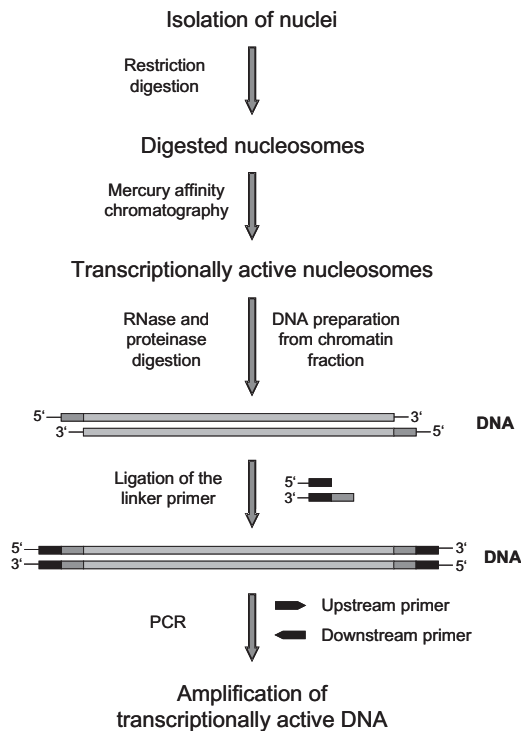
3.2.4.10.1 Introduction

Subtractive RNA hybridisation (Lee *et al.*, 1991) can be quite useful for the examination of the mRNA content of different cell types. However, in practice, the use of subtractive techniques to search for low abundance cDNA is often quite difficult because the expression levels of different mRNAs within a population of cells differ vastly, whereas the expression level of the mRNA of interest may be very similar within the two compared populations of cells. Using genomic DNA can reduce these difficulties because the copy numbers of genomic DNA are ordinarily similar in contrast to the copy numbers of cDNA. But as the use of the complete genome is very complicated, a method was developed that preferentially selects transcriptionally active genes (Cui *et al.*, 1999).

3.2.4.10.2 Principles of the technique

The method for the isolation of differentially expressed genes by subtractive hybridisation of transcriptionally active genomic DNA fragments is a combination of two formerly established techniques: the purification of transcriptionally active genomic DNA (Allegra *et al.*, 1987) and the subtractive hybridisation technique developed by Nikolai Lisitsyn and co-workers (Lisitsyn *et al.*, 1993). Figures 3.31 and

Figure 3.31 Steps of the isolation of transcriptionally active DNA fragments. The chromatin of isolated nuclei is digested with a restriction enzyme (step 1). The transcriptionally active nucleosomes are purified using mercury affinity chromatography (step 2). DNA fragments are prepared by ribonuclease and proteinase digestion (step 3). Ligation of linker primers (step 4) and PCR amplification using the linker region as upstream and downstream primers (step 5) enables the use of the linker capture subtraction technique (figure 3.32). Isolation and amplification are performed for two different cell pools, treated cells and control cells.



3.32 illustrate the principle of the subtractive hybridisation of transcriptionally active genomic DNA fragments procedure.

3.2.4.10.2.1 Purification of transcriptionally active genomic DNA

From two different cell pools, treated cells and control cells, nuclei were isolated with modified standard protocols (Antalis & Godbolt, 1991; Chen-Cleland *et al.*, 1993; Cui *et al.*, 1999). The chromatin is digested with a restriction enzyme (*e.g.*, *Mbo* I). The transcriptionally active digested genomic DNA of both cell populations is purified using mercury affinity chromatography (Allegra *et al.*, 1987; Chen & Allfrey, 1987). This purification technique is based on the principle that the thiol groups of histone H3, normally buried deep within the nucleosome core, are exposed to the solvent in transcriptionally active chromatin (Johnson *et al.*, 1987; Prior *et al.*, 1983) and therefore, are easily accessible for binding to the organomercurial agarose column (Walker *et al.*, 1990). DNA fragments are prepared from column-binding chromatin fractions by ribonuclease (RNase) and proteinase digestion. Isolation of transcriptionally active genomic DNA fragments has been successful for eukaryotic cell lines (Allegra *et al.*, 1987; Chen & Allfrey, 1987; Chen *et al.*, 1990), in particular yeast (Walker *et al.*, 1990) and human cancer cell lines (Boffa *et al.*, 1990; Sterner *et al.*, 1987).

3.2.4.10.2.2 Subtractive hybridisation

Using the purified DNA fragments isolated from the two populations of cells, a subtractive hybridisation approach is employed to isolate and clone those genes that are differentially expressed (figure 3.32). The procedure includes repeated steps of hybridisation of an excess of DNA from one source lacking the sequences of interest (called the 'driver') with DNA from the source containing the desired sequences (called the 'tester'). Subtractive hybridisation is performed using the representative difference analysis method (RDA) that is described in a previous section in this chapter (Lisitsyn *et al.*, 1993; Lisitsyn, 1995). RDA combines subtractive hybridisation with kinetic enrichment to purify DNA restriction fragments present in one population but not in another. RDA is used here as an improved variation called the linker capture subtraction technique (LCS) that is also described in a previous section in this chapter (Yang & Sytkowski, 1996).

Tester DNA and driver DNA are ligated to linker primers compatible with the induced restriction sites (figure 3.31). After PCR amplification of the linked DNA, the PCR products of the driver are digested by the nuclease (*e.g.*, *Mbo* I) to cleave the linker region (figure 3.32). For each hybridisation/amplification cycle, the tester DNA and an excess of the driver DNA are mixed, heat-denatured and subsequently hybridised (figure 3.32). Single-stranded DNA including sticky ends is degraded by mung bean nuclease. Enrichment of the target sequences is performed by PCR using the linker primer regions. As tester/driver hybrids and re-annealed driver fragments do not have the linker region on both ends, these hybridisation products cannot be amplified. The hybridisation/amplification cycle is repeated up to three times. Enrichment of subtraction products is visualised by polyacrylamide gel electrophoresis (figure 3.33). Enriched DNA fragments are

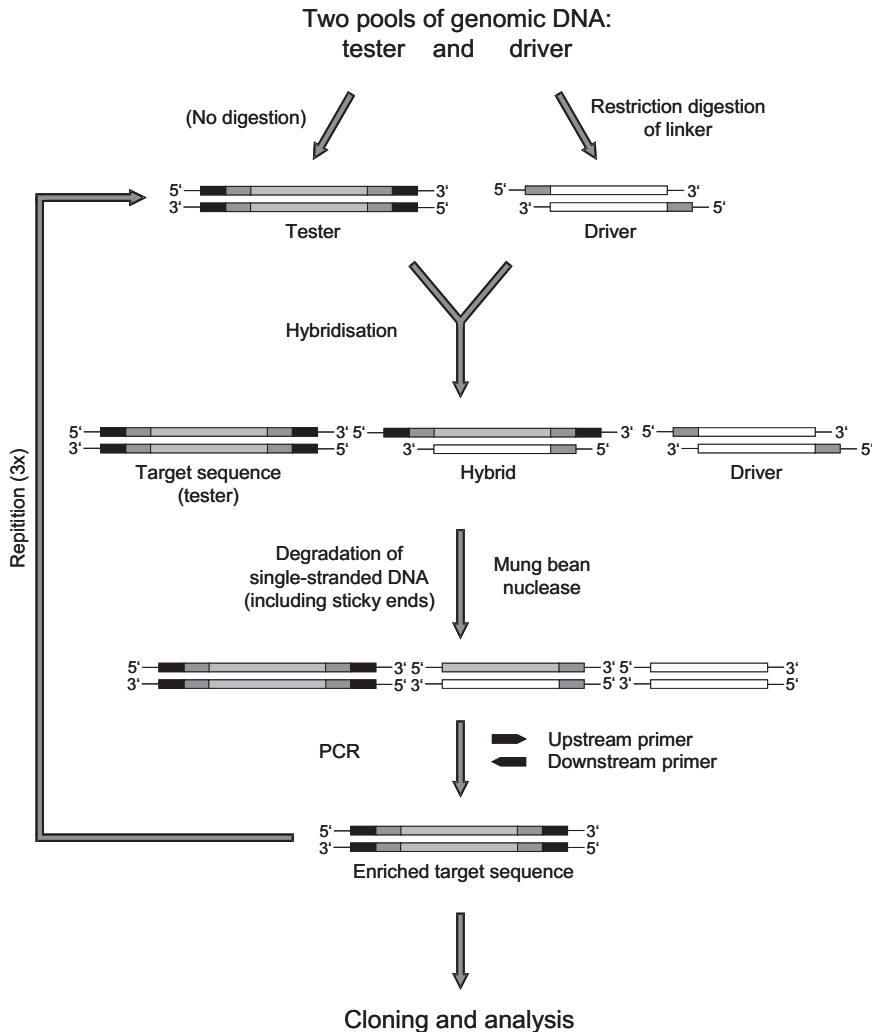


Figure 3.32 Steps of subtractive hybridisation using RDA and LCS. After PCR amplification of the linked DNA (compare figure 3.31) the PCR products of the driver are again digested by the restriction nuclease to cleave the linker region from the PCR products (step 1, right arrow). For each hybridisation/amplification cycle, the tester DNA and an excess of the driver DNA are hybridised (step 2). Single-stranded DNA including sticky ends are degraded by mung bean nuclease (step 3). Enrichment of the target sequences is performed by PCR using the linker primer regions (step 4). The hybridisation/amplification cycle is repeated up to three times. Enriched DNA fragments are cloned and analysed (step 5, compare figure 3.33).

cloned into a vector to establish a small subtraction library. Cloned fragments are used for Northern blot analysis with total RNA probes to verify the subtraction products (figure 3.33) (Cui *et al.*, 1999). The use of different (or more than one) restriction enzymes permits the discovery of additional genes.

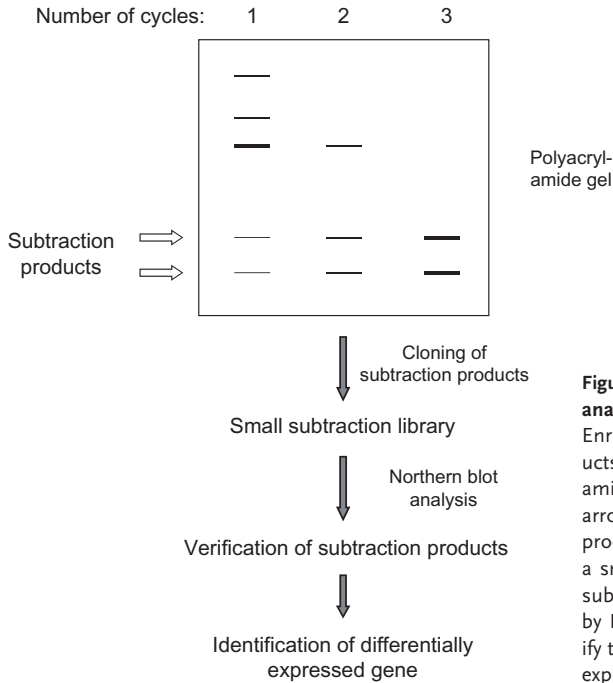


Figure 3.33 Visualisation and analysis of subtraction products. Enrichment of subtraction products is visualised by polyacrylamide gel electrophoresis (white arrows). Detected subtraction products are cloned to establish a small subtraction library. The subtraction products are tested by Northern blot analysis to verify the detection of a differentially expressed gene.

3.2.4.10.3 Examples and typical results

The method was used to study the mechanism of growth factors, in detail to isolate genes that are up or down-regulated by certain growth factors. As an example, Kunyuan Cui and co-workers isolated genes up-regulated by the hormone erythropoietin in erythroid cells (Cui *et al.*, 1999). Erythropoietin regulates the proliferation and differentiation of erythroid cells of this cell line. Therefore, Rauscher murine erythroleukemia cells induced with erythropoietin were used as the tester and control cells not treated with erythropoietin of the same lineage as the driver. Subtraction products were visualised by agarose gel electrophoresis (figure 3.34). These products were cloned, tested and sequenced.

Subtraction cycle

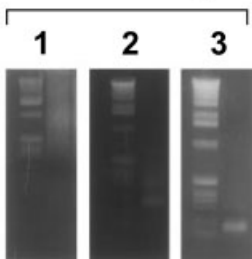


Figure 3.34 Agarose gel analysis of subtraction products induced by erythropoietin in Rauscher murine erythroleukemia cells (Cui *et al.*, 1999). Enrichment from cycle one to three of subtraction products is clearly visible. Cycle numbers are given above. Lanes are for each panel: marker left; subtraction products right. Reproduced with kind permission from *Methods* (Cui *et al.*, 1999).

3.2.4.10.4 Discussion

3.2.4.10.4.1 Advantages of the procedure

The method has some advantages compared to techniques that make use of either mRNA or total genomic DNA subtractive hybridisation. A principle advantage is that interference by background levels of cellular mRNA is avoided. Often, growth factor-induced changes in mRNA are too small to be identified using cDNA subtraction in the presence of the total population of mRNA. Vast differences in mRNA abundance can adversely affect the subtractive hybridisation efficiency. Genomic DNA is more easily obtained and manipulated and the copy numbers of purified genomic DNA fragments are much more uniform than in the mRNA population. By selecting only transcriptionally active chromatin as a starting material, changes in gene expression may be identified more completely, whereas the use of the complete genome as a starting material increases the complexity of the subtraction step. However, subtractive hybridisation using the complete genome has been successful in the case of large deletions on the mutated chromosome (Kunkel *et al.*, 1985; Nussbaum *et al.*, 1987). Furthermore, RDA was also developed to detect genetic lesions in cancer using the complete genome although described to be a complex procedure (Lisitsyn & Wigler, 1995; Lisitsyn, 1995).

3.2.4.10.4.2 Limitations

However, some limitations and problems may occur using this method. Attention should be paid to the possible sources of artefacts, especially in the subtraction steps. Contamination and PCR bias can occur. Some PCR products present after subtraction and enrichment of the target may result from the stochastic nature of the process itself. Each candidate must be tested using, for example, Northern blot analyses. Furthermore, DNA fragments of lower molecular weight are preferred to be enriched due to PCR preferences. As genomic DNA contains both exon and intron regions, dot blot analysis of spotted DNA fragments using cDNA as a probe can also be useful to distinguish clones including exon regions.

3.2.4.11 SuRF – Subtracted RNA fingerprinting

3.2.4.11.1 Introduction

Subtractive hybridisation, one of the earliest methods developed to identify differentially expressed mRNAs between two samples (Sagerström *et al.*, 1997), has been successfully applied in the identification of several differentially expressed genes. In this process, common messages are selectively removed from two RNA populations and unique ones are enriched. However, the original method of subtractive hybridisation proved to be highly labour-intensive and ineffective in that only a small fraction of the total complement of differentially expressed genes could actually be identified. Advances such as differential display RT-PCR (DDRT-PCR) (Liang & Pardee, 1992) and RNA arbitrarily primed PCR fingerprinting (Ralph *et*

al., 1993) represented significant improvements over the original method as the expression of a large number of mRNAs could be analysed from a small number of samples, especially in the case of DDRT-PCR. Furthermore, these methods were able to assure global screening of expressed genes between samples without requiring any prior information on the types of genes expressed by a given tissue or cell type. However, there remained inherent complexities and a number of limitations, including lack of reproducibility, false-positives, redundant identification of genes and masking of signals of low abundance mRNAs by high abundance ones. Thus, while DDRT-PCR methods theoretically can provide a near 100 percent screening efficiency, the extent to which such methods can identify poorly expressed genes is unknown. A recently developed method termed reciprocal subtraction differential display (RSDD), which combines the resolution of subtractive hybridisation with the power of DDRT-PCR (Kang *et al.*, 1998) has been shown to be very effective in monitoring differentially expressed genes. However, it has many of the same limitations listed above and is labour-intensive. For example, it is necessary to generate two λ -ZAP™ cDNA libraries representing ‘tester’ and ‘driver’ mRNAs and an additional two subtracted libraries. It requires isolation of poly(A)⁺ mRNA as the starting material, double strand cDNA synthesis, digestion with restriction enzymes, ligation of linkers and the acquisition of subtracted plasmids.

In this section, a mutually subtracted RNA fingerprinting (SuRF) method is described that utilises the intrinsic powers of both DDRT-PCR and subtractive hybridisation. SuRF does not have any of the above limitations of RSDD. Importantly, the method eliminates the need to generate multiple cDNA libraries since SuRF is applied directly on RNA. SuRF has been used to isolate differentially expressed genes between cultured normal osteoblast and osteosarcoma cells in order to identify osteosarcoma-associated marker genes. The method identified a large number of known and unknown genes (candidate markers) for the disease. Investigations are currently underway to identify potential breast carcinoma-associated markers using SuRF coupled with PACS (preferential amplification of coding sequences), a technique which is discussed in a separate section in this book.

3.2.4.11.2 Principle and basics

3.2.4.11.2.1 Isolation of RNA from cultured cells

Total RNA from FOB, an immortalised normal osteoblast cell line, was used as a reference (Harris *et al.*, 1995), and MG63, an osteogenic sarcoma-derived cell line (Heremans *et al.*, 1978), was isolated by phenol/chloroform extraction method (Chomczynski & Sacchi, 1987). RNA samples are treated with DNase to ensure destruction of contaminating DNA (Sanyal *et al.*, 1997). Then, RNA is quantified in a ultraviolet spectrophotometer and the integrity is evaluated by electrophoresis in an agarose gel prior to cDNA synthesis.

3.2.4.11.2.2 Mutually subtracted RNA fingerprinting procedure

Figure 3.35 shows a scheme illustrating the principle of the SuRF method. In the first step of the SuRF procedure, poly(A)⁺ mRNA is isolated from total RNA using oligo(dT) primers attached to paramagnetic beads. Complementary DNA is synthesised from the mRNA/oligo(dT)-bead complex by Moloney murine leukemia virus (MMLV) reverse transcriptase. To degrade RNA from the RNA/DNA hybrid, sodium hydroxide is added to the above reaction. After incubation of the reaction

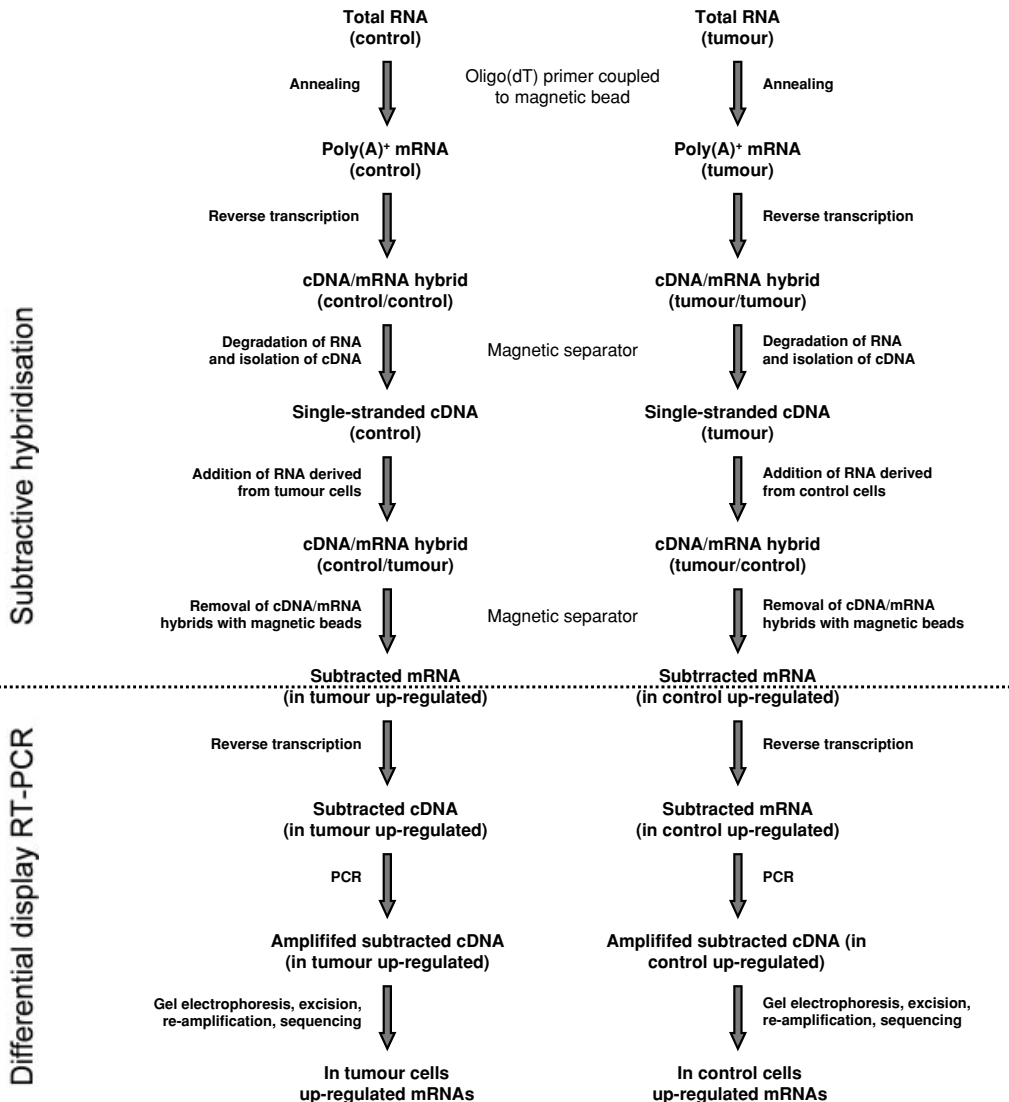


Figure 3.35 Flowchart of the subtracted RNA fingerprinting (SuRF) method. The principle of this method is described in detail in the text.

mixture with Tris buffer and hydrochloric acid at room temperature, the cDNA is then recovered as a pellet by use of a magnetic separator. Following isolation of single-stranded cDNA, an excess of total RNA from a source from which common RNAs needs to be removed is added. This RNA/cDNA mix is denatured and incubation at room temperature is performed to allow hybridisation of RNA and cDNA strands. A magnetic separator is then used to collect the supernatant, which contains the subtracted RNA, which is subsequently precipitated for further processing. The RNA is then converted to cDNA with standard oligo(dT) primers (without magnetic beads) or random hexamer primers by using a standard protocol described by Sarkar *et al.* (1990).

For mRNA fingerprinting, a PCR is carried out with different combinations of an oligo(dT)₂₃V (V is adenine, cytosine, or guanine, but not thymine) primer as reverse primer and the double restriction site primer (DRSP) as forward primer (for details, see table 3.5). After PCR, a stop buffer (containing formamide, EDTA, xylene cyanol and bromophenol blue) is added after which the PCR product is denatured. A sample of the denatured material is electrophoresed on a six percent polyacrylamide gel containing urea. The gel is dried after electrophoresis and subjected to auto-radiography.

3.2.4.11.2.3 Extraction, cloning and sequencing of gel-embedded DNA

DNA of differentially expressed bands is extracted after excision from the gel. Then, DNA is reamplified by PCR with the same combination of primers used at the fingerprinting step and cloned into a vector. Inserts in clones are amplified by PCR using primers from the vector and sequenced in an automated sequencer. Messenger RNAs are identified using the FASTA or BLAST search algorithms (Altschul *et al.*, 1990; Pearson & Lipman, 1988).

3.2.4.11.3 Typical results, examples and discussion

SuRF was developed in order to efficiently identify osteogenic sarcoma-associated molecular markers (and has since been applied in the identification of breast carcinoma-associated genes among others). The expression pattern of most genes is

Table 3.5 List of primers used for subtracted RNA fingerprinting (SuRF). N is either adenine, cytosine, guanine, or thymine. The following abbreviation is used in this table: DRSP, double restriction site primer.

<i>Number</i>	<i>Oligo(dT) primer</i>
1	5'-T ₂₃ A-3'
2	5'-T ₂₃ G-3'
3	5'-T ₂₃ C-3'
<i>Number</i>	<i>DRSP primer (containing restriction site sequence at both ends)</i>
1	5'-GAATTC NNN TCGA-3'
2	5'-GAATTC NNN GCGC-3'
3	5'-GAATTC NNN GGCC-3'
4	5'-GAATTC NNN CCGG-3'

believed to remain unaltered between normal and corresponding tumour cells/tissues. These 'common' genes have no implication as tumour markers. Further, an excess of such genes can greatly compromise identification of differentially expressed genes in two important ways. First, redundant identification of genes could lead to a high frequency of false-positives and an excess of common genes decreases the chance of identifying poorly-expressed genes. In the conventional differential display RT-PCR method (Liang & Pardee, 1992), there is no provision for eliminating common mRNAs between samples. It is reasoned that removal of the common mRNAs between two samples should greatly improve the outcome of an mRNA fingerprinting method. The oligo(dT) primer coupled with magnetic beads serves as the key reagent for mutual subtraction of common mRNAs. The subtracted mRNAs are then subjected to PCR fingerprinting with restriction site oligonucleotides (Sarkar *et al.*, 1993; Weber *et al.*, 1998) as 'random' primers and an oligo(dT) primer (without magnetic bead attached to it as anchor primer). Use of magnetic oligo(dT) primer before PCR fingerprinting (differential display) serves two other important purposes: first, it allows cDNA synthesis of only mRNAs and it allows sequestration and subsequent isolation of cDNA from a mixture of RNA and cDNA. A flowchart illustrating the SuRF principle is presented in figure 3.35.

In order to identify candidate markers for osteogenic sarcoma, SuRF was applied between MG63 (an osteogenic sarcoma-derived) and FOB (human fetal osteoblast) cell lines. A representative differential display pattern between SuRF and standard differential display RT-PCR is presented in figure 3.36. The left panel represents standard differential display RT-PCR whereas the right panel represents SuRF. In this experiment, an oligo(dT)₂₃C and 5'-Eco R1 NNN GGCC-3' (double restric-

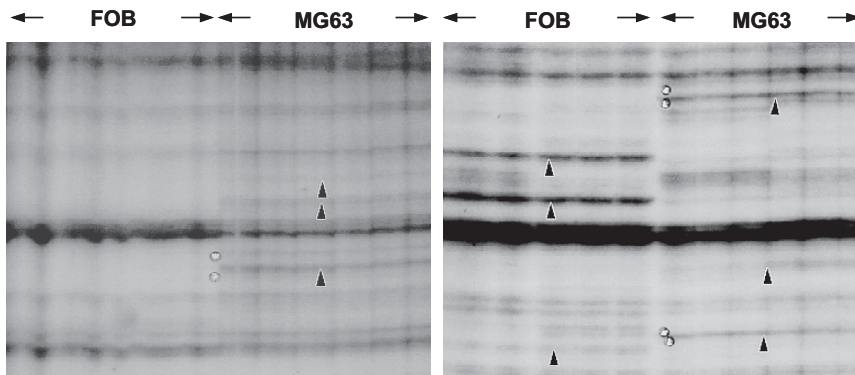


Figure 3.36 Identification of differentially expressed mRNAs by subtracted RNA fingerprinting (SuRF). *Left panel:* Conventional differential display RT-PCR pattern between FOB and MG63 RNAs (see text for details about the cell lines). *Right panel:* SuRF display of the two RNA samples. Both conventional differential display RT-PCR and SuRF were carried out with the same set of primers. Upward arrows indicate

potentially differentially expressed sequence tags. For both methods, PCR was carried out in duplicate and a sample from each tube was loaded on the sequencing gel in triplicate. After electrophoresis, the gel was dried and processed for auto-radiography. In this case, auto-radiographic exposure was for 24 hours at -80°C .

tion site primer, DRSP) primers were used for PCR in duplicate, and amplified samples from each tube were then loaded in triplicate on the sequencing gel. Note that although a large number of amplified DNA bands are produced, the number of differentially expressed bands obtained by standard differential display RT-PCR appears to be few while that of SuRF are considerably less complex than that of the standard differential display RT-PCR. Furthermore, potential differentially expressed bands produced by SuRF are more abundant and are easier to identify. Thus, results from this experiment suggest that removal of common mRNAs prior to differential display considerably reduces background and aids in the generation of a simpler display pattern. Some differentially expressed bands present in the profile obtained by standard differential display RT-PCR appear to be absent from the profile obtained by SuRF. While this could be interpreted as evidence that SuRF fails to identify some *bona fide* differentially expressed mRNAs detectable by standard differential display RT-PCR, it was not attempted to establish if they represent true differentially expressed mRNAs. Since it is known that standard differential display RT-PCR produces a high frequency of false-positives, it is difficult to conclude if those bands represent truly differentially expressed mRNAs or are artefacts. It is entirely possible that a differentially expressed gene tag identifiable by one method can be missed by another, since no single method can identify all differentially expressed genes between two samples. To conclusively answer this question, a detailed study could be undertaken to compare the efficiency of various methods directed to identify differentially expressed mRNAs.

In the experiment presented in figure 3.36, cDNAs from 15 micrograms of total RNA from FOB cells (competitor) were used to subtract from five micrograms total RNA from MG63 cells (tester) to obtain subtracted MG63 RNA. Subtracted FOB RNA was generated using the reverse approach. A competitor to tester ratio of 3:1 is routinely used for SuRF, however, competitor to tester ratios of 4:1 or 1:1 have been found to produce very similar results as obtained with the 3:1 ratio (data not shown). It is possible that the overall SuRF pattern may be different at other ratios of tester to competitor RNAs. Different competitor to tester ratios may also be required to obtain optimal SuRF pattern when using RNAs from other cell lines or tissues. However, as a rule of thumb, more competitor than tester RNA is used for subtracting common mRNAs. It is assumed that more molecules of tester cDNA should facilitate more effective removal of the common mRNAs at the subtraction step. Use of poly(A)⁺ mRNA instead of total RNA might further improve the outcome of SuRF (Fuchs *et al.*, 2000). In addition to mutual subtraction, SuRF can also be employed following subtraction of a common RNA sample (competitor) from multiple tester RNAs with the objective of identifying 'sample-specific' genes. For example, RNA from fibroblasts (or from any other desired source) may be subtracted as a common competitor to identify liver, spleen or kidney-specific genes.

Presently, more than 500 differentially expressed coding sequence tags (dCSTs) were identified for osteogenic sarcoma using SuRF and PACS (described in a separate section in this chapter) methods. Differential expression of nearly 100 of these genes/dCSTs have been confirmed by performing RT-PCR with specific primers.

A database has been created to store these dCSTs, which were defined as candidate molecular markers for osteogenic sarcoma. The database does not contain other apparently unimportant sequences such as repeat sequences and rRNA sequences. Nearly 50 percent of the sequences represent known genes while many sequences do not have any match with sequences in the GenBank (Sarkar *et al.*, 1993; Sarkar *et al.*, manuscript in preparation). All the sequences will be presented in a publicly-accessible website that is currently under construction.

3.2.4.11.4 Concluding remarks

No single method for the identification of differentially expressed genes currently available is 100 percent comprehensive. Many genes expressed in the human genome have not been identified despite intense efforts, presumably due to their low copy number. Since human mRNAs contain an average of 200 adenosine residues at their 3'-end (Lewin, 1997), standard oligo(dT) priming in reverse transcription results in the inclusion of varying lengths of poly(dA/dT) sequences at the 3'-end of cDNA templates. Wang *et al.* (2000) have observed that the presence of these poly(dA/dT) hybrids contributes significantly to the problem of loss of low copy number messages due to random annealing between poly(dA) and poly(dT) sequences of unrelated cDNAs. This leads to the formation of tangled double-stranded complexes consisting of multiple cDNA templates. These hybrid complexes are subsequently lost in normalisation/subtraction applications, which are required to reduce high abundance copies and increase the representation of low abundance copies to identify genes expressed at low levels (Bonaldo *et al.*, 1996). However, upon removal of hybrids, copies of many genes inappropriately annealed can be lost. Low level copies are particularly affected. A strategy has been devised to overcome this obstacle, which has been demonstrated to provide a much higher degree of novel gene identification than the current popular techniques (Wang *et al.*, 2000). In this approach, special 3'-anchored oligonucleotide primers are constructed which have only eleven (dT) residues at their 5'-end followed by a different anchoring nucleotide or pair of nucleotides at the 3'-end as follows: 5'-(dT)₁₁(dR)-3' (where R is adenine or guanine) or 5'-(dT)₁₁(dC/dV)-3' (where V is adenine, guanine or cytosine). This greatly increases the probability of primer annealing at the 5'-most region of the poly(A)⁺ tract nearest the 3'-end of an open reading frame. The SuRF method obtains differentially expressed genes from the 3'-noncoding ends. It is speculated that the overall efficiency of SuRF could be enhanced markedly by using the anchoring oligo(dT) primers described in the strategy above in cDNA synthesis.

While it has not been directly demonstrated that the SuRF method has the ability to identify differential expression of low-copy mRNAs, investigations have clearly shown that SuRF has the potential to identify differentially expressed mRNAs at high frequency. A direct demonstration of differential expression of low-copy mRNAs can be extremely challenging. A rational strategy for this would require prior knowledge of low-copy transcripts and an experimental model known to document differential expression of the transcripts. Alternatively, an artificial model containing low levels of synthetic transcripts can be developed to determine

the power of SuRF or related methods. Investigations of this nature are underway. Rapid identification of abundant as well as low-copy differentially expressed mRNAs will be important in understanding the mechanism regulating a biological process. The SuRF method should find wide application in resolving such issues. Although other methods conceptually comparable to SuRF are known in the literature (e.g., Hakvoort *et al.*, 1994), it is thought that SuRF offers a simple strategy to identify differentially expressed genes after normalisation by removing mutually common mRNAs between samples.

3.2.4.12 SSH – suppressive subtractive hybridisation

3.2.4.12.1 Introduction

Solution hybridisation analysis of genomic DNA was first used over 30 years ago to identify families of repetitive sequences and interspersed repetitive and non-repetitive sequences (Britten & Kohne, 1968). Current methods are capable of analysis of much smaller quantities of material than the original protocol. In 1987, subtractive hybridisation was designed as a technique to compare two populations of mRNA and to identify genes that are expressed in one cell or tissue type but not in another (Brown *et al.*, 1997; Candaliere *et al.*, 1999; Hillarby *et al.*, 1996). The method of subtractive hybridisation is often combined with RT-PCR (Brady & Iscove, 1993).

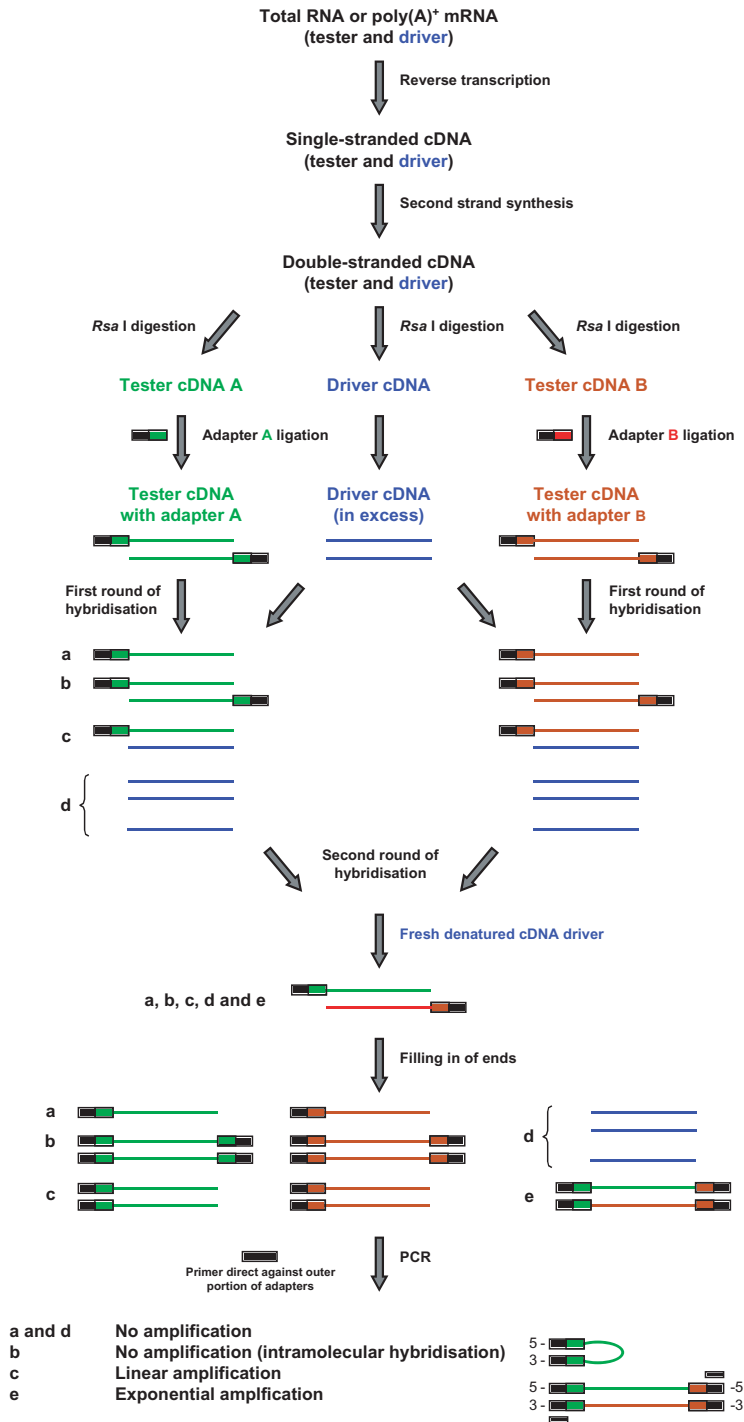
3.2.4.12.2 Principle and basics

The first step is to isolate the RNA populations that are to be compared. These may be either from the same tissue or cells treated in different ways or from two different types of tissue or cells. Both mRNA populations are then converted to cDNA. The cDNA that contains the differentially expressed transcripts of interest is referred to as tester, the reference cDNA is referred to as driver. Tester and driver are then hybridised and the hybrids are removed. The remaining non-hybridised cDNAs represent genes that are present in the tester but absent from the driver mRNA.

In suppression subtractive hybridisation (figure 3.37), both cDNAs are digested with an enzyme that yields blunt ends. The tester cDNA is subdivided into two por-

Figure 3.37 Flow diagram of the suppression subtractive hybridisation (SSH) procedure. In a first step, cDNA is prepared from total RNA or poly(A)⁺ mRNA isolated from different samples (tester and driver). Following synthesis of double-stranded cDNA, tester and driver cDNAs are separately digested using, for example, *Rsa* I to obtain shorter blunt-ended fragments. The tester cDNA is then divided into two pools. The complementary DNA of each pool is ligated to a different adapter (green and red), whereas the driver cDNA (blue) does not undergo ligation. Subsequent hybridisation of tester cDNA with an

excess of driver cDNA leads to equalisation and enrichment of differentially expressed sequences among single-stranded tester molecules. A second hybridisation step and the addition of fresh denatured driver cDNA generates templates for PCR amplification. Due to the suppression effect occurring in the PCR step, only differentially expressed cDNAs are amplified exponentially, whereas other cDNAs are either not amplified or amplified in linear fashion only. The figure is adapted from Diatchenko *et al.* (1996) with permission from *Proceedings of the National Academy of Sciences of the USA*.



tions, each of which is ligated to a different cDNA adapter. The ends of the adapter do not have a phosphate group, so only one strand of each adapter attaches to the 5'-ends of the cDNA. The two adapters have stretches of identical sequence to allow annealing of the PCR primer once the recessed ends have been filled in.

Two rounds of hybridisation are then performed. In the first round, an excess of driver is added to each sample of the tester. The samples are heat-denatured and allowed to anneal, generating different types of molecules in each sample (type 'a', 'b', 'c', and 'd' molecules in figure 3.37). The concentration of high and low abundance sequences is equalised among the type 'a' molecules as re-annealing is faster for the more abundant molecules due to the second order kinetics of hybridisation (James & Higgins, 1985). At the same time, the single-stranded type 'a' molecules are significantly enriched for differentially expressed sequences, as cDNAs that are not differentially expressed form type 'c' molecules with the driver.

During the second round of hybridisation, the two primary hybridisation samples are mixed together without denaturation. Only the remaining equalised and subtracted single-stranded tester cDNAs can re-associate to form type 'e' hybrids. These are double-stranded tester molecules with different ends, corresponding to the sequences of adapters A and B (green and red in figure 3.37). Fresh denatured driver cDNA is added to further enrich fraction 'e' for differentially expressed sequences. After filling in the ends by DNA polymerase, the type 'e' molecules, which represent differentially expressed tester sequences, have different annealing sites for the nested primers on their 5' and 3'-ends.

The entire population of molecules is then subjected to PCR to amplify the differentially expressed sequences. During PCR, type 'a' and 'd' molecules lack primer annealing sites and are therefore not amplified. Due to the suppression PCR effect, most type 'b' molecules form a pan-like structure that prevents their exponential amplification. Type 'c' molecules have only one primer annealing site and are amplified in linear fashion. Only type 'e' molecules, which have two different adapters, can be amplified exponentially. These are the equalised, differentially expressed sequences.

Next, a secondary PCR amplification is performed using nested adapter PCR primers to further reduce any background PCR products and enrich for differentially expressed sequences which are finally identified by sequencing.

3.2.4.12.3 Example

Diatchenko *et al.* (1996) developed the new and highly effective method of suppression subtractive hybridisation (SSH) for the generation of subtracted cDNA libraries. Including a normalisation step which equalises the abundance of cDNAs within the target population and a subtraction step which excludes the common sequences between tester and driver population in their model, they enriched for rare sequences over 1,000-fold in one round of subtractive hybridisation (Diatchenko *et al.*, 1996). The authors documented the usefulness of the method by generating a testis-specific cDNA library and using the subtracted cDNA mixture as a hybridisation probe to identify homologous sequences in a human Y chromosome cosmid library. The human DNA inserts in the isolated cosmids were further

confirmed to be expressed in a testis-specific manner suggesting that the method of suppression subtractive hybridisation is applicable to a wide range of molecular genetic and positional cloning studies for the identification of tissue-specific, developmental or other differentially expressed genes (Diatchenko *et al.*, 1996).

In the experiment, the double-stranded driver cDNA was synthesised from ten different human poly(A)⁺ mRNAs (heart, brain, placenta, lung, liver, skeletal muscle, kidney, spleen, thymus and ovary) by reverse transcription according to a commercial kit's protocol and pooled in equal proportions. The tester was prepared from *Rsa* I or *Hae* III-digested double-stranded testis cDNA as described above and diluted. The diluted tester was then ligated to two different adapters in separate ligation reactions.

The subtractive hybridisation was carried out as described above and for each subtraction there were carried out two PCR amplifications as described above. The PCR products were then analysed by agarose gel electrophoresis. The authors finally cloned and analysed the products by sequencing. Furthermore, the cloned inserts were ³²phosphorus-labelled by random priming and used as probes for Northern hybridisation of human multiple tissue Northern blots.

The efficiency of the subtraction was tested by labelling of the mixtures of non-subtracted and subtracted testis-specific cDNAs which were hybridised to Northern blots of multiple human tissues. The non-subtracted probe hybridised strongly to all RNA samples on the blot and possibly to the common or homologous species of mRNAs among the human tissues. The subtracted and therefore testis-specific cDNA probe hybridised strongly to the testis-derived RNA and only very weakly to all others. This implies that there was achieved a high level of enrichment of testis-specific cDNA and, at the same time, a drastic reduction of highly abundant cDNAs (Diatchenko *et al.*, 1996). The efficiency was further substantiated after cloning and sequencing the testis-specific cDNAs. Ten randomly selected clones were used as probes for further Northern blot analysis on human multiple tissue. All probes revealed unique mRNAs only being expressed in the testis. Also GenBank and EMBL database analyses revealed high homology to known human testis-specific RNAs.

For this specificity, the authors used the subtracted cDNA mixture to screen a cosmid library constructed from flow-sorted human Y chromosomes to successfully identify functional sequences expressed in the testis. A total of 224 out of 3,072 cosmids showed specific hybridisation of which the products were further analysed and evaluated.

This study shows that the technique of subtractive hybridisation seems to be a simple as well as efficient method for generation of cDNAs being highly enriched for differentially expressed genes. All clones analysed gave positive hybridisation signals with the observed RNA in Northern blot analysis.

Since its development in 1996, SSH has become a widely used method for analysing gene expression. This is demonstrated by more than 200 citations in the PubMed database using the phrase "suppression subtractive hybridization" and more than 530 citations of Luda Diatchenko's article describing the SSH procedure for the first time (Diatchenko *et al.*, 1996).

3.2.4.12.4 Alternative approaches and modifications

As an alternative, SSH can be performed using an avidin/biotin-based modification (Sive & St. John, 1988). As described in Byers *et al.* (2000), the driver cDNA in this approach is photo-biotinylated by irradiation using a sunlamp, and Tris/EDTA-saturated 2-butanol and chloroform are used to extract the biotinylated products. Biotinylation of the driver promotes strand breakage during the hybridisation step as the smaller driver fragments hybridise faster than the larger ones. Following hybridisation, the hybridised and non-hybridised biotinylated driver cDNA species are removed by streptavidin precipitation and phenol/chloroform extraction. The subtracted products are used for two or three further rounds of subtraction.

Alternative methods for removal of common sequences include the use of biotinylated primers in the PCR amplification of driver cDNA, the use of streptavidin-coated magnetic beads to remove biotinylated driver cDNAs, and the ligation of a specific primer to the tracer cDNAs, such that only fragments derived from the tracer are amplified and cloned. The first round of subtraction removes the rapidly hybridising repetitive classes of DNA (approximately 40 percent of the genome), while the second and third rounds remove more slowly hybridising sequences, leaving only genes that are unique or massively up-regulated in the tracer. If the technique is working well, less than five percent of the starting cDNA sequences remain at the end (all: Byers *et al.*, 2000).

3.2.4.12.5 Discussion

The use of poly(A)-based RT-PCR prevents loss of cDNAs during PCR steps of the subtractive hybridisation procedure (Brady & Iscove, 1993; Hillarby *et al.*, 1996) and is improving the yield of this method (Byers *et al.*, 2000). The technique relies on hybridisation of cDNA sequences present in both cDNA populations which are then removed ('subtracted') from the reaction. The population of cDNA species that remains is enriched for sequences preferentially expressed in one cell or tissue type.

The advantages of suppression subtractive hybridisation include, especially in combination with RT-PCR, the requirement for very small amounts of mRNA, the ability to detect mRNAs which comprise as little as 0.01 percent of the total mRNA, and the ability to detect novel genes (Byers *et al.*, 2000). Another advantage is that no radioactivity is involved.

The major drawback of SSH is related to the nature of subtraction kinetics. Abundant genes hybridise faster and to a greater level of completion than rare ones, making them more amenable to subtraction. Unfortunately, many interesting regulatory genes are of low abundance. In addition, often several rounds of hybridisation are necessary to avoid retention of sequences that are not differentially expressed.

SSH is also technically demanding and identification of isolated genes by library screening is labour-intensive. A significant loss of cDNA occurs through biotinylation, though this can now be overcome by use of biotinylated PCR primers for amplification of driver cDNAs (Pardinas *et al.*, 1998), or by using paramagnetic oligo(dT) beads (Meszaros & Morton, 1996). Though the method is designed to con-

serve rare RNAs, a consequence of the kinetic nature of the subtraction is that exceedingly rare RNAs may be lost. Conversely, complete removal of cDNAs common to both driver and tracer pools is often not possible leading to the possibility of false-positives, and the need for confirmatory expression studies by, for example, RT-PCR or Northern blot analysis. In practice, the technique reliably isolates genes enriched at least ten-fold in the tester pool compared to the driver (Byers *et al.*, 2000). Furthermore, complete denaturation of tracer cDNA species is required to avoid its persistence after hybridisation with biotinylated driver cDNA, resulting in false-positives. Another aspect is that clones with very high signals may correspond to multiple isolates of abundant mRNAs such as mitochondrial mRNAs (Byers *et al.*, 2000).

A further potential disadvantage is that under standard conditions a few micrograms of poly(A)⁺ mRNA from both RNA populations are needed. In some special cases it might be difficult to obtain such quantities of RNA. This problem can be solved by an incorporated amplification step for both tester as well as driver to generate sufficient quantities of both cDNA samples before initiating the subtractive hybridisation itself. Unfortunately this may result in the loss of some sequences (Diatchenko *et al.*, 1996).

Another point to be mentioned is that each subtractive hybridisation experiment will detect differences only in one direction, whereas other methods such as differential display RT-PCR allow detection of genes whose expression is up or down-regulated (Carulli *et al.*, 1998). Furthermore, only two mRNA populations can be directly compared with each other simultaneously (Zhang *et al.*, 1998). On the other hand, subtractive hybridisation is more sensitive in detecting rare mRNA species.

Overall, however, suppression subtractive hybridisation is an acceptable method to obtain differentially expressed mRNA populations or regulated genes providing appropriate confirmatory evidence, for example, by RT-PCR or Northern blot analyses (Carulli *et al.*, 1998).

3.2.5

Multiplexed DNA and RNA analysis on fluorescent microbeads

3.2.5.1 Introduction

Multiplex analysis of DNA or RNA has become increasingly relevant for the detection of single nucleotide polymorphisms (SNP) and gene expression monitoring. It is usually performed in spatially oriented formats, such as microtiter plates or DNA arrays. The use of fluorescent microbeads as a solid phase for hybridisation based assays represents a recent alternative format. The initial use of fluorescent microbeads was intended to improve immunoassays (Lisi *et al.*, 1982; Vignali, 2000), and this has remained the area of greatest interest until now. However, it was shown a few years ago that bead arrays or suspension arrays are also suitable for nucleic acid analysis (Fulton *et al.*, 1997). There are a small number of commercially available systems and assays from Luminex (Austin, Texas, USA; FlowMetrix™), and BD Pharmingen (San Diego, California, USA; Cytometric

Bead Array™). While most marketed assays are designed for immunoassays, the Luminex system has been used broadly for nucleic acid testing.

3.2.5.2 Principle of bead technology

The basic principle of this type of assay is a chemical reaction using a fluorescent colour-coded microbead as the solid phase. The reaction may be enzymatic in nature, an antigen/antibody interaction, or a nucleic acid hybridisation. The only prerequisite is that the reaction adds or removes a fluorescent signal that differs from the intrinsic colour-coding signal. In the Luminex system, 100 differently coded beads are available. Beads are coded in red while the reaction signal is usually green. However, other colours may serve the same purpose. Multiplexing is achieved by using many differently coded beads. Each of these beads is specific for a defined reaction, for example, different oligonucleotide hybridisations. Thus, a set of 100 uniquely labelled microbeads is equivalent to a 100 spot DNA microarray. Specific populations of microbeads are identified according to their fluorescent properties in a conventional flow cytometer or a similar dedicated flow system. The signal of the chemical reaction is measured as a distinct fluorescence intensity by a third fluorescence detector (figure 3.38). Since three colour capability is standard in commercially available flow cytometers, most laboratories equipped with a flow cytometer will require only appropriate software for data analysis in order to use this method. Bead populations are identified exactly the same way as specific white cell populations, for example, monocytes, in a flow cytometer. The intensity of the hybridisation signal would be comparable to the expression of a surface molecule, for example, DR class II major histocompatibility complex (HLA-DR), on the monocyte surface.

The major potential advantage of this type of assay is its multiplexing capability combined with the comparably small reagent consumption. In addition, it is conceivable that the flow cell could be integrated into a fully automated system. Bead

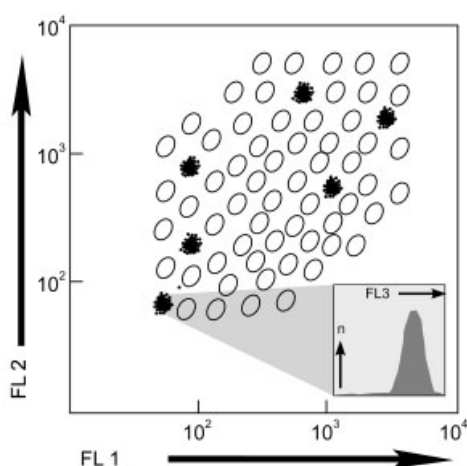


Figure 3.38 Identification of bead populations by flow analysis and measurement of reaction signal. From a large set of fluorescent microbeads, six have been used and were identified according to their fluorescent properties FL1 and FL2. The intensity of the reaction signal is detected as a third fluorescence (FL3) and is depicted as a histogram in the insert.

arrays can easily be customised for the purposes of the user, which is usually not the case for conventional cDNA arrays. The protocols for covalently linking oligonucleotides to the beads can be performed in every laboratory equipped for nucleic acid work. Oligonucleotides are usually coupled to the beads *via* a spacer to permit optimal access to the probe. Common spacers in nucleic acid testing are 15 to 18 carbon atom spacers. Other spacers have been also devised, for example, C6/T5 (six carbon atoms, five thymidine molecules). Luminex offers a so called 'Unilink™' spacer which it claims to be superior to other molecules in terms of coupling efficiency.

Another advantage claimed for the technology is that the reaction kinetics approach those of a fluid phase. Whether the microbead environment really mimics fluid phase conditions for the molecular interactions is not perfectly clear. In terms of molecular sizes, the microbeads still represent a 'large' solid phase. Nevertheless, it appears that the bead-based assays are much closer to fluid phase conditions than for instance microtiter plate assays. The use of so-called ZipCode™ oligonucleotides to capture reaction products to the beads permits the use of real fluid phase hybridisation conditions in a first reaction with later separation of products, which may in fact be advantageous.

Several protocols have been proposed for nucleic acid analysis. Both direct and competitive assay formats can be designed. However, most formats now use direct hybridisation because it has proven to be more sensitive and easier to handle than the competitive formats. The system has been successfully used for mutation detection, RNA quantification, and the detection of nucleic acids. Figure 3.39 summarises the principles of nucleic acid hybridisation tests using microbead technology. For most purposes, an amplification step is required to produce sufficient quantities of the target sequence. This is usually a PCR reaction. However, RNA polymerase-driven amplification is also conceivable for gene expression analysis. In this respect, the 'suspension arrays' resemble conventional DNA microarrays very much.

3.2.5.3 Mutation and SNP detection

Several methods have been devised to detect single or multiple single nucleotide polymorphisms (SNPs) by suspension arrays. Although this book is devoted to gene expression, these methods are described briefly here for the sake of completeness. They are all based on specific hybridisation comparable to DNA arrays. However, they can be divided in two major principles (see also figure 3.39).

One group of assays uses allele or SNP-specific sequences attached to the bead that under the appropriate hybridisation conditions bind directly to the target sequence obtained by PCR. In this case, signal intensity on the bead increases when the target DNA contains the complementary sequence. Alternatively, the assay contains a reporter oligonucleotide which hybridises to the complementary oligonucleotide on the bead (figures 3.39.A and 3.39.B). If the PCR product is complementary to the reporter oligonucleotide, it will be captured and signal intensity on the bead is decreased. This competitive format was proposed initially, but is ap-

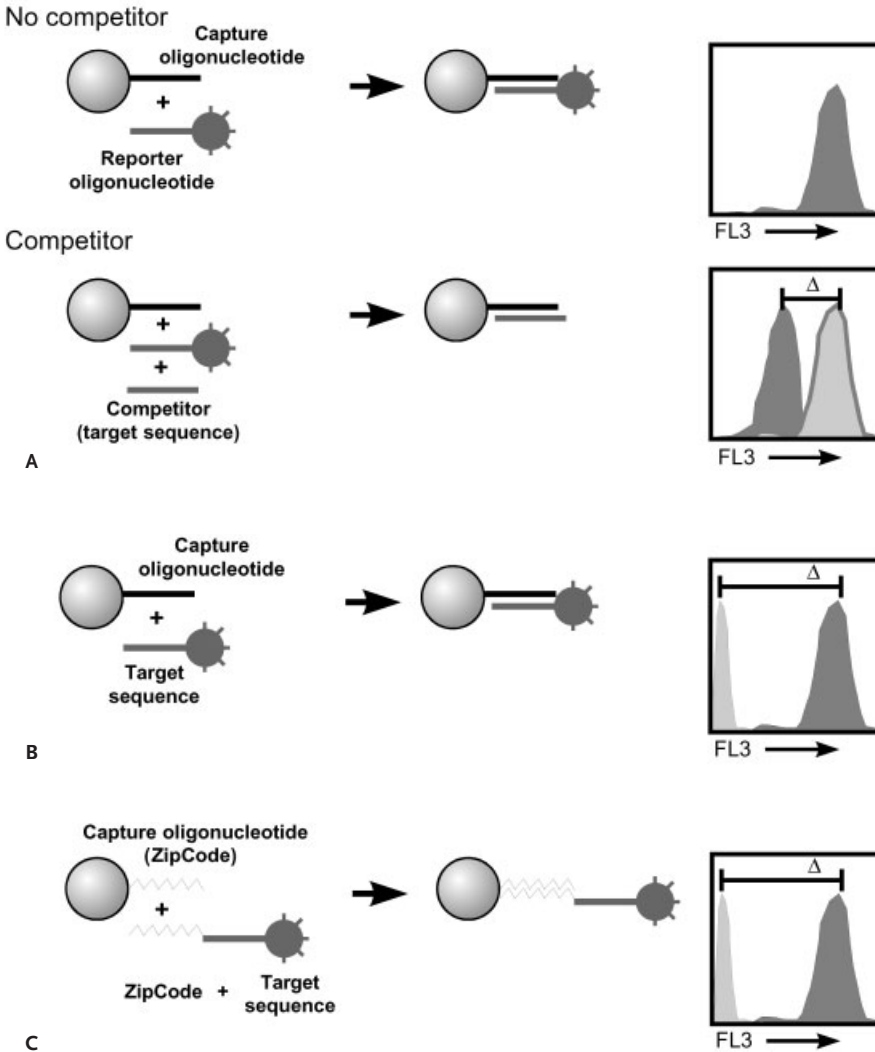


Figure 3.39 Schematic representation of the different approaches to nucleic acid analysis by suspension arrays. **A)** Competitive assay format. **B)** Direct assay format. **C)** ZipCode™-directed assay format. The ZipCode™-based protocol is usually preceded by a more complicated amplification procedure than the

other formats which rely on differential, sequence specific hybridisation. The symbol for a fluorescent label represents also biotin tags that will permit the addition of a streptavidin-coupled fluorophore in a separate step. The following abbreviation is used in this figure: FL, fluorescent property.

parently not superior to the direct format. Recently, Dunbar & Jacobson (2000) reported the simultaneous detection of the five most common cystic fibrosis mutations using Luminex technology. They amplified patient DNA in a multiplex PCR, followed by capture of the amplicons to allele specific capture probes bound to microbeads. The fluorescent label is added after hybridisation *via* a streptavidin tag, immediately before flow analysis.

The second group of assays uses various formats in the initial reaction, for example, oligonucleotide ligation or allele-specific amplification (figure 3.39.C) (Iannone *et al.*, 2000; Ye *et al.*, 2001). Common to these reactions is that a ZipCode™ sequence is attached to the final labelled reaction product providing the complementary sequence is present. In an additional step, the ZipCode™-bearing reaction products are captured by beads coated with the complementary ZipCode™ sequence. The advantage of this system is that only a limited number of ZipCode™-bearing beads needs to be prepared. In fact, the maximal number is the number of different colour-coded beads available. These beads can be subjected to a rigorous quality control scheme to ensure proper loading with the ZipCode™ oligonucleotide. Furthermore, it is easier to establish hybridisation conditions suitable for base pairing of the various ZipCode™ oligonucleotides than for base pairing of many different gene sequences. In particular, ZipCode™ oligonucleotides can be modified to obtain optimal hybridisation properties, which is not the case for genomic sequences surrounding a mutation of interest.

Both types of assays have been shown to work in principle. In particular, both are suitable for multiplexing. However, it should be pointed out that the theoretical number of 100 different beads has not yet been reached by any of the groups working on this topic. This means that a 100-fold multiplexing has not been achieved as yet. Nevertheless, this seems to be a technical rather than a fundamental problem. Thus, it can be expected that it will soon be possible to verify 100 different sequences in a single run. This is sufficient for most clinical applications of mutation screening. This is exemplified by the analysis of the RET (rearranged during transfection) protooncogene for mutations in families with multiple endocrine neoplasia type 2A (MEN2A) or familial medullary thyroid carcinoma (FMTC) (Lackner *et al.*, 1999). Here, some 40 mutations are responsible for the vast majority (more than 99.9 percent) of all cases. Diagnosis is possible after amplification of the relevant exons and subsequent SNP analysis with about 40 different beads.

3.2.5.4 Expression monitoring

Another application of the suspension array system is the analysis of gene expression in a multiplex format. Here, the use of a direct rather than a competitive hybridisation assay is necessary to achieve optimal sensitivity of the system. However, low abundance mRNA species have not yet been quantified reproducibly by this system without prior amplification. In this respect, the suspension arrays are not substantially different from or superior to conventional DNA arrays.

One disadvantage of the system is that the number of mRNAs to be analysed is limited to a maximum of one hundred, or one thousand providing a third colour

can be introduced for the identification of the microbeads. This is far fewer than is possible using high-density cDNA or oligonucleotide arrays. Thus, the technique is not suitable for studying expression of a large number of genes, for example, in screening tumours for aberrant expression patterns. However, the technique may be used to analyse cells or tissues for the expression of a specific set of genes. This has recently been demonstrated for the analysis of gene expression in *Arabidopsis* species by Yang *et al.* (2001). The results obtained by the bead array system were similar to conventional DNA arrays, showing that suspension arrays represent a viable alternative to DNA arrays providing only a limited number of genes need to be monitored.

An appealing feature of the bead array technology is its potential to combine RNA quantification with protein measurement on the same platform. Thus, it should be feasible to establish protocols for the quantification of mRNA and protein, for example, for various cytokines in a multiplexed format.

3.2.5.5 Detection of nucleic acids

The detection of specific nucleic acids, for example, in the diagnosis of infectious diseases or the detection of tumour cells, is a further interesting application of suspension arrays (figure 3.40). In fact, the technology easily lends itself to detect specific amplicons generated by conventional PCR or other techniques. Amplified viral nucleic acid may be probed for specific sequence motifs, thus providing for a fast identification and differentiation of subtypes. There has been already proof of principle that viral nucleic acids can be detected with high efficiency (Smith *et al.*, 1998). The identification and partial quantification of human immunodeficiency virus (HIV), hepatitis C virus (HCV) and *Herpes simplex* virus (HSV) after appropriate amplification by PCR has been reported. The ability of the bead technology to identify subtype specific sequences in one assay is of particular interest.

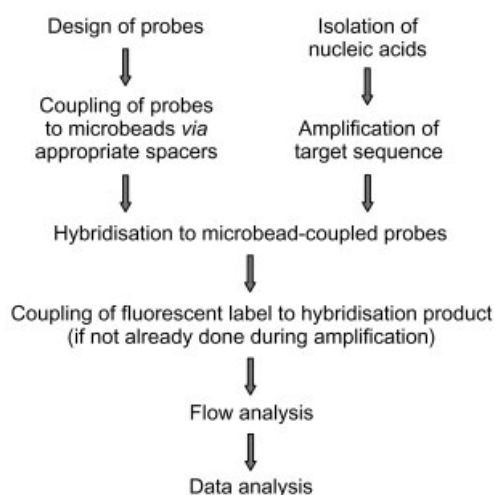


Figure 3.40 Flow-chart of the general procedure for nucleic acid analysis by suspension arrays. See text for details.

3.2.5.6 Conclusions

Suspension arrays provide a highly promising tool for the simultaneous, multiplexed analysis of intermediate numbers of nucleic acids. Their major advantage over conventional DNA arrays is their potential for customised assays and the ensuing lower cost. Furthermore, the technology does not depend on highly expensive dedicated systems, such as DNA array scanners. Since the final step of fluorescence quantification is similar for all analytes, it is even possible to use the same flow cell for various applications from immunochemistry to nucleic acid diagnostics, which could change the face of the current clinical pathology laboratory. At present, the methodology is clearly still at a rather early stage. Nevertheless, it is very likely that we will see a rapid increase in its use and the development of pre-analytical modules that will be necessary for routine use.

3.2.6

RPA — Ribonuclease protection assay

3.2.6.1 Introduction

The ribonuclease protection assay (RPA) is a highly sensitive and specific method for the detection of mRNA species and their structural characterisation. Today, a great variety of convenient RPA kits are commercially available and the method is widely used for the detection of specific RNA species or the study of mRNA splicing pathways or alternative transcription initiation or polyadenylation events. The procedure was first reported by Zinn *et al.* (1983) and subsequently described in detail by Melton *et al.* (1984). It represented an improvement of the S1 nuclease mapping technique (Berk & Sharp, 1977; Weaver & Weissman, 1979) which had been used in many laboratories before. The development of the RPA technique was made possible by the isolation and characterisation of the phage-encoded RNA polymerase of the *Salmonella typhimurium* phage SP6 (Butler & Chamberlin, 1982) and the mapping and cloning of a bacteriophage promoter sequence (Kassavetis *et al.*, 1982; Melton *et al.*, 1984). Subsequently, the RNA polymerase produced by the related *Escherichia coli* phages T3 and T7 have been similarly employed (Davanloo *et al.*, 1984). The RPA technique has a number of advantages over classical S1 nuclease mapping: (i) probe preparation is much more convenient because gel purification can normally be omitted; (ii) unlike end-labelled or nick-translated probes, the specific activities of RNA probes depend exclusively on the specific activity of the labelled ribonucleoside triphosphate, not on the efficiency of the enzymatic reaction; (iii) the *in vitro* transcription reaction yields larger quantities of the probes which have a much higher specific activity than classical end-labelled S1 probes. This is particularly important because it increases the sensitivity of detection dramatically; (iv) the probes are single-stranded and therefore they cannot reanneal; (v) they generate more stable duplexes than would a DNA probe; and (vi) treatment of RNA/RNA duplexes with ribonuclease is more reliable and reproducible than is treatment of RNA/DNA hybrids with S1 nuclease, the latter being sensitive to temperature and enzyme concentration.

3.2.6.2 Basics and principle

The ribonuclease protection assay was made possible by the discovery and characterisation of DNA-dependent RNA polymerases from the bacteriophages SP6, T7 and T3, and the elucidation of their cognate promoter sequences. These polymerases are ideal for the synthesis of high-specific-activity RNA probes from DNA templates because they exhibit a high degree of fidelity for their promoters, polymerise RNA at a very high rate, transcribe long segments efficiently, and do not require high concentrations of rNTPs. Thus, cDNA fragments of interest can be subcloned into plasmids which contain bacteriophage promoters, and the constructs can then be used as a template for synthesis of radiolabelled or biotinylated antisense RNA probes. The principle of the ribonuclease protection assay is illustrated in figure 3.41.

Bacteriophage RNA polymerases possess several properties that make them well suited for the preparation of high-specific-activity hybridisation probes (for review, see Chamberlin & Ryan, 1982). They are single subunit enzymes that are relatively stable and easy to purify. Moreover, because their genes reside on phage genomes, they have been fairly straightforward to clone and express in uninfected *Escherichia coli*, thereby increasing the ease and economy of purification. Second, they polymerise RNA at an exceedingly high rate – 200 to 300 nucleotides per minute – approximately ten times faster than *Escherichia coli* RNA polymerase and even faster than most DNA polymerases. Thus, large amounts of probe are easily prepared. Third, they are very specific in their action, recognising fairly long promoter sequences that are unlikely to appear fortuitously in other DNA. Therefore, the probes are very homogeneous in sequence and usually require little further purification.

To make an RPA probe, it is necessary to first subclone a fragment containing the sequences of interest downstream of a phage promoter. The sequence to be analysed must be cloned such that the RNA produced by the phage polymerase is complementary to the RNA to be analysed. Ideally, this construct should be able to be digested with a restriction enzyme to produce a linear template that will be transcribed into a 100 to 300-base runoff transcript. For example, if one wished to analyse the level of appropriately initiated transcription from a promoter, a DNA fragment from that promoter with endpoints at +150 and –100 could be cloned such that the +150 site is immediately adjacent to the phage promoter. Cleaving the resultant clone at a restriction site adjacent to the –100 site would then allow runoff synthesis of a 250-base probe that would give a 150-base signal in the assay (figure 3.42).

3.2.6.3 Procedure and typical results

The procedure starts with cloning the probe DNA into a cloning vector containing a T3 or a T7 bacteriophage promoter upstream of the cloned fragment. Many of the currently used cloning vectors contain both T3 and T7 promoters flanking their multiple cloning sites (for example, see the pBluescript series from Stratagene, La Jolla, California, USA; figure 3.43). Once the DNA fragment has been cloned,

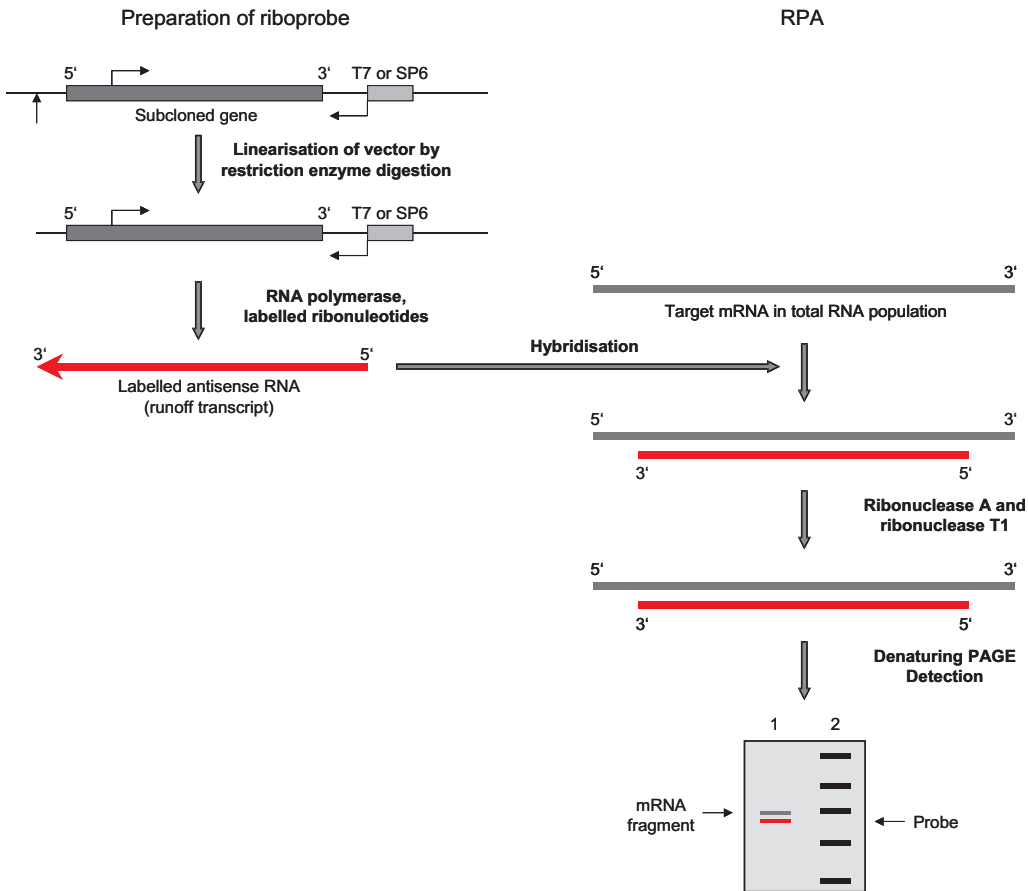


Figure 3.41 Flow-chart illustrating the principle of the ribonuclease protection assay (RPA). In a first step, a labelled antisense riboprobe is constructed by *in vitro* transcription using a cDNA template subcloned in a vector containing at least a T7 or SP6 promoter to allow RNA polymerase-mediated *in vitro* transcription. The purified, labelled antisense riboprobe is hybridised to a pool of total RNA of interest. The reaction mix, containing the resulting perfect matched double-stranded RNA/RNA hybrid molecules and unpaired, single-stranded RNAs, is usually treated with a mixture of ribonucleases A and T1. Both ribonucleases digest single-stranded RNAs specifically. In contrast to ribonuclease T1, RNase A, which is isolated

from bovine pancreas, cleaves phosphodiester bonds at pyrimidine nucleotides. RNase T1, which is isolated from *Aspergillus oryzae*, cleaves at guanine nucleotides only. This treatment results in a digestion of all single-stranded RNAs, *i.e.* all RNAs that are not complementary to the riboprobe and the single-stranded overhangs of the mRNA/riboprobe hybrid. Following RNase treatment, protected fragments are separated electrophoretically by denaturing polyacrylamide gel electrophoresis (PAGE) and detected by auto-radiography or the biotin detection system. Using this procedure it is possible to quantify multiple mRNAs in a complex mixture of total RNA in a single experiment.

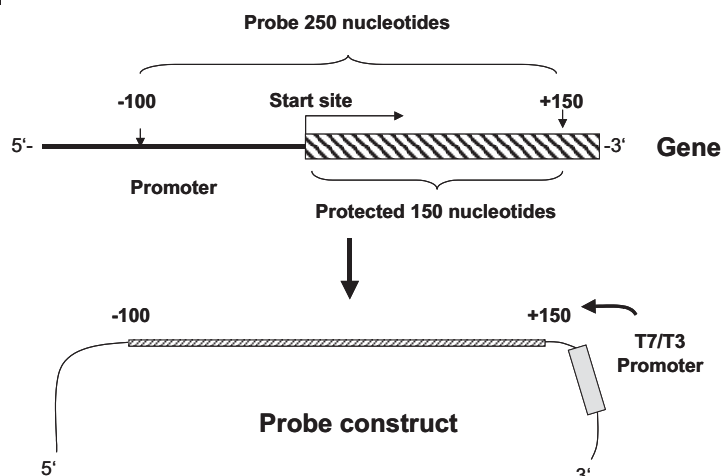
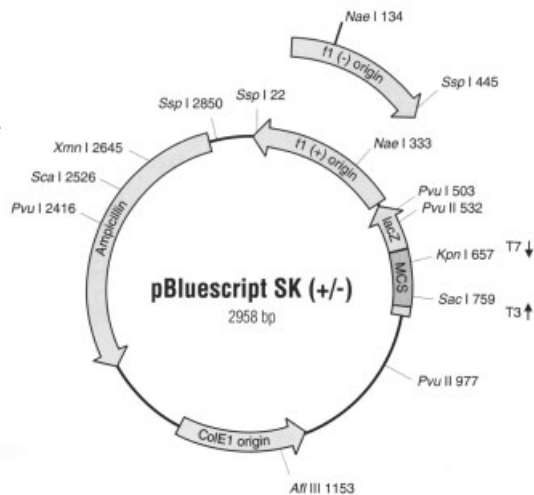


Figure 3.42 Example for a design of a typical ribonuclease protection assay (RPA) for the detection of correctly initiated transcripts. The sizes and positions relative to the start site of the gene are approximations and may vary

considerably depending on the situation. The cloning vector may be any plasmid containing an appropriate SP6, T3 or T7 RNA polymerase promoter in the context of the probe insert.

Figure 3.43 Map and sequence of the multi-cloning site of the pBluescript SK vector (Stratagene, La Jolla, California, USA). The vector provides a particularly useful template for the synthesis of ribonuclease protection assay (RPA) riboprobes, since T3 and T7 RNA polymerase promoters are located 5' and 3' from the multi-cloning site, thus allowing transcription from both strands of the insert.



5' GGAAACAGCTATGACCATGATTACGCCAGCTCGAATTAACCTCTCAATAGGGAACAAAGCTGGAGCTCCACCGGGTGGCGCCGCTCTAGAACTAGTGATCCCCGGGCTGCAGG
3' CCTTTGTCTGACTGCTACTAATGCGGTTTCGAGCTTTAATTGGAGTGAATTCCTGTTTTCGACCTCGAGGTGGCGCCACCGCGGAGATCTTGTATCACTAGGGGGCCGACGCTCTTAA

816 β-Galactosidase →

759

5' AATTCGATATCAAGCTTATCGATACCTGACCTCGAGGGGGGGCCGGTACCCCAATTCGCCCTATAGTGAGTGTATTAACTTCACTGGCGCTGTTTACAA 3' (+)
GCTATAGTTCGAATAGCTATGGCAGCTGGAGCTCCCCCGGCGCATGGGTTAAGCGGGATATCACTAGGATTAAGTGAACCGGCAAGAAATGTT 5' (-)

657 ← +1 T7 promoter

one has to be sure about the orientation of the insert because the *in vitro* transcribed probe must be complementary to the mRNA of interest. After a plasmid preparation, which should be absolutely RNase-free, the plasmid is linearised by complete digestion with an appropriate restriction enzyme that cuts downstream of the inserted DNA fragment. Under standard condition, one microgram of linearised plasmid is incubated with ten units of either T3 or T7 RNA polymerase in a 25 microlitres volume of reaction buffer in the presence of 400 micromolar rNTPs. Depending on the desired specific activity of the probe, the non-labelled rNTPs are replaced by rNTPs that are labelled in α -position with ^{32} phosphorus or biotinylated rNTPs which are available from, for example, Roche Diagnostics GmbH (Mannheim, Germany). Probe synthesis requires at least 30 minutes at 37 °C but is preferentially be performed for as long as three hours.

Over half of the label in the transcription reaction is typically incorporated into the probe (about 10^8 counts per minute). If substantially lower incorporation is achieved, the specific activity of the probe is usually just as high, but the yield is reduced. Therefore, such a probe should work just as well as a probe from an efficient transcription reaction. However, in practice, greater backgrounds may result when probes from low-incorporation reactions are used. If the probe is purified by gel electrophoresis, yields are lower, typically 2 to 5×10^7 counts per minute which should be enough for at least 40 standard hybridisations.

The hybridisation signal resulting from a fairly low abundance mRNA in ten micrograms of total RNA is in most instances detectable after an overnight exposure using an intensifying screen. Often, shorter exposures or exposures without a screen will suffice. Including gel purification, the whole procedure, including probe electrophoresis and elution, takes about six to eight hours. It is usually convenient to precipitate the RNA samples with ethanol for hybridisation during this time. It is important to note that radiolabelled probes are not very stable. Therefore, they should be used on the same day they are synthesised. Thus, a convenient arrangement is to prepare the probes and set up the hybridisations the same day. The following morning, nuclease digestions and gel electrophoresis are performed. Digestion with RNase is performed for one to two hours. Subsequently, the samples are purified and subjected to electrophoresis under denaturing conditions. It is important to include appropriate size markers in the analysis to enable size determination of the fragments that are protected from RNase digestion.

Apart from the classical radioactive RPA assay, several non-isotopic formats have been developed. The most important represents the biotin-based RNA labelling procedure. Although the biotin detection system requires additional steps, *i.e.* the transfer of the nucleic acids from the polyacrylamide gel to a positively charged nylon membrane, it is very convenient because neither exposition to potentially hazardous radiation occurs nor is radioactive waste produced. Moreover, biotinylated probes can be stored for several months at -80 °C because they are much more stable than radiolabelled probes.

A more recent development consists of the lysate ribonuclease protection assay. Ambion, Inc. (Austin, Texas, USA) offers a test kit (called 'Direct Protect™') for the detection, quantification and characterisation of RNA directly in crude lysates from

cultured cells or solid tissues without the need for prior RNA isolation. The procedure is quick, and takes place in a single test tube. Since there are no precipitation or transfer steps, one does not have to worry about RNA losses from small samples, differential recoveries between samples, or partial degradation during extended handling. A labelled RNA probe is hybridised directly with cellular RNA in guanidine thiocyanate which is a chaotropic salt that rapidly solubilises cells and tissues while inactivating cellular ribonucleases and enabling direct hybridisation. When tissue is precious or limiting, single samples can be hybridised with multiple probes. After hybridisation, excess RNA probe and sample RNA are removed by digestion with a mixture of ribonucleases and the products are then precipitated. Protected fragments are separated on a denaturing or native polyacrylamide gel and either transferred to a membrane for non-isotopic detection or exposed to film directly when radioactive probes are used. The sensitivity is comparable or even superior to conventional methods. Using a lysate ribonuclease protection assay, Strauss & Jacobowitz (1993) have been able to detect 43 femtograms of mRNA per microgram of protein of a specific, moderately abundant mRNA from as few as two one micrometre diameter punches, containing only 40.9 ± 2.8 micrograms of protein. The group found that the amount of specific message in the brain punches was reproducible and correlated well with that quantified by *in situ* hybridisation histochemistry (without the weeks to month exposure times often needed for the latter technique). A typical result obtained with the Direct Protect™ system is shown in figure 3.44.

3.2.6.4 Conclusions

The most important advantage of the RPA technique relates to the fact that the methods not only gives information about the structures of specific mRNAs but at the same time also allows their quantification. Although it is possible to analyse several mRNA species in a single run, careful optimisation is required for each case. Thus, the procedure is not well suited to serve as a high-throughput technique.

3.2.7

Virtual Northern blot

3.2.7.1 Introduction

Differential gene expression in time and/or location is a fundamental phenomenon found in all biological systems. The knowledge of those genes that are expressed in response to external stimuli during specific periods of development, during tumourigenesis, or in specific parts of the body (Franz *et al.*, 1998) should assist us in understanding the molecular basis underlying these processes. These scientific problems could be analysed with a huge variety of different methods. Among them are subtractive hybridisation, polymerase chain reaction (PCR)-based mRNA fingerprinting, differential display RT-PCR (DDRT-PCR) or, most notably, DNA array analysis. Due to the inherent amplification steps, some of these

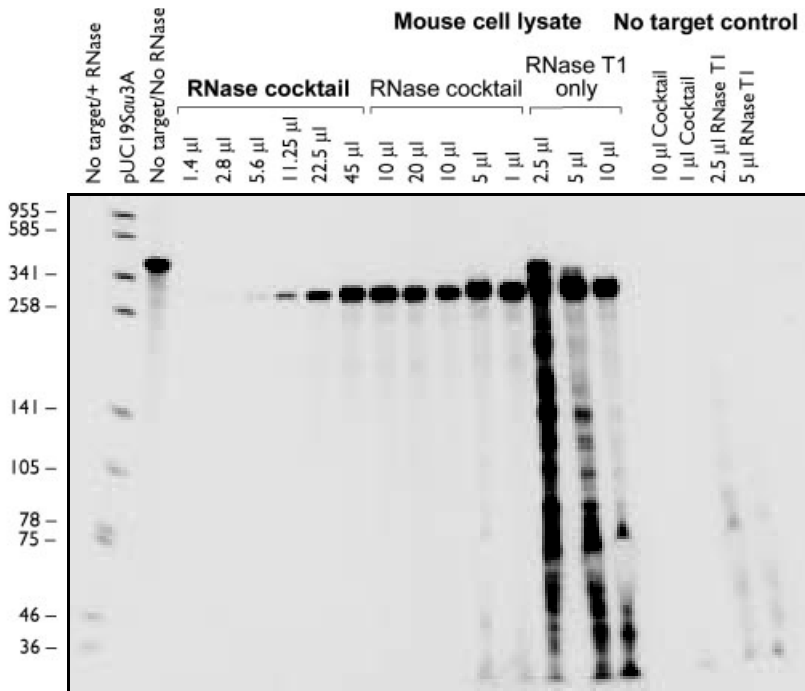


Figure 3.44 Example for a typical result obtained with the Direct Protect™ Lysate Ribonuclease Protection Assay from Ambion, Inc. (Austin, Texas, USA).

The indicated amounts of lysate from mouse cells were directly hybridised to 9.0×10^4 counts per minute of a ^{32} phosphorus-labelled glyceraldehyde-3-phosphat dehydrogenase (GAPDH) control probe (specific activity: 5×10^8 counts per minute per microgram). The ribonuclease digestion was per-

formed according to the Ambion instruction manual under varying conditions, as indicated. The protected fragments were electrophoresed on a five percent polyacrylamide gel containing eight molar urea (0.75 millimetres thick). The dried gel was exposed to X-ray film for two hours with an intensifying screen. The figure is reproduced from the Direct Protect™ Lysate RPA instruction manual with permission from Ambion, Inc. (Austin, Texas, USA).

methods have the potential to start with minimal amounts of tissue sample. All these high-throughput approaches that screen numerous clones at the same time have the inherent problem of inaccuracies and false-positives. Verification steps using alternative approaches are therefore required. Usually, Northern blots or RNase protection assays are the methods of choice. For both methods a relatively large amount of material is needed, which excludes various fields of interest from this kind of verification.

To circumvent these problems, methods with intrinsic amplification steps are required. Conventional PCR reactions amplify DNA regions surrounded by known sequences. For global amplification processes of cDNA populations, only the 3'-region of the RNA is accessible, as known sequences could be included during first-strand cDNA synthesis. The nature of the 5'-end is unknown, a problem that must

be solved to enable global cDNA amplification. A number of different approaches allow tagging of 5'-ends to obtain information about the corresponding sequences. If this tagging of 5'-ends is done, the cDNA is enclosed in known sequences, a prerequisite for PCR-based amplification approaches. In addition to the characterisation of 5'-ends, they are extremely useful for the amplification of complete and complex cDNA mixtures. At least four different approaches are currently available for the characterisation of 5'-mRNA ends. The first is homopolymeric tailing using terminal deoxynucleotidyl transferase and dGTPs (TdT tailing) (Schaefer, 1995). This aligo (dG) tail is recognised by an oligo(dC) primer, thus tagging the 3'-end of the first-strand cDNA with a sequence known to be suitable for PCR amplification. More recently, two other approaches have been introduced. An oligonucleotide is added to the first-strand cDNA or an oligoribonucleotide is added to the 5'-end of the mRNA, both of which reactions are catalysed by the T4 RNA ligase (ligation-anchored and RNA ligase-mediated; LA and RLM, respectively) (Franz *et al.*, 1999; Schaefer 1995). The fourth approach, the so-called cap-finder cDNA depends on the ability of reverse transcriptase to add nucleotides to the 3'-end of the newly synthesised cDNA after completion of first-strand synthesis (Franz *et al.*, 1999). As the Moloney murine leukemia virus (MMLV) reverse transcriptase preferentially adds cytosines (Peliska & Benkovic, 1992), a second primer that contains a guanine-stretch at its 3'-end is added to the reaction mixture. This anneals to the complementary region and serves as a template for the reverse transcriptase, leading to a template switch, thus delivering cDNA with known sequences at the 5' and 3'-terminal end (Franz *et al.*, 1999; Schramm *et al.*, 2000).

3.2.7.2 Principles and basics

To perform virtual Northern blots, it is important to start with equal amounts of RNA. RNA quantities ranging from a maximum of one microgram to the amount obtained from a few cells are used for the first-strand cDNA synthesis. RNA is isolated according to standard protocols. Obtaining RNA of high quality is of crucial importance for the entire process. All the following steps are performed on a solid support (Roeder, 1998 and 2000). This allows rapid buffer exchanges with no risk of losing time or precious material. Coupling to a solid support is achieved *via* biotinylated oligo(dT) primers that are coupled to avidin-coated beads. Among the different methods outlined above, we will focus on two approaches: a rapid one and a sophisticated, more reliable one. The rapid approach usually yields excellent results. It is based on the cap-finder technique outlined in figure 3.45 (Franz *et al.*, 1999). Its main advantage is its simple experimental set-up that minimises errors introduced during cDNA synthesis. The other approach is more demanding, as a sequence of enzymatic manipulations precedes cDNA synthesis and PCR amplification (Barnes, 1994). Nevertheless, it has various fields of applications where it is superior to other approaches. As only full-length amplicons are generated during this process, more accurate results than those obtained with conventional Northern blots are obtained.

3.2.7.3 Typical results

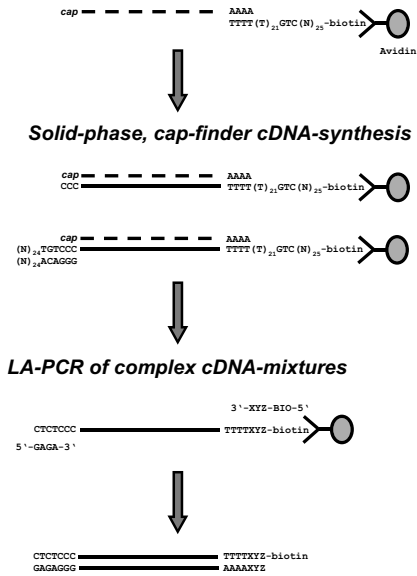
To take advantage of the benefits of the virtual Northern blot approach, one precaution must always be kept in mind. During PCR amplification, a shift in the relative concentration of transcripts shifts towards small or low abundance transcripts may occur. These detrimental effects of PCR amplifications are especially encountered if the amplification process reaches the saturation phase. To avoid this, it is necessary to perform pilot PCR reactions with increasing cycle numbers (figure 3.46). Using this simple approach, it is easy to identify those conditions that are ideally suited for the production of virtual Northern blots. The PCR reaction is in the exponential phase if each additional cycle doubles the amount of DNA product as detected on a gel or by real-time RT-PCR. If this is taken into account, virtual Northern blots give excellent results both regarding the relative amounts of transcripts and the length of the amplicons. A typical virtual Northern blot is shown in figure 3.47. Hybridisation and reactions conditions are as for conventional Southern blots. Both radioactive and non-radioactive strategies can be used. The information that can be extracted from such an experiment is almost identical to that obtained from conventional Northern blot experiments. In this example, a virtual Northern blot analysis performed in *Caenorhabditis elegans* treated with tetrabutyl hydroxypoxide is shown. Hybridisation with a probe derived from worm peroxiredoxin, using equal amounts of cDNA revealed signals that differ in their intensity. Although only small amounts of starting material were available, differences in expression of the peroxiredoxin gene under different experimental conditions could be observed.

3.2.7.4 Discussion

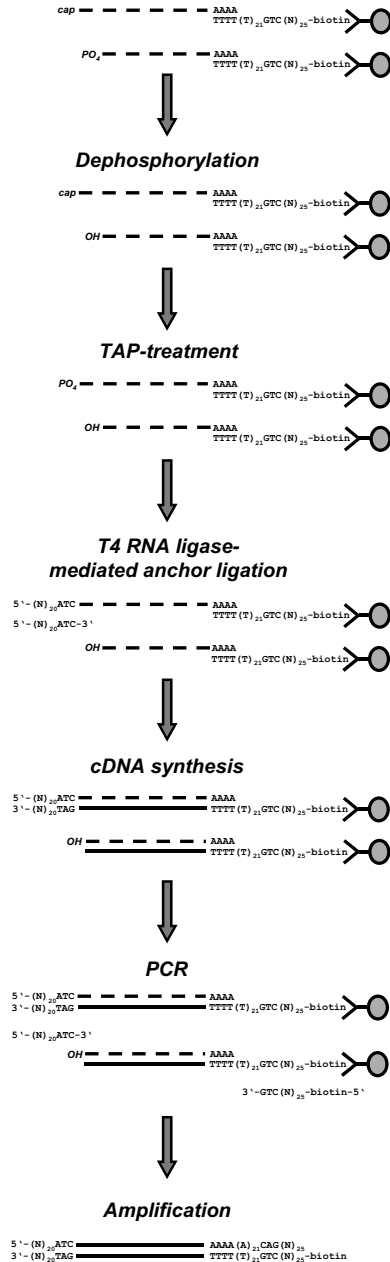
Virtual Northern blots are an ideal alternative to conventional Northern blots or the ribonuclease protection assay. If performed properly, they provide the same amount of information as conventional Northern blots. This includes information on the length of the amplicon as well as on the relative abundance of the mRNA/cDNA population of interest. Virtual Northern blotting can be used even if starting material is very limited or if the material is too precious to load it onto a Northern blot. Alternative approaches such as RT-PCR are more error-prone than virtual Northern blotting. In addition, important information regarding the length of the transcript is lost in these studies.

As numerous different methods may be used to produce the virtual Northern blots, the advantages and disadvantages of each approach should be taken into account. These methods differ significantly in their experimental complexity. The cap-finder cDNA approach outscores the other approaches because only a single experimental step, the first-strand cDNA synthesis precedes the long and accurate PCR (LA-PCR) process. All other approaches consist of a chain of enzymatic reactions, which may introduce numerous failures. The ultimate approach regarding the amplification of complex cDNA mixtures is outlined in figure 3.45. Only complete cDNAs participate in the amplification process. But it should be kept in mind that this approach is experimentally demanding.

Cap-finder



RLM



◀ **Figure 3.45 Schematic outline of the production of virtual Northern blots using either the cap-finder approach (left) or the RLM (T4 RNA ligase mediated)-based amplification of complex cDNA mixtures (right).** The cap-finder cDNA synthesis approach requires only a single PCR reaction. By the unique ability of Moloney murine leukemia virus (MMLV) reverse transcriptase to add several cytosines to the 5'-end of the cDNA is exploited to tag the 5'-end with a known sequence. The addition of an oligonucleotide that contains at least three guanines at its 3'-end leads to hybridisation with the cytosines added by the reverse transcriptase to the 5'-end of the cDNA. The reverse transcriptase now accepts the 5'-part of the annealed oligonucleotide as a new template and completes the cDNA complementary to the added oligonucleotide. Whereas the cap-finder approach requires only a single reaction until PCR-amplification is possible, the RLM ap-

proach requires a set of different enzymatic reactions. Its rationale depends on the enzymatic removal of the 5'-cap followed by ligation of a short DNA or RNA anchor. The first of a number of enzymatic steps is the removal of phosphate groups from incomplete mRNA populations, which are devoid of the cap structure. As the next step, tobacco acid pyrophosphatase (TAP) removes the 5'-cap from the mRNA, leaving a phosphate residue on the mRNA. This is required to ligate an RNA or DNA oligonucleotide to these phosphorylated mRNA populations. Only those RNAs which originally possessed an intact cap structure now bear this oligonucleotide. After conventional cDNA synthesis, a PCR employing oligonucleotides derived from the known 5' and 3'-ends amplifies the entire cDNA population. The following abbreviation is used in this figure: LA-PCR, long and accurate PCR.

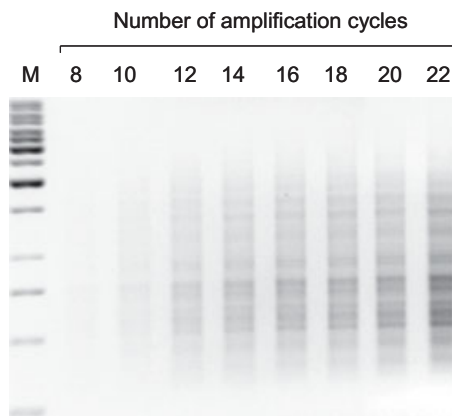


Figure 3.46 Whole cDNA amplification using increasing numbers of PCR cycles. First-strand cDNA produced by the cap-finder method from 0.5 micrograms of total RNA derived from *Entamoeba histolytica* was subjected to a long and accurate PCR (LA-PCR) amplification process using increasing numbers of PCR cycles. In the region between cycles ten and 22, cDNA amplicons were identified visually. Between cycles ten to 16 an almost linear increase of the re-

action product is seen. From cycle 18 onwards, the total amount of cDNA produced remained almost constant, indicating saturation of the amplification process. To ensure that the relative concentrations of cDNA populations are conserved in these samples, it is necessary to restrict the PCR amplification process to the exponential phase. The test PCR using different numbers of cycles is required to exclude false results.

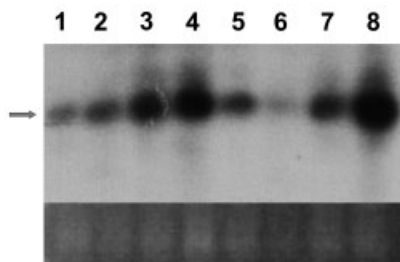


Figure 3.47 Virtual Northern blot of a peroxiredoxin gene from the nematode *Caenorhabditis elegans*. Nematodes confronted with different concentrations of tetrabutyl hydroperoxide were used for the experiments. Following RNA isolation, the cDNA was amplified according to the cap-finder approach. Equal amounts of cDNA were loaded (see lower part of the figure; ethidium-bromide staining). The cDNA is visible as a weak smear with distinct bands representing highly abundant transcripts (only a small part of the gel is displayed). Hybridisation (upper

part of the figure) with a probe derived from a peroxiredoxin gene (F09E5.2) showed different signal intensities (arrow) representing different expression levels of the corresponding gene. The worms were treated with increasing concentration of tetrabutyl hydroperoxide (*lanes 1 and 5*: control; *lane 2*: 0.5 millimolar for one hour; *lane 3*: one millimolar for one hour; *lane 4*: 1.5 millimolar for one hour; *lane 6*: 0.5 millimolar for three hours; *lane 7*: one millimolar for three hours; *lane 8*: 1.5 millimolar for three hours).

It is possible to perform virtual Northern blots with RNA obtained from a few cells (about 50 to 100 cells). The material is sufficient for 30 PCRs (100 microlitres total volume, each resulting in 30 to 50 micrograms DNA), enough to perform more than 100 virtual Northern blots. These virtual Northern blots have the same performance as conventional Northern blots performed using microgram quantities of mRNA. It must be kept in mind that PCR-based approaches may lead to changes in the composition of the cDNA populations, introducing failures in the verification of differential expression. If all precautions are followed carefully, especially when the PCR process is in the exponential range, it is usually possible to circumvent these problems. In the authors' hands, the virtual Northern blot approach was capable of detecting three to five-fold differences in gene expression.

In summary, virtual Northern blots produced by either of the methods described above may serve as an alternative to conventional Northern blots in many fields of molecular biology, allowing analyses to be performed even when starting material is limited. Use of virtual Northern blotting need not be restricted to the verification of differential expression patterns, but may prove useful in all circumstances where the amount of RNA is limited.

3.3

PCR-based methods

3.3.1

ADDER – Amplification of double-stranded cDNA end restriction fragments

3.3.1.1 Introduction

Numerous methods are currently available for the comparison of mRNA populations and many of them are discussed in this book. This section focuses on a recently developed technique, dubbed ADDER (amplification of double-stranded cDNA end restriction fragments), and compares its advantages and drawbacks to those of other commonly employed approaches. The emphasis is placed on the analysis of highly complex mammalian mRNA populations, since this is a particularly demanding challenge. In order to set a basis for this discussion, this section begins with a short description of the difficulties encountered in the qualitative and quantitative description of mammalian gene expression patterns.

The transcriptome describes the identities and expression levels of all RNA species accumulating in a given cell type or organism (see chapter 1). In the literature, however, this term is frequently restricted to mRNA populations. In fact, although genes encoding proteins produce only a small percentage of cellular transcripts, they account for the bulk of RNA complexity. The entire human genome contains approximately 30,000 protein-encoding genes (Lander *et al.*, 2001, Venter *et al.*, 2001). On the basis of various independent experimental approaches, including kinetic hybridisation assays (Ryffel & McCarthy, 1975), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995) and mRNA display methods (*e.g.*, Kornmann *et al.*, 2001, Liang & Pardee, 1992) it can be estimated that about half of these genes are expressed in a given cell type. However, the transcripts of these genes accumulate to highly variable levels. Thus, the mRNAs produced by only a few genes may make up most of the mRNA mass. For example, in a mouse exocrine pancreas cell, a few transcripts account for most of the mRNA mass (Otto Hagenbüchle and Peter K. Wellauer, personal communication).

Not all cell types are as highly specialised as exocrine pancreas cells. However, it can be assumed that in most cell types, one third of all expressed genes (approximately 5,000) produce about 90 percent or more of the mRNA mass, while the remaining two thirds (approximately 10,000) contribute only 10 percent of the cellular mRNA quantity (Ryffel & McCarthy, 1975; Yamashita *et al.*, 2000). Hence, on average, most mRNA species each contribute about one hundred thousandth of the total mRNA transcriptome. A liver cell (parenchymal hepatocyte) harbours about 50 picograms (5×10^{-11} grams) of total RNA and about 1.5 picograms of mRNA (Edward E. Schmidt and Ueli Schibler, unpublished data). Assuming an average molecular mRNA length of 2,000 nucleotides and an average molecular mass of 340 daltons per nucleoside monophosphate, the total number of mRNA molecules per hepatocyte can be estimated to be about 1.3×10^6 . Thus, even in a large cell like a parenchymal hepatocyte, most mRNA species belong to a frequency distribution peaking at about ten molecules per

cell, implying that thousands of transcripts accumulate to even lower cellular concentrations.

The low abundance of most mRNA transcripts renders the complete and accurate description of a cellular transcriptome very difficult. Matters get worse when one takes into consideration that most tissues are composed of many different cell types. In RNA preparations from such heterogeneous tissues, the number for any cell type-specific transcript is further reduced by the dilution of the cells of interest by other cell types not expressing this transcript. As a consequence, techniques aimed at the identification of all or most transcripts present in the sample require an extremely high sensitivity and specificity of signal detection.

3.3.1.2 Principles and basics

3.3.1.2.1 Experimental design of the ADDER procedure

The ADDER differential display technology relies on the synthesis and PCR amplification of double-stranded cDNA restriction fragments complementary to the 3'-terminal moieties of mRNAs. A related cDNA display method, systematic differential display, has been described by Ke *et al.* (1999), but this procedure involves a restriction enzymes cutting at a distance from the recognition sequence. Although *Mbo* I was used in the original ADDER protocol, any frequently cutting restriction enzyme should be suitable for the generation of ADDER double-stranded cDNA libraries (see below).

Figure 3.48 schematically displays the eight experimental steps of the ADDER procedure: (i) The polyadenylated RNA within whole cell RNA is reverse-transcribed using an anchored oligonucleotide primer carrying a recognition sequence for the infrequently cutting restriction endonuclease *Asc* I and a 5'-terminal biotin group. (ii) The resulting mRNA/cDNA hybrid is converted into double-stranded cDNA using DNA polymerase 1 and ribonuclease H, and the double-stranded cDNA is immobilised on paramagnetic streptavidin beads. (iii) After washing the beads (by magnetic attraction and release) in an appropriate restriction buffer, the adsorbed double-stranded cDNA is cleaved to completion with a frequently cutting restriction enzyme, such as *Mbo* I. (iv) After washing and resuspension of the beads in ligation buffer, the cleaved end of the double-stranded cDNA is ligated to an excess of a double-stranded oligonucleotide adapter. (v) The beads are washed, resuspended in *Asc* I restriction buffer, and the double-stranded cDNA is released from the beads by digestion with *Asc* I. (vi) The released double-stranded cDNA fragments are PCR amplified with appropriate forward and backward primers to generate a stock for the differential displays. (vii) Aliquots of the double-stranded cDNA stock are PCR amplified in 192 reactions with combinations of twelve forward display primers and 16 backward display primers. The 17-mer forward primers vary in the two nucleotides complementary to the ultimate and penultimate nucleotides of the mRNA (*i. e.*, the two nucleotides preceding the poly(A) site), while the decamer backward primers vary in the two nucleotides following the restriction site used to generate 3'-terminal cDNA fragments. In order to increase the speci-

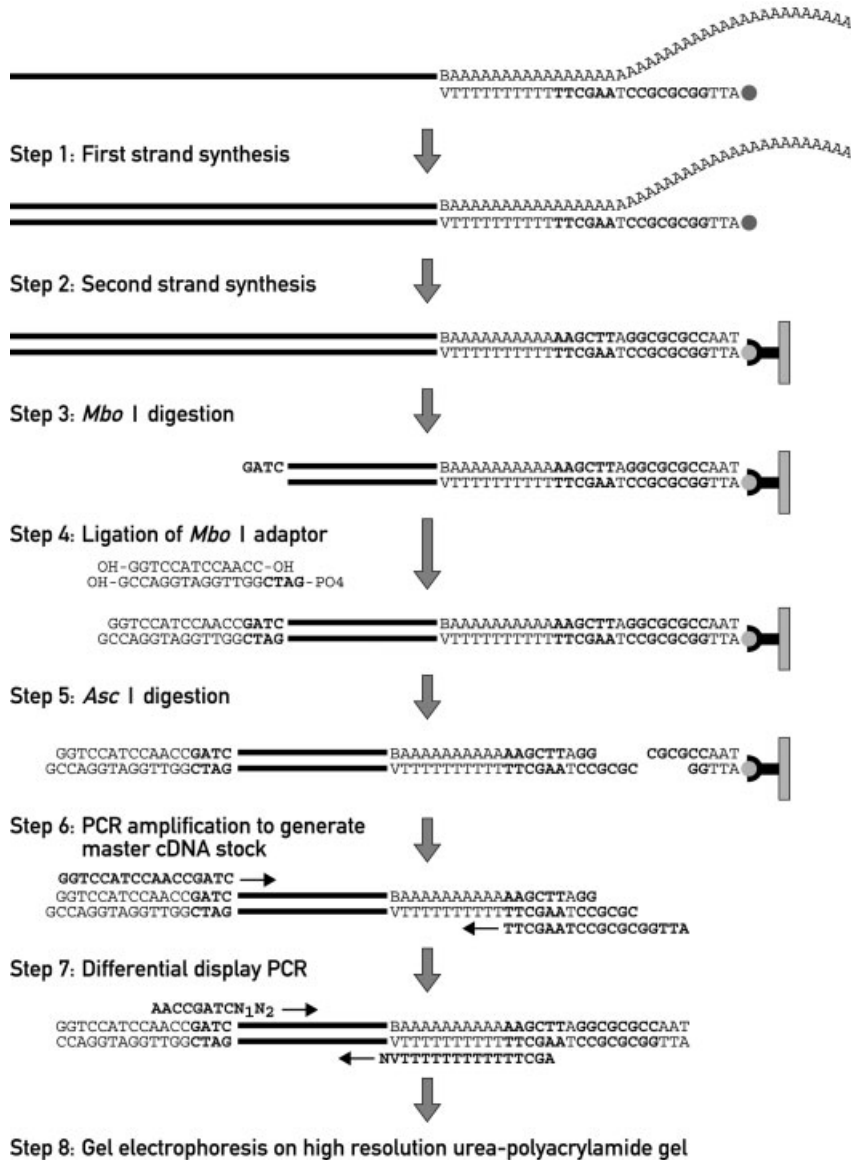


Figure 3.48 Flow diagram of the amplification of double-stranded cDNA end restriction fragments (ADDER) procedure. The eight steps required to perform ADDER differential display are described in detail in the text. In this figure, B is cytosine, guanine, or thymine, V is cytosine, guanine or adenine, and N is either adenine, cytosine, guanine, or thymine).

ficity in the display reactions, touch-down PCR (Don *et al.*, 1991) is used in the 192 PCR reactions with display primers. The accuracy of the ADDER chemistry has been examined and confirmed in the original report (Kornmann *et al.*, 2001).

ADDER was successfully used to compare as many as twelve different mRNA populations. This necessitates the electrophoretic display of 2,304 PCR reactions on about 60 high-resolution urea/polyacrylamide gels and can be completed in two weeks by a single person. In fact, the most time-consuming part in any differential display method is the cloning and identification of candidate cDNAs for differentially expressed mRNA species, in particular if these are numerous.

3.3.1.2.2 Identification of differentially expressed transcripts

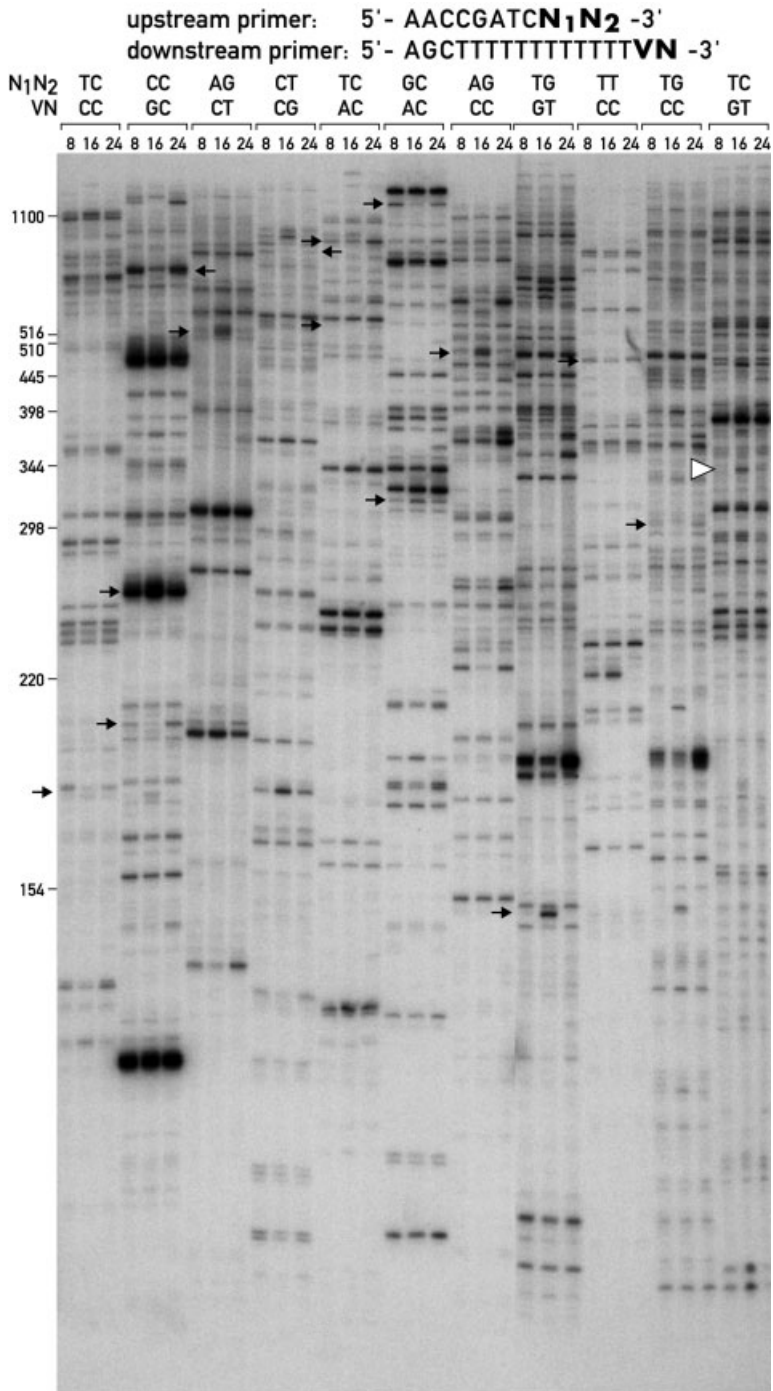
Candidate cDNAs for differentially expressed mRNAs are eluted from dried gel slices, PCR amplified with full-length primers (*i.e.*, primers complementary to the entire lengths of the linkers contained in the cDNA fragments of the master stock (see step 7 in figure 3.48), cleaved with *Asc* I and *Mbo* I (or another restriction enzyme 'X' used to generate the 3'-terminal cDNA fragments), purified by agarose gel electrophoresis, and directionally cloned into a modified pKS plasmid vector cleaved with *Asc* I and *Bam* H1 (or an enzyme generating an end compatible for ligation to ends generated by enzyme 'X'). In most cases, the desired restriction fragment excised from the polyacrylamide gel is still contaminated with multiple other species, and the abundance of these contaminating species may increase during the PCR reaction required to produce double-stranded DNA suitable for cloning. As a consequence, the identification of the bacterial colonies carrying the desired recombinant plasmids requires further screening. This is accomplished by the digestion of cDNA inserts obtained from multiple individual recombinant colonies with a mixture of several frequently cutting endonucleases. In most (but not all) cases, the wanted inserts harbour the restriction fragments that are most abundant in the PCR product before cloning.

3.3.1.3 Typical results and examples

Originally, ADDER was used for the isolation of circadian mRNA sequences (figure 3.49), but since then this technique was applied to other biological problems (figure 3.50).

Figure 3.49 Identification of mRNAs with circadian accumulation in mouse liver. Aliquots of double-stranded cDNA libraries were amplified with the display primers indicated on top of the figure and size-fractionated on a five percent denaturing urea/polyacrylamide gel alongside size markers (positions indicated on the left of the panel). The day times at which the animals were sacrificed to prepare liver RNA are given on top of the panel. The positions of cDNA fragments that were cloned and sequenced are indicated by arrows. All of the mRNAs complementary to these cDNAs showed circadian accumulation profiles, when examined by Northern blot hybridisation or

ribonuclease protection assays. The cDNA fragment indicated by a white arrow-head corresponds to a transcript that is present at about seven to eight copies per cell (5×10^{-6} of the mRNA population) (Kornmann *et al.*, 2001). Several additional fragments from potentially circadian mRNAs can be seen in the auto-radiograph. However, these fragments have not yet been cloned and the corresponding mRNAs have thus not been examined for cyclic accumulation by Northern blot or RNase protection assays (see Kornmann *et al.*, 2001, and figure 3.50). The figure is reproduced from Kornmann *et al.* (2001) with permission from *Nucleic Acids Research*.



The hunt for transcripts with circadian accumulation in the liver (figure 3.49) has revealed cDNA fragments for about 300 candidate genes (Kornmann *et al.*, 2001), of which approximately 60 have been cloned and sequenced. The careful abundance determination for one of the cycling transcripts has revealed that ADDER can reveal mRNA species contributing a fraction of less than 5×10^{-6} to the total population of polyadenylated mRNAs.

In order to compare the efficacies of ADDER display and Affymetrix® (Santa Clara, San Diego, USA) oligonucleotide microarray hybridisation in the identification of circadian genes, whole cell murine liver RNAs harvested at 3 a.m. and 3 p.m. were subjected in duplicate to microarray analysis. These hybridisation experiments were performed in collaboration with Deborah Keenan and Fiona

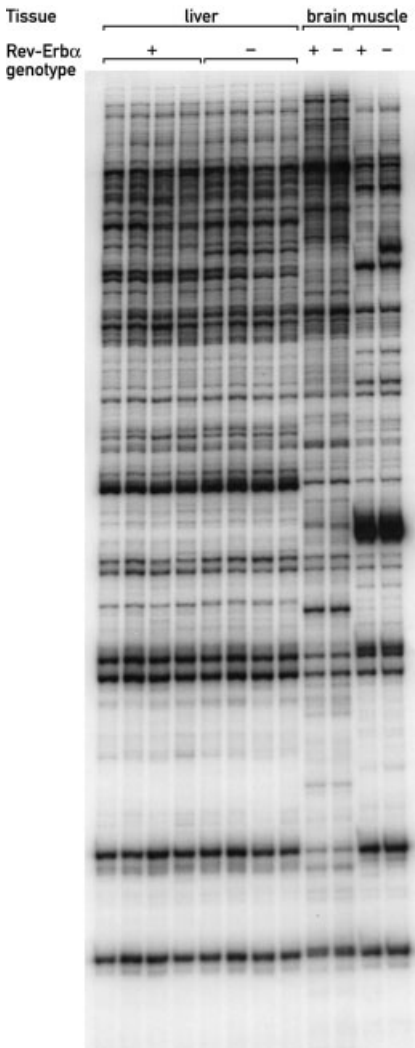


Figure 3.50 Identification of Rev-Erb α target genes in three tissues of fed and starved mice. Rev-Erb α is a nuclear orphan receptor that represses the transcription of its target genes when bound to DNA recognition within *cis*-acting regulatory regions. Some target genes are only repressed by Rev-Erb α when animals are starved (Nicolas Preitner, Francesca Damiola, and Ueli Schibler; unpublished observation). In order to identify target genes of the transcription factor Rev-Erb α in liver, skeletal muscle (hind leg), and brain, mRNA populations from starved wild type (+/+) and Rev-Erb α knockout (-/-) mice were compared by amplification of double-stranded cDNA end restriction fragments (ADDER) differential display as described in figure 3.49. The auto-radiograph shows the display products of twelve RNA populations obtained with the primer pair 5'-CCA ACC GAT CCA-3' and 5'-AGC TTT TTT TTT GA-3'. In order to reduce individual differences, RNA pools harvested from multiple animals were employed, and four independent pools were used for the display of liver RNA. The pools were obtained by mixing equal quantities of whole cell RNAs from four mice in the case of liver, and eight animals in the case of brain and skeletal muscle. Note that one cDNA species (marked by an arrow) is present in higher amounts in the liver and muscle samples of Rev-Erb α -deficient mice. Hence, the corresponding mRNA is likely encoded by a gene whose transcription is repressed by Rev-Erb α . In brain, this gene does not appear to be regulated by Rev-Erb α .

Brew of Affymetrix® UK Ltd. (High Wycombe, United Kingdom), using Mu74Av2 arrays that contain oligonucleotide features for 12,488 mouse transcripts (unpublished data). This analysis revealed 43 transcripts showing two-fold and higher differences between the two populations isolated at different day times. The differentially expressed transcripts include several mRNAs with well-established circadian expression patterns (the transcription factor DBP (D site of albumin promoter-binding protein), Rev-Erb α (member 1 of group D of the nuclear receptor subfamily 1, NR1D1; also called thyroid hormone receptor α 1-like, THRAL), Rev-Erb β (member 2 of group D of the nuclear receptor subfamily 1, NR1D2), cryptochrome 1 (Cry1; also called photolyase 1, PHLL1), brain and muscle aryl hydrocarbon receptor nuclear translocator (ARNT)-like protein 1 (BMAL1; also called aryl hydrocarbon receptor nuclear translocator-like protein, ARNTL), and 17 of the 55 circadian transcripts identified by the ADDER procedure. The remaining 38 cyclic transcripts revealed by the ADDER screen were not represented on the Affymetrix® microarray (15 transcripts), did not produce hybridisation signals (five transcripts), were not expected to yield different signals at the two chosen time points (13 transcripts), or did not produce significantly different hybridisation signals (five transcripts).

Figure 3.50 shows an ADDER comparison of twelve different RNA preparations. In this experiment, ADDER has been used to screen for transcripts that are expressed differentially after starvation in tissues of mice with a different genetic background. This experiment (with 192 primer pairs) has revealed that the nuclear orphan receptor Rev-Erb α affects the transcription of many genes only in starved animals (Nicolas Preitner and Ueli Schibler, unpublished results).

3.3.1.4 Validation of differentially expressed species identified by ADDER

The validity of the results produced by ADDER was examined for all 55 cloned and sequenced cDNA candidates for circadian mRNAs. To this end, liver RNA samples harvested at four-hour intervals around the clock were examined by Northern blot hybridisation or ribonuclease protection assays for the circadian accumulation of the transcripts complementary to the cloned and sequenced candidate cDNAs. For 51 of the 55 cloned cDNA probes, a circadian expression with an amplitude of at least two-fold could be confirmed, while the remaining four probes yielded no detectable signal in these assays. Interestingly, one of these unconfirmed circadian transcripts also produced different signals in the Affymetrix® microarray experiment. At present, it is not known why this transcript could be detected by ADDER and Affymetrix® microarray analysis, but not by ribonuclease protection assays, which in general provides a higher sensitivity than ADDER or microarray hybridisation.

3.3.1.5 Comparison of ADDER with other procedures

3.3.1.5.1 Accuracy, cost, and labour-intensiveness

In the authors' hands, the ADDER procedure generated few false-positives, and these were usually the result of genetic polymorphisms. Obviously, a single nucleotide polymorphism (SNP) can either eliminate or generate a recognition site for the enzyme(s) used in the production of 3'-terminal cDNA fragments. As a consequence, in the ADDER display of transcriptomes from genetically heterogeneous sources bands may appear or disappear that are unrelated to the expression levels of the corresponding transcripts. ADDER has also revealed an unexpectedly high frequency of small insertions and deletions within the 3'-untranslated regions of mRNAs. Surprisingly, SNPs and length heterogeneities have been observed even in RNA samples from different individuals of the same mouse inbred strain (strain 129^{OLA}; Benoît Kornmann, unpublished observation). The problems associated with such genetic heterogeneity can be reduced by the parallel display of multiple samples and/or by using RNA samples pooled from multiple individuals (see figure 3.50). SNP heterozygosity is expected to affect ADDER and SAGE to a similar extent, since both involve the generation of double-stranded cDNA restriction fragments. In contrast, small insertions or deletions should perturb SAGE to a lesser degree, as this method only analyses short sequence tags. Neither SNPs nor small length polymorphisms should influence the outcome of DNA or oligonucleotide microarray experiments, irrespective of whether Affymetrix® or printed DNA arrays are used. Hence, microarray hybridisation experiments are expected to be considerably less vulnerable to artefacts caused by genetic heterogeneity than ADDER or SAGE.

Preliminary comparative experiments conducted in collaboration with Affymetrix® UK Ltd. have revealed a relatively high level of sensitivity and specificity for both techniques (see above). Thus, some transcripts contributing about 10^{-5} parts to the total mRNA populations were detectable by both procedures. However, the TATA box-binding protein (TBP) mRNA, which makes up about 2×10^{-6} parts of the total liver mRNA population (Schmidt & Schibler, 1995), did not yield a signal in the Affymetrix® oligonucleotide microarray, and it was not examined whether this transcript produces a visible band in the ADDER display. The two major advantages of ADDER over Affymetrix® microarray hybridisation are its independence from nucleotide sequence information and its low cost. Obviously, the profiling of transcriptomes by microarray hybridisation is only applicable for transcripts with known sequences. Moreover, in contrast to Affymetrix® microarray hybridisation assays, ADDER experiments are affordable for most academic research laboratories, even if large sample numbers have to be compared.

The obvious advantages of Affymetrix® microarray hybridisation over ADDER are the rapid acquisition of data and their automated interpretation with regard to sequence identity and signal strength. The insensitivity of this technique towards nucleotide polymorphisms adds another important benefit to this technique.

ADDER and SAGE also have their relative advantages and drawbacks when compared to each other. The major advantage of ADDER is its low cost, while the major benefit of SAGE is the direct generation of sequence data.

3.3.1.5.2 Completeness of transcript detection

Different causes can lead to losses in the detection of transcripts in ADDER, SAGE, and DNA microarray hybridisation. The major limitations in DNA microarray technologies are the difficulty of manufacturing DNA arrays containing probes for all genes and the generation of meaningful hybridisation signals for low abundance mRNAs. The production of a SAGE sequence tag relies on the presence of a *Nla* III restriction site within a given double-stranded cDNA, and transcripts not containing such a recognition sequence are lost from the population. Moreover, this technique necessitates the sequencing of millions of sequence tags, if the profiling of rare transcripts is aspired. ADDER also suffers from limitations that can lead to the loss of cDNA species. As with SAGE, transcripts whose cDNAs do not contain the chosen restriction site within the 3'-terminal mRNA moiety will get lost in the ADDER display. For any frequently cutting restriction site this leads to the elimination of about 20 percent of all transcripts (Kornmann *et al.*, 2001). The ability to detect cDNA species of the remaining 80 percent of transcripts can further be jeopardised by several factors, including the co-migration of multiple transcripts during the electrophoretic display and unequal PCR amplification for different templates. In fact, PCR reactions follow highly complex kinetics. During the initial exponential phase, the number of amplified molecules follows the equation

$$N_x = N_0 [1 + E]^x,$$

where N_x is the number of molecules after x PCR cycles, N_0 is the number of molecules before amplification, and E is the efficiency (the value of E is between 0 and 1) of producing a copy from a given DNA template. Small variations in E can result in large differences in the ratio of N_x/N_0 of these templates after a certain number of PCR amplification cycles. Moreover, as shown in figure 3.51, E decreases during the PCR reactions and eventually approaches zero. This progressive reduction in E can be caused by the exhaustion of substrates, a decrease in *Taq* DNA polymerase activity, and the competition between templates and primers in their annealing with the complementary strand, once the template has reached a critical concentration (Heid *et al.*, 1996). If E is small and/or diminishes rapidly during the PCR amplification of a given template, this template is likely to get lost from the population.

3.3.1.6 Perspectives

ADDER is a sensitive and inexpensive approach towards the profiling of complex transcriptomes, in particular if many samples are to be compared. Nevertheless, several aspects of this method could be significantly improved. These include

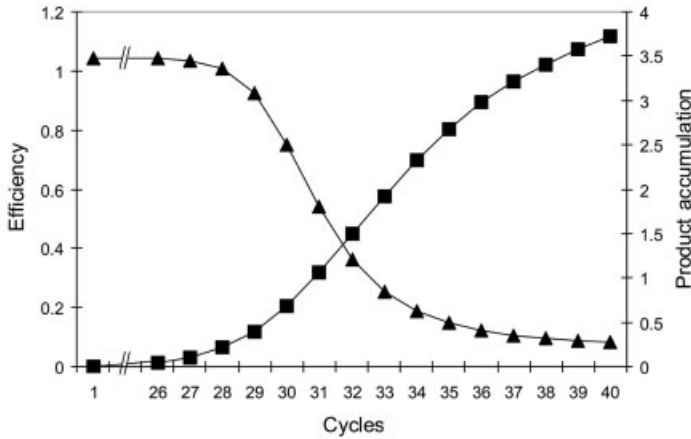


Figure 3.51 The efficiency of PCR amplification decreases during the reaction. To examine changes in the efficiency E during a PCR reaction, a curve of cDNA amplification (murine cryptochrome 1, Cry1, or photolyase 1; PHLL1) generated by TaqMan® real-time RT-PCR was analysed. The accumulation of the PCR product was measured as fluorescence and the values are plotted in arbitrary unit. The efficiency was empirically determined by using the equation:

$$E = \left[\frac{\text{accumulation}_N}{\text{accumulation}_{N-1}} \right] - 1$$

for individual PCR cycles N yielding fluorescence values higher than the background. An empirical best-fit equation was then used to obtain an idealised curve for the efficiency. The extrapolation of E for cycle 1 yielded a value of 1.05. In reality, E cannot exceed 1. The shape of the curve $E = f(\text{number of PCR cycles})$ can change for different amplicons.

the completeness of transcript representation, the labour-intensiveness of data acquisition, and the identification of transcripts whose sequences are already present in databases. In future attempts, remedies for these limitations will be sought by performing ADDER displays in parallel with multiple enzymes, and by automating both the data acquisition and its analysis.

With currently available pipetting robotics and automated capillary electrophoresis, it should be possible to assemble and display thousands of display reactions in a reasonably short time span. This should permit the execution of ADDER displays with multiple restriction enzymes in parallel and hence the reduction of cDNA loss. If ADDER libraries were prepared with four different restriction nucleases, the loss imposed by statistical probability should be only 0.16 percent $[(0.2)^4 \times 100 \text{ percent}]$. The risk of losing a given cDNA species due to unequal PCR amplification efficiencies should also be greatly reduced, since one of the four chosen restriction enzymes would be likely to generate a short and thus readily amplifiable fragment. Likewise, genetic micro-heterogeneity should no longer pose major problems, as SNPs are unlikely to occur in more than one restriction site, and as the probability of encountering small deletions and/or insertions diminishes with fragment size. Most importantly, however, the sequence information generated by using ADDER display libraries obtained by using multiple re-

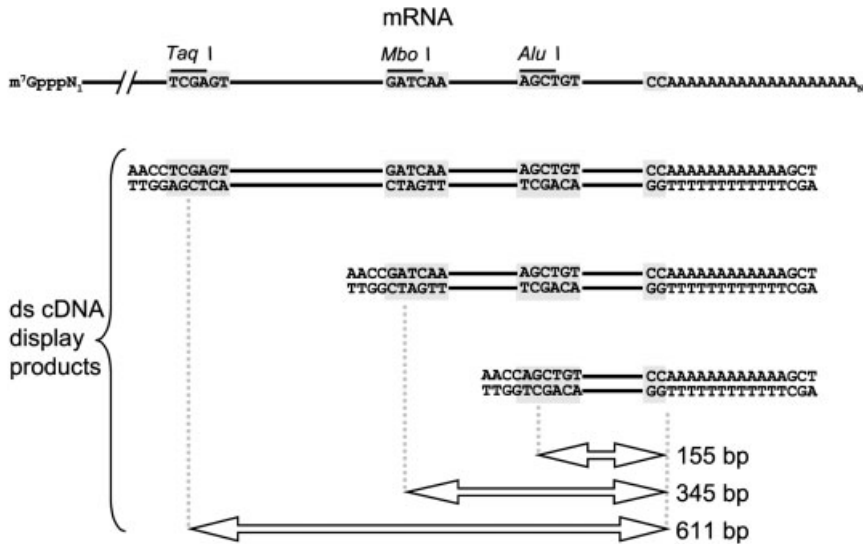


Figure 3.52 Sequence information available for the identification of amplification of double-stranded cDNA end restriction fragments (ADDER) cDNA fragments. The following hypothetical scenario was assumed: Two mRNA populations are displayed in parallel using ADDER cDNA libraries obtained with the four restriction enzymes *Alu* I (recognition sequence: 5'-AGCT-3'), *Mbo* I (recognition sequence: 5'-GATC-3'), *Rsa* I (recognition sequence: 5'-GTAC-3'), and *Taq* I (recognition sequence: 5'-TCGA-3'). One transcript that is more abundant in one population than in the other (labelled as mRNA in the cartoon) contains recognition sequences for three of the four enzymes (*Alu* I, *Mbo* I, and *Taq* I) at the indicated positions upstream of the polyadenylation sites, but no recognition sequence for *Rsa* I (within the 3'-terminal moiety). Using appropriate size markers on the display gel, the

length of the cDNA fragments can be estimated with an accuracy of $\pm$ five percent. The PCR reactions with the forward display primer 5'-AGC(A)₁₂CC-3' and the indicated backward primers should yield fragments of about 170 to 190 base pairs for the *Alu* I library, 350 to 385 base pairs for the *Mbo* I library, and 600 to 660 base pairs for the *Taq* I library, and these fragments should be more prominent in one of the two compared libraries. Even if each of these primer pairs produced multiple fragments corresponding to other differentially expressed mRNAs, it should still be possible to identify the hypothetical mRNAs shown here in sequence databases, if its sequence is known. Efforts to establish the adequate software are underway. The following abbreviations are used in this figure: bp, base pairs; ds, double-stranded.

restriction enzymes should establish a unique characteristic fingerprint for most differentially expressed transcripts (figure 3.52). By developing appropriate computer programs, it may be feasible to assign such fingerprints unambiguously to known mRNA sequences stored in public databases. In recent experiments, ADDER libraries produced with the two frequently cutting restriction enzymes (*Mbo* I and *Rsa* I) in parallel were successfully displayed, and the list of restriction endonucleases used in establishing display libraries may be further extended at will. Therefore, the conversion of ADDER into an automated high-throughput technology should be achievable in the near future.

3.3.2

ATAC-PCR – Adapter-tagged competitive PCR**3.3.2.1 Introduction**

The concept of using PCR amplification for mRNA quantification dates back to the late 1980s (Becker-André & Hahlbrock, 1989; Innis *et al.*, 1990). Ten years later, it continues to be a complicated procedure. The major challenge is constructing the precise standard curve required for accurate quantification. One of the most sophisticated among several techniques is real-time RT-PCR (Heid *et al.*, 1996). Once experimental conditions are determined, this method allows high-throughput analysis of single genes. However, it also requires a sensitive calibration experiment for each gene, which renders it impractical for handling a large number of genes.

Adapter-tagged competitive PCR (ATAC-PCR), an advanced form of quantitative competitive PCR (Kato, 1997), is a more recent technique that overcomes the above limitations. Reactions are carried out in a single tube as in standard PCR, permitting analysis of a large number of samples. In this section, the underlying principles, characteristics and large-scale applications of ATAC-PCR are described.

3.3.2.2 Principle and basics**3.3.2.2.1 The principle of ATAC-PCR**

In studying gene expression, the absolute amount of expression of a particular gene is often less important than relative expression level. ATAC-PCR was designed to measure relative expression levels of a particular mRNA among various samples. The method was originally developed for use with two RNA samples and two synthetic adapters. Figure 3.53 schematically displays the experimental steps of the ATAC-PCR procedure: (i) Double-stranded cDNA is prepared using a biotin-labelled oligo(dT) primer and standard methods. (ii) The cDNA is then cut with a restriction enzyme. (iii) Cohesive end adapters are added with T4 DNA ligase to the restricted cDNA. The adapters share a common end sequence, but each contains a spacer region of different length. (iv) Equal amounts of each ligated cDNA sample are mixed together. (v) cDNA is then recovered with streptavidin-coated paramagnetic beads, and (vi) amplified using an adapter primer and a gene-specific primer. The adapter primer shares the same sequence as the adapter's common region. (vii) Amplified products are separated by denaturing polyacrylamide gel electrophoresis. The products arising from each cDNA sample are separated due to the variable length of the spacer region. The relative expression levels in the original mRNA samples can then be deduced from the ratio of amplified products.

The original method with two adapters used titration experiments for accurate quantification and required several reactions for a single quantification. Further experiments revealed that up to seven different adapters can be used successfully. In this multiple adapter set, three adapters are designated for different amounts of a

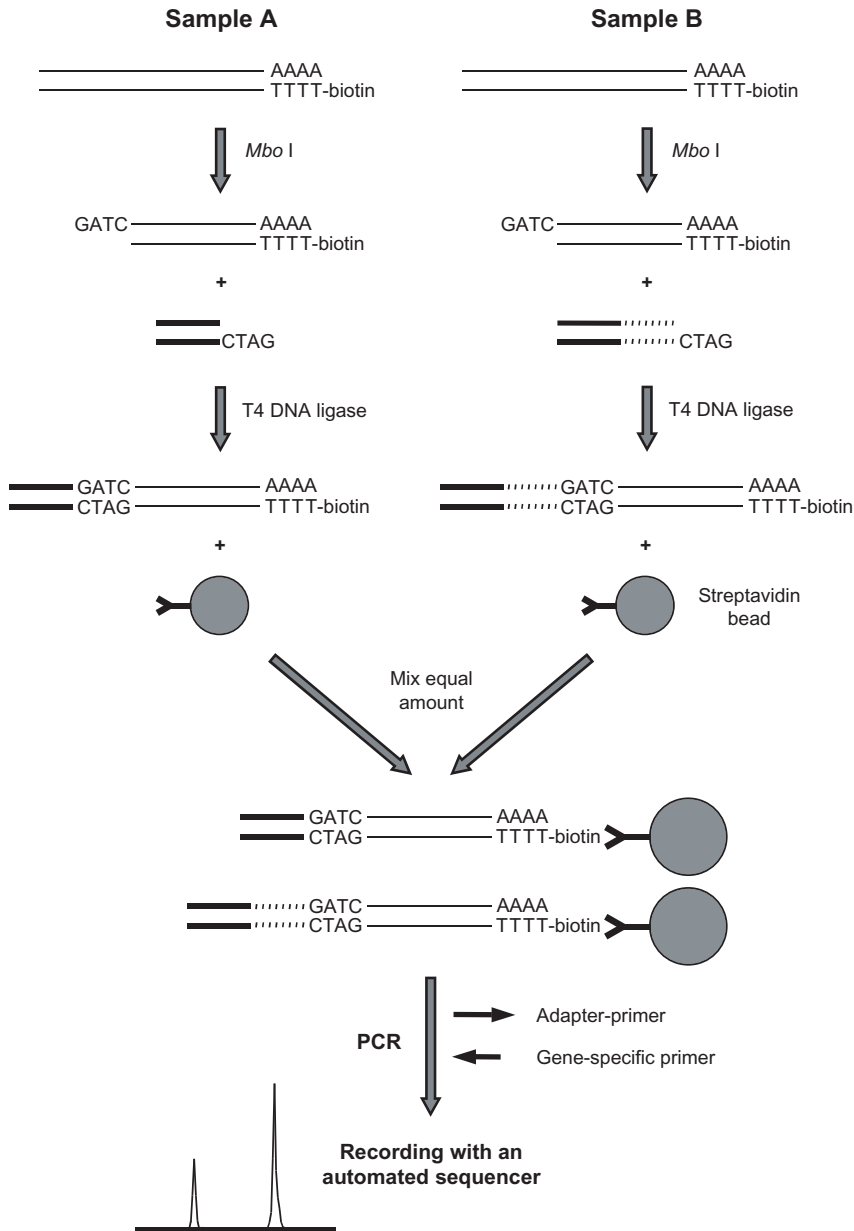


Figure 3.53 Schematic representation of the adapter-tagged competitive PCR (ATAC-PCR) using *Mbo* I as a restriction enzyme. Details are described in the text.

control cDNA, while the others are used to sample cDNAs (Matoba *et al.*, 2000a). In figure 3.54, one, three, and ten portions of the cDNA sample (with different adapters) are included in each PCR reaction. With the control cDNAs, a calibration curve is constructed, and accurate measurements of sample RNA expression levels can subsequently be performed. This advanced version of the technique has enabled accurate quantification of multiple samples in a single-tube reaction.

It should be noted that sample binding to paramagnetic beads is an essential step. The absolute amount of original cDNA template is very small, and non-specific template binding to the tube wall cannot be neglected. To overcome this problem, templates are allowed to bind to streptavidin-coated paramagnetic beads before removal from the tube. Since the total bead surface area is much larger than that of the tube wall, non-specific binding to the wall is likely to be negligible. Thus, quantitative template recovery is achieved. A variant of ATAC-PCR, iAFLP (introduced amplified fragment length polymorphism) (Kawamoto *et al.*, 1999), uses a vector primer as a carrier instead of beads.

Because ATAC-PCR uses only one specific PCR primer for each gene, the size of amplified products must be monitored in order to ensure product identity. Use of an automatic DNA sequencer is preferred for simultaneous quantification and size determination of a large number of samples. For the cDNA digestion, *Mbo* I, *Taq* I, *Hap* II and *Nla* III have been confirmed as usable. In practice, any restriction enzymes producing cohesive ends can be used.

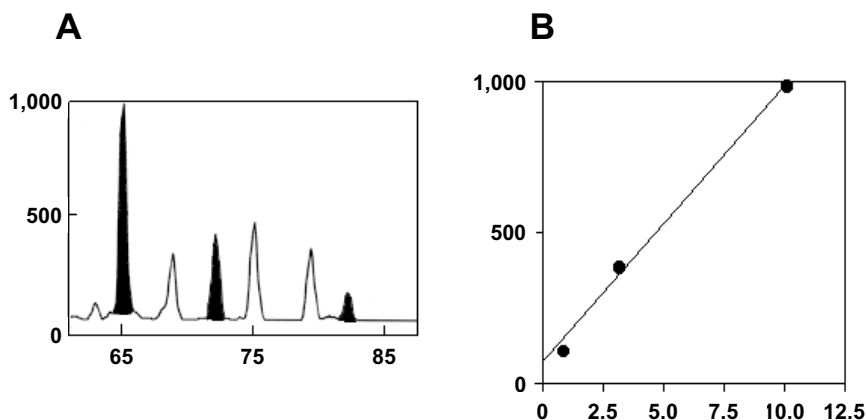


Figure 3.54 Example of adapter-tagged competitive PCR (ATAC-PCR). The gene is an ATP synthase-coupling factor. **A)** Electropherogram of ATAC-PCR. Black peaks depict the internal controls; three different amounts of cDNA made from whole brain (minus cerebellum) of an adult mouse (referred to as whole adult brain hereafter). Specifically, the peaks are (from left to right) ten, three, and one portions of whole adult brain cDNA. White peaks are of cerebellar samples at different stages of devel-

opment: from left to right, four-day cerebellum, twelve-day cerebellum, and six-week cerebellum. The vertical axis depicts the fluorescence intensity obtained by ABI® 377 sequencer (Applied Biosystems, Weiterstadt, Germany); while the horizontal axis shows fragment size. **B)** Calibration curve made with the internal controls in figure 3.54.A. Vertical axis, the fluorescence intensity; horizontal axis, expression level relative to transcript in whole adult brain.

3.3.2.2.2 Characteristics of the technique

The current version of ATAC-PCR, which has been routinely used in several laboratories, is capable of testing a large number of genes while maintaining a high degree of accuracy. Alternative techniques such as real-time RT-PCR set their calibration curves using serially diluted standard DNA. While these methods have similar accuracy across a wide range of concentrations, in ATAC-PCR the precise calibration curve is set at physiological concentrations because it uses native RNA as a standard. Thus, its accuracy is highest around the physiological concentrations of interest. A thorough comparison of both techniques has revealed many minute differences detected by ATAC-PCR that could not be detected by real-time RT-PCR.

The other distinct advantage of ATAC-PCR is its relative insensitivity to RNA degradation. Because only the 3'-end of mRNA is monitored, the technique is not hindered by template degradation. This feature makes ATAC-PCR an excellent choice for analysing clinical samples, which are not necessarily taken under ideal conditions.

A single assay requires no more than one nanogram of total RNA, with a detection limit likely to be one in 1,000,000. To date, this technique has been found to be successful on about 70 percent of expressed genes in mammalian tissues tested. This success rate is similar to that of real-time RT-PCR.

3.3.2.2.3 Large-scale analysis

In addition to facilitating routine laboratory practices, ATAC-PCR offers an entirely new method of gene expression profiling. Since ATAC-PCR shares several features with DNA sequencing, large-scale application is not difficult for laboratories equipped to handle large-scale DNA sequencing.

This technique is most useful when combined with EST sequencing (Matoba *et al.*, 2000b). First, a 3'-directed cDNA library is constructed using RNA purified from the target tissue, and a large number of clones (4,000 to 10,000) are subjected to single-pass sequencing. Use of 3'-directed cDNA libraries is well-suited for ATAC-PCR, enabling sequence reads from a restriction enzyme site. Expressed sequence tag (EST) sequencing has two major advantages. One is that selected genes are guaranteed to be expressed in the target tissue. Designing gene-specific primers exclusively for selected genes reduce labour and project cost. The other advantage is that these sequences guarantee accuracy of the 3'-end of the transcripts. In databases found in public domain, mRNA or cDNA sequences often lack correct 3'-end sequences. In addition, differential splicing is common, and the 3'-ends in target tissues are not necessarily the same as those in database sequences. For these reasons, ATAC-PCR combined with EST sequencing results in a higher success rate.

The entire process requires facilities capable of large-scale DNA sequencing. In particular, access to automatic DNA sequencers with fragment analysis software is essential. Here, ABI® Prism 3100 or 3700 DNA analysers from Applied Biosystems GmbH (Weiterstadt, Germany) are the instruments of choice.

3.3.2.3 Examples

To date, ATAC-PCR has been used for gene expression profiling (*i.e.*, analyses of a large number of genes) in a number of systems. Examples include postnatal development of the cerebellum (Matoba *et al.*, 2000c) and dentate gyrus in the mouse (Saito *et al.*, 2002), as well as apoptosis in a cultured cell line (Wyttenbach *et al.*, 2000) and breast cancer in humans (Iwao *et al.*, 2002). Here, the work on postnatal cerebellar development is discussed.

In rodents, development of the cerebellar cortex begins just after birth. Granule cell proliferation peaks in the first week, while cell migration and axon elongation peak in the second week (Altman & Bayer, 1996). Morphological changes are completed in the third week, and synaptic maturation occurs slowly until about six weeks (Altman & Bayer, 1996). This process was analysed in mouse through gene expression profiling with ATAC-PCR.

In order to define the system in terms of relevant genes, 3'-end-directed cDNA libraries were first constructed using RNA purified from mouse cerebellum at four days, twelve days and three weeks. More than 15,000 clones were sequenced, and 7,728 unique sequences were obtained (Matoba *et al.*, 2000b). Next, more than 2,500 genes were selected for primer design, after prioritising on the basis of both abundance and knowledge of gene function. The ATAC-PCR assay was performed with RNA purified from mouse cerebellum at two days, four days, eight days, twelve days, three weeks and six weeks. RNA purified from the cerebellum at six weeks was used as a control. The sample was assayed at least twice at each time point using different calibrations. Ultimately, complete expression data for 1,869 genes were acquired.

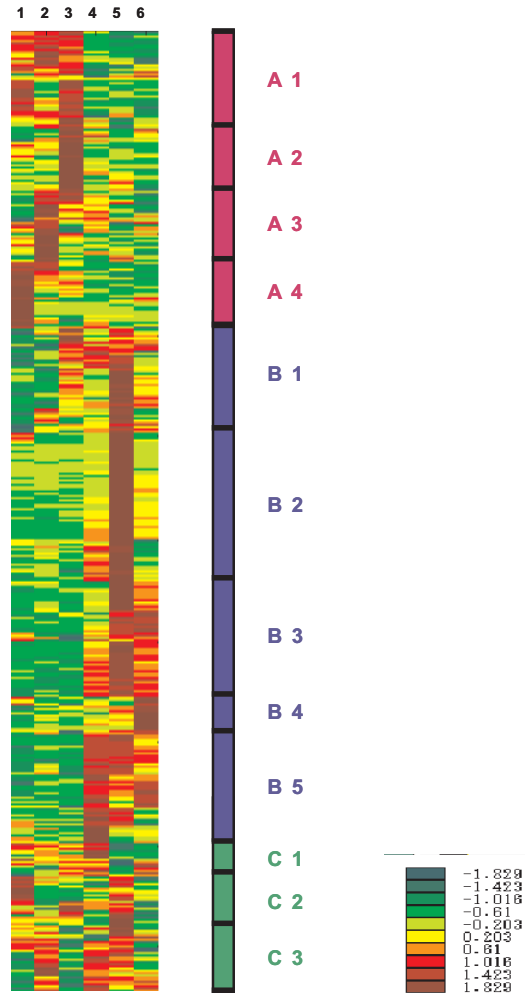
After processing the data, the results were organised into hierarchical clusters for analysis. The Wards method was used with/for standardised data as the clustering algorithm; genes were classified according to their developmental expression patterns. Clustering was truncated at twelve cluster levels, a schematic representation of which is shown in figure 3.55. These twelve clusters were grouped into three major groups: One characterised by elevated gene expression at an earlier developmental stage (group A); one characterised by elevated gene expression at later developmental stages (group B); and one characterised by miscellaneous expression patterns (group C).

Gene expression patterns were then tried to be correlated with gene function. Comparable studies on budding yeast have used the functional classification of the Munich Information Center for Protein Sequences (MIPS, Forschungszentrum für Umwelt und Gesundheit, GmbH, Neuherberg, Germany) (Tavazoie *et al.*, 1999). This database is useful for studying cellular functions, but less well-suited for analysis of the mammalian nervous system.

The in-house EST collection (Brain EST Database, BED) of the Nara Institute of Science and Technology (Nara, Japan) contains more than 1,600 known mammalian genes, corresponding to about one-fourth of all the known mouse genes (UniGene Build 85). Each gene is labelled with up to four keywords representing unique functions. The total number of keywords used is about 100. The keyword list is available at <http://love2.aist-nara.ac.jp>. In total, 1,053 known genes were as-

Figure 3.55 Cluster analysis of 1,869 genes using their expression patterns.

The 1,869 genes are aligned vertically: the data matrix is standardised to z-score (*i.e.*, converted to zero mean and unit variance) and schematically represented. Each row shows the expression pattern of each gene, and the columns represent time points after birth: two days after birth (1), four days (2), eight days (3), twelve days (4), three weeks (5), and six weeks (6). Expression levels are indicated by colour, with the scale shown at bottom right: dark red, yellow and dark green represent high, middle and low expression levels, respectively. Clustering is truncated at twelve cluster levels. A1 to A4 are characterised by elevated expression at the early stages (from two days to eight days) and decline; five clusters, B1 to B5, are characterised by low expression at early stages followed by elevated expression at later stages (from twelve days to six weeks); three clusters, C1 to C3, are characterised by other complicated patterns.



sayed and the correlation between gene function and expression pattern using keyword distribution was examined. The keywords 'cerebellum-specific' and 'cerebellum-dominant' are based on the cerebellum:cerebrum expression level ratio: 'cerebellum-specific' refers to those genes with a ratio greater than 20, 'cerebellum-dominant' to those with a ratio between ten and 20. From the three-group classification (groups A, B and C), eleven particularly informative keywords were selected (figure 3.56). The expression breakdown clearly indicates that ribosomal proteins and cancer-related genes are highly expressed in the early stages of development. Genes related to brain function are highly expressed at a later stage of development and/or in the adult. Genes involved in carbohydrate metabolism and RNA synthesis are also elevated at a later developmental stage. The re-

	A group	B group	C group	Gene number
Cancer-related	10	4	1	15
Ribosomal protein	35	5	14	54
RNA synthesis	0	11	1	12
Carbohydrate metabolism	3	15	0	18
Brain	33	104	15	152
Ion channel & transporter	2	17	1	20
Synapse component	1	14	2	17
Neurotransmitter receptor	1	9	0	10
Oligodendroglia	0	7	1	8
Cerebellum-dominant	8	55	4	67
Intracellular ion transporter	1	7	4	12
Total gene number	584	997	288	1,869

Figure 3.56 Functional categories active in specific groups of gene expression patterns during mouse cerebellar development. Each box represents either the number of genes belonging to each group or a functional category (keyword). Dark shading indicates statistically significant expression activity.

sults agree well with the anatomical and physiological characteristics of each developmental stage.

3.3.2.4 Discussion

As a routine laboratory practice, ATAC-PCR offers the most accurate quantification of relative gene expression. It is most readily usable with an automatic DNA sequencer, but can also be performed using conventional polyacrylamide gel electrophoresis with fluorescent dyes or radioisotopes. The major disadvantage of ATAC-PCR is the inherent difficulty in quantifying absolute levels of gene expression. Other popular techniques such as Northern hybridisation offer some information on absolute expression levels. Although using artificial control DNA at specified concentrations would enable absolute quantification, this is too tedious experimentally for practical laboratory use. Template preparation, which includes cDNA synthesis, restriction enzyme digestion and adapter ligation, is a complex process that might be too difficult for laboratories not accustomed to molecular biology. Development of self-contained kits should enable use of ATAC-PCR in a broader range of laboratories.

Large-scale application of ATAC-PCR has several characteristics in common with DNA microarrays (Schena *et al.*, 1995), currently the most popular technique for gene expression profiling. Both techniques require clones or sequence information of expressed genes to measure the expression level of each gene. In microarrays,

background hybridisation is indistinguishable from genuine hybridisation, while quantification is not necessarily accurate. By contrast, the quantification made with ATAC-PCR is guaranteed by internal standards. An obvious advantage of ATAC-PCR over microarrays is accuracy, as the resulting data is highly quantitative. However, the major disadvantage of ATAC-PCR is processivity. The rate-limiting step for both techniques (not including oligonucleotide microarrays) is PCR. From a single PCR reaction, ATAC-PCR can yield data for five samples. By contrast, a single PCR reaction with a DNA microarray can yield data for more than 100 samples. The running costs of both techniques are similar.

In conclusion, ATAC-PCR is one of the best techniques for cases that require accurate measurement of expression levels, from just a few to thousands of genes. As an expression profiling technique, it offers a useful alternative to DNA microarrays.

3.3.3

cDNA-AFLP – cDNA-based amplified-fragment length polymorphism fingerprinting

3.3.3.1 Introduction

Amplified fragment length polymorphism (AFLP) is a technique for the visualisation of genomic DNA polymorphisms in genomes. The method is based on restriction enzyme digestion with two enzymes differing in the frequency of recognition sites within a given genome, followed by the ligation of anchors (also termed adapters) on the sticky ends left by restriction enzyme digestion. PCR amplification, using primers corresponding to the anchors with one or more additional nucleotides extending beyond the restriction recognition site and into the target fragments, is used to create the final fingerprint. Initially, the method was used to construct genetic linkage maps in segregating populations and to identify individual organisms with unknown genotypes.

As a further extension of the application, the method was applied to cDNA with the aim of visualising gene expression rather than detecting DNA polymorphism (cDNA-based amplified-fragment length polymorphism fingerprinting, cDNA-AFLP). In this method, mRNA is isolated, for example, from different tissues or various developmental stages and cDNA is synthesised. The cDNA is then processed in the same way as in genomic AFLP (Vos *et al.*, 1995) to prepare a template that is amplified to generate an RNA fingerprint. The intensity of individual signals in the fingerprint is then taken as a measure of the strength of expression of a gene corresponding to the band or transcript-derived fragment (TDF).

The method of cDNA-AFLP can be adapted and tuned to suit a wide range of systems. Several successful applications have been presented in a wide range of biological systems, and some minor adaptations of the procedure have been published (Bachem *et al.*, 1998; Dellagi *et al.*, 2000; Hendriks *et al.*, 1991; Money *et al.*, 1996). As a combination of expression profiling and genetic mapping, cDNA-AFLP was recently applied to RNA isolated from mapping populations of diploid potato, and, thus, making it possible to show that a transcriptome map can be constructed in this way (Brugmans *et al.*, 2002).

The following contribution is written to discuss the numerous issues that arise in the application of cDNA-AFLP and explain the individual variants of the method. A detailed step-by-step protocol can be found on <http://www.dpw.wau.nl/pv/aflp/cDNA-AFLP%20Protocol.htm> and in the published literature (Bachem *et al.*, 1996 and 1998).

3.3.3.2 Principle and basics

The cDNA-AFLP method selectively displays transcript-derived restriction fragments as bands on polyacrylamide gels. Amplification is achieved by providing cDNA fragments with anchors at their termini serving as primer sites in subsequent PCR amplifications. From the genes likely to be expressed in an organ, tissue or cell type, only a fraction of the transcripts can be practically visualised in a single fingerprint by gel electrophoresis. To achieve a selective reduction, two strategies are taken. Firstly, the restriction enzyme used to digest the cDNA is chosen to limit the number of transcripts visualised. Secondly, by using different lengths of so-called 'selective bases' on the primer termini, a high level of tuning in the number of targeted fragments can be achieved per amplification.

As with all such techniques, the fidelity of the RNA fingerprint in representing the expression of genes depends on the cumulative influence of each step of the protocol. In the first instance, this is dependent on an efficient and reliable method of RNA extraction.

3.3.3.2.1 RNA preparation

The RNA isolation protocols used should be suitable for the organism and targeted to the tissue used in the experiment. Since cDNA-AFLP requires only small amounts of starting RNA, isolation can be attempted from very small tissue samples. Users report few problems in RNA isolation when using animal systems. Several plant tissues used in the authors' laboratory, however, have proved recalcitrant for high-quality RNA extraction. Frequently, polysaccharides and polyphenols tend to be a particular problem. Elimination of many such contaminants can frequently be achieved during the enrichment of poly(A)⁺ mRNA using paramagnetic beads (Bachem *et al.*, 1998).

3.3.3.2.1.1 Total RNA versus poly(A)⁺ mRNA enrichment

Although ribosomal RNA should not take part in cDNA synthesis when poly(dT) primers are used, results have shown that a purification of poly(A)⁺ mRNA greatly reduces the occurrence of rRNA-derived fragments in the RNA fingerprints (figure 3.57, step 2). A further advantage of poly(A)⁺ mRNA enrichment using poly(dT)₂₅V-coated magnetic beads (V is either adenin, cytosine or guamin, but not thymine) is also the virtual elimination of large molecular weight DNA frequently present as a contaminant in crude total RNA preparations. When using cDNA-AFLP fingerprinting for prokaryotes, other strategies for synthesising cDNA may be employed (Dellagi *et al.*, 2000). Several kits for poly(A)⁺ mRNA isolation are available and all appear to be suitable for cDNA-AFLP template produc-

tion. However, the authors use streptavidin-coated paramagnetic beads and attach biotinylated oligo(dT)₂₅V, where the V nucleotide anchors the cDNA at the 5'-end of the poly(A) tail. Due to the inevitable losses of RNA incurred during poly(A)⁺ mRNA purification, it may be advisable to use total RNA as input for cDNA synthesis where RNA quantity may be a limiting factor. In this case, it is important to include a DNase I treatment of the total RNA to avoid genomic DNA (gDNA)-derived bands in the fingerprint.

3.3.3.2.2 Solid support cDNA synthesis

Isolated poly(A)⁺ mRNA can be used with the addition of the oligo(dT)₂₅V as primer for reverse transcription. Complementary DNA can be synthesised while the poly(A)⁺ mRNA is still attached to the beads. This procedure has the advantage of eliminating the 3'-end of the cDNA after release of the 5'-end by restriction enzyme digestion (figure 3.57, step 4.A). In a further variation of the procedure (figure 3.57, step 4.B), cDNA synthesis on a solid support is used to eliminate secondary fragments from cDNAs with multiple restriction sites should this be required. In this variant of the method, however, a different oligonucleotide should be used for poly(A)⁺ mRNA capture, where a recognition site for a (rare) restriction enzyme is incorporated in order to release the desired fragment from the beads after elimination of 5'-fragments (see section above and figure 3.57, step 5).

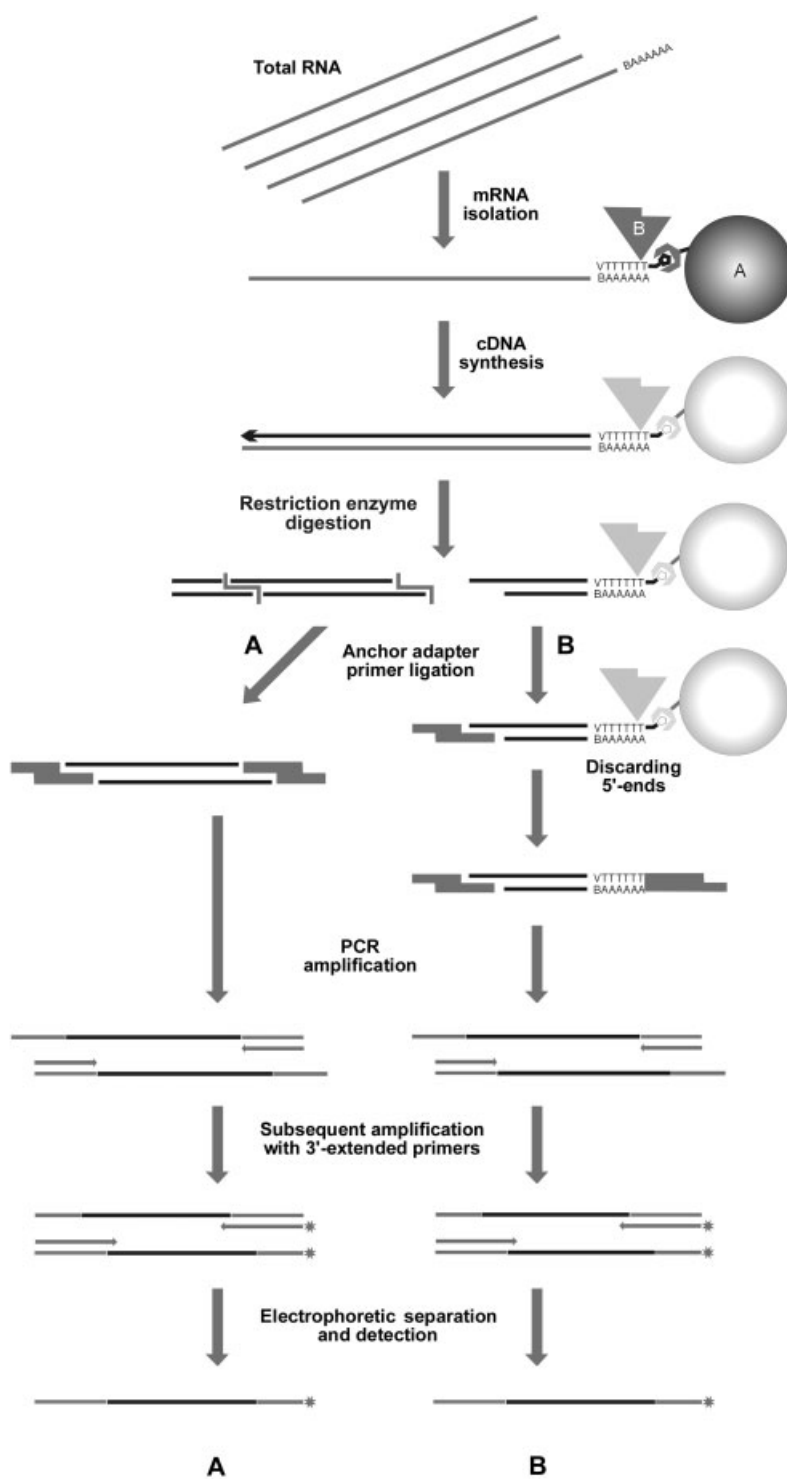
3.3.3.3 Template preparation

Template preparation from double-stranded cDNA can be divided into three stages: (i) restriction enzyme digestion, (ii) anchor ligation, and (iii) pre-amplification.

3.3.3.3.1 Restriction enzyme digestion and choice of enzymes

The digestion of cDNA has two main aims. The first is to provide sticky ends for the efficient ligation of anchors to the ends of restriction fragments. The second is to reduce the size of DNA fragments for separation using polyacrylamide gel electrophoresis (between 50 and 800 base pairs). This aim can be achieved in two consecutive steps. In the first step, the cDNA is digested with a rare cutting enzyme such as an enzyme with a six-nucleotide recognition sequence. The second enzyme is used to generate fragments of the desired size, and for this, a restriction enzyme recognising four nucleotides is usually used. Due to the differential frequency of restriction sites of the two enzymes in cDNA, the most fragments produced by double digestion carry sticky ends for the frequent cutting enzyme only. These fragments are, however, not visualised, since only the primer for the rare cutting enzyme site is labelled in the final PCR thereby revealing only fragments carrying either both sites or only sites for the rare cutter.

The selection of enzymes is a crucial point for optimising results since it will affect the number of different sequences represented in the fingerprint. Ideally, the rare cutter would cut every cDNA once. However, evaluation of 38 six-base cutting enzymes on a set of 191 full-length potato cDNAs shows that the highest achiev-



◀ **Figure 3.57 Diagrammatic representation of the cDNA-based amplified-fragment length polymorphism fingerprinting (cDNA-AFLP) method and its variants.** The AFLP procedure can be divided into the following steps: (i) Total RNA is isolated and (ii) enriched for poly(A)⁺ mRNA using an oligo(dT)₂₅V primer (V is adenine, cytosine and guanine, but not thymine) attached to paramagnetic beads (marked with an 'A') via a biotin/streptavidin linkage. Optionally, the oligo(dT)₂₅V primer may be engineered to contain a restriction enzyme recognition site (marked with a 'B') to facilitate display of a single band for every expressed transcript. (iii) RNA isolation is followed by double-strand cDNA synthesis. (iv) DNA digestion with two different restriction enzymes is carried out to produce restriction

fragments between 50 and 800 base pairs (A). Where a restriction site has been incorporated into the oligo(dT)₂₅V primer, the 5'-cDNA fragments are discarded before digestion with the second enzyme (B). (v) The restriction fragments receive anchors on their termini. (vi) PCR amplification using primers corresponding to the anchor sequences is used to bulk the targeted cDNA restriction fragments. (vii) In a subsequent round of amplification, a reduced set of restriction-fragments is amplified using cDNA specific 3'-extensions on the primers. (viii) Single-stranded DNA fragments (transcript-derived fragments, TDFs) are electrophoretically separated and visualised by auto-radiography, staining or fluorescence depending on the labelling method employed.

able digestion frequency for this organism is about 50 percent (table 3.6). This figure drops to around 33 percent when incomplete sequences obtained from an EST database are used for the analysis (data not shown). Since reverse transcription is frequently incomplete, giving rise to shortened cDNAs, the actual digestion frequency that can be experimentally achieved is probably somewhat below 50 percent. Several strategies have been considered to increase the hit-rate for the rare cutter enzyme without generating a large number of fragments per cDNA. One of those is to use more than one rare cutter generating sticking ends compatible with the same anchor (*e.g.*, *Bam* HI, *Bgl* II, *Bcl* I) or to use enzymes with a recognition sequence including one or more wobble bases. However, here, degenerate primers need to be used in the amplification steps to account for the variety of possible sequences.

The data shown in table 3.6 also show that there are large differences in the restriction frequency of the different enzymes. Because plant genomes are generally adenine/thymine-rich, it is not surprising that enzymes with adenine/thymine-rich recognition sites have, in general, a higher digestion frequency. However, this does not necessarily hold true for all organisms. The effectiveness of the restriction enzymes should therefore be determined in order to optimise the results. The distribution of the fragments generated is also likely to be dependent on the enzyme combination used. Since the complexity of transcribed sequences increases in the coding region, guanine/cytosine-rich recognition sites are likely to be more abundant around the middle of the cDNAs whereas those with adenine/thymine-rich sites are more likely to be located in the untranslated regions (UTRs).

When cDNA synthesis is carried out on a solid support, a number of additional options exist during digestion. The restriction fragments released from the beads can be collected and the 3'-ends of the cDNAs discarded, thereby reducing the complexity of the template. In the same way, the rare cutter can be used in a

Table 3.6 Frequency of restriction enzyme recognition sites in full-length cDNAs of *Solanum tuberosum* (in alphabetical order).

Enzyme	Recognition sequence	Number	Percentage
Aat II	G↓ACGT↓C	14	7.33
Aci I	C↓CG↑C	118	61.78
Acl I	AA↓CG↑TT	41	21.47
Afl II	C↓TTAA↑G	50	26.18
Age I	A↓CCGG↑T	28	14.66
Apa LI	G↓TGCA↑C	32	16.75
Ase I	AT↓TA↑AT	87	45.55
Asp 718I	G↓GTAC↑C	39	20.42
Asu II	TT↓CG↑AA	67	35.08
Avr II	C↓CTAG↑G	31	16.23
Bam HI	G↓GATC↑C	59	30.89
Bcl I	T↓GATC↑A	80	41.88
Bgl II	A↓GATC↑T	92	48.17
Bse PI	G↓CGCG↑C	8	4.19
Bsp 120I	G↓GGCC↑C	14	7.33
Bsp 1407I	T↓GTAC↑A	58	30.37
Bsp HI	T↓CATG↑A	75	39.27
Bsp LU11I	A↓CATG↑T	69	36.13
Cla I	AT↓CG↑AT	52	27.23
Eco RI	G↓AATT↑C	97	50.79
Hin dIII	A↓AGCT↑T	108	56.54
Hin P1I	G↓CG↑C	124	64.92
Hpa II	C↓CG↑G	128	67.02
Mae I	C↓TA↑G	179	93.72
Mae II	A↓CG↑T	141	73.82
Mbo I	↓GATC↑	188	98.43
Mfe I	C↓AATT↑G	95	49.74
Mlu I	A↓CGCG↑T	14	7.33
Mse I	T↓TA↑A	189	98.95
Nar I	GG↓CG↑CG	7	3.66
Nco I	C↓CATG↑G	79	41.36
Nde I	CA↓TA↑TG	59	30.89
Nhe I	G↓CTAG↑G	40	20.94
Pst I	C↑TGCA↓G	76	39.79
Ppu 10I	A↓TGCA↑C	64	33.51
Pvu I	CG↑AT↓CG	19	9.95
Sac I	G↑AGCT↓C	64	33.51
Sac II	CC↑GC↓GG	11	5.76
Sal I	G↓TCGA↑C	24	12.57
Spe I	A↓CTAG↑T	57	29.84
Sph I	G↑CATG↓C	44	23.04
Taq I	T↓CG↑A	176	92.15
Tsp EI	↓AATT↑	189	98.95
Xba I	T↓CTAG↑A	70	36.65
Xho I	C↓TCGA↑G	49	25.65
Xma I	C↓CCGG↑G	16	8.38
Xma III	C↓GGCC↑G	18	9.42

↓ site of top strand nick

↑ site of lower strand nick

first round of digestion and only the DNA retained on the beads exposed to the frequent cutting enzyme after discarding the fragments generated in the first digestion. This procedure will generate only one fragment for each cDNA that contains sites for both enzymes (Breyne & Zabeau, 2001).

By incorporating a restriction enzyme site in the primer used for cDNA synthesis and initial digestion with the frequent cutting enzyme, all 5'-fragments of the cDNAs can be discarded leaving a single fragment for every cDNA attached to the beads. This fragment can then be released from the beads by digestion with second (rare cutting) enzyme. Theoretically, this procedure also yields a single transcript-derived fragment for every cDNA available in the poly(A)⁺ fraction that has a site for the frequent cutting enzyme. It should be noted, however, that a 3'-bias is established in this variation of the method that is likely to produce fragments from 3'-UTRs. Due to the low level of conservation in such sequences between homologues, using this procedure in a species with a little sequence data available in the databases will make the direct identification of transcripts difficult by sequence similarity (Bachem *et al.*, 1996; Money *et al.*, 1996).

3.3.3.3.2 Anchor and primer design

The addition of anchors (frequently termed adapters) to the ends of the cDNA restriction fragments creates universal primers-annealing sites for subsequent PCR steps. The anchors should be designed to have single-stranded overhangs compatible with sticky ends produced by the restriction enzymes. The adjacent sequences should prevent reconstitution of the restriction sites after ligation of the anchors allowing restriction enzymes to remain active during ligation, thus preventing concatemerisation of fragments. Overall, the anchors should have a melting temperature/dissociation temperature (T_m/T_d) of around 50 °C – most importantly, the anchors for both sites should produce close to identical annealing temperatures for the primers used in the PCR. Furthermore, the anchors should not allow the formation of secondary structures and have an internal temperature profile that inhibits 3'-mismatching during PCR.

3.3.3.3.3 Pre-amplification

After ligation of the anchors to the ends of the cDNA-restriction fragments, the ligation mixture (primary template) may be used directly for a round of pre-amplification. This procedure uses primers that are complementary to the lower strands of the anchors and extend into the remains of the restriction sites in the cDNA. The PCR will produce copies of all available transcript-derived fragments and will produce a sufficient quantity of template (secondary template) for the subsequent selective fragment amplification (SFA). Results show that pre-amplification does not significantly alter the fingerprint when compared to the selective fragment amplification directly on the primary template (Bachem *et al.*, 1998). Thus, where a large amount of RNA starting material is available, pre-amplification may not be required. Repeated pre-amplification of a secondary template is not advised since this will result in a skewing in the representation of transcript-derived fragments towards those with higher amplification efficiency.

3.3.3.4 Fingerprint generation

For the production of the final fingerprint, primers are used that carry extensions at the 3'-end such that amplification occurs only where the corresponding bases are available in the cDNA. This reduces the number of bands visualised in the fingerprint. Since only a single fragment per transcript is required, the primer corresponding to the rare-cutting end of the fragment is labelled. The number of bands that can be accurately scored in a single lane is between 50 and 100. This target can be achieved by altering the number of selective bases at the ends of the primers for every biological system under investigation. In simple eukaryotes such as yeast and nematodes, one selective base at each end may be sufficient to achieve the required number of bands whereas higher eukaryotes require at least a two-base extension at each end.

In general, extending the length of an AFLP primer by one base will decrease the number of visualised bands about four-fold. In practice, this means that when comparing a fingerprint with a two base extension at both ends (+2, +2) with a fingerprint produced with +2 and the four possible +3 base extensions, most of the bands produced with the higher selectivity can be found in the (+2, +2) fingerprint. However, a number of additional bands will generally be amplified in the (+2, +3) fingerprint that appear to have no counterpart in the original (+2, +2) fingerprint. The same phenomenon will also hold for additional additions of selective bases. The basis for this phenomenon remains unclear (see discussion).

3.3.3.5 Band isolation and verification

Transcript-derived fragments of interest can be isolated from gels of fingerprints generated by either auto-radiography or silver staining by excising the relevant section from the polyacrylamide gel and by re-amplifying the DNA extracted from the slice using the same primers. When using fluorescent labelling in conjunction with automatic sequencing apparatus, the primer combinations yielding transcript-derived fragments of interest are re-run using radioactive labelling, and the bands are isolated as described. Several methods of DNA isolation from polyacrylamide gels have been described, however, electro-elution was found to be the most consistent method (Bachem *et al.*, 1998). After re-amplification of the DNA, fragments may be directly sequenced or first cloned into an appropriate vector and then sequenced. It is advised to sequence several independent clones to verify that a single homogenous band has been isolated. In order to further verify that the isolated transcript-derived fragments correspond to the observed bands, it is possible to re-run the RNA fingerprint using primers with longer extensions (+3) corresponding the sequence determined in the isolated fragment. The targeted transcript-derived fragment should be present with the same expression pattern in the repeated fingerprint.

3.3.3.6 Typical results

In the authors' laboratory, cDNA-AFLP fingerprinting was applied to the study of the potato tuber life-cycle. The life-cycle of the potato tuber can be subdivided into a number of developmental phases including organogenesis, tuber bulking, dormancy, sprouting and resource mobilisation. Although the potato tuber is a relatively simple organ consisting essentially of a modified underground stem, its life-cycle requires the coordinated control of a complex set of interlocked physiological processes and metabolic pathways. Such processes are likely to be regulated by a large number of genes. Using tissue culture-derived tissues, RNA samples were taken at regular intervals of this developmental process and the RNA finger-

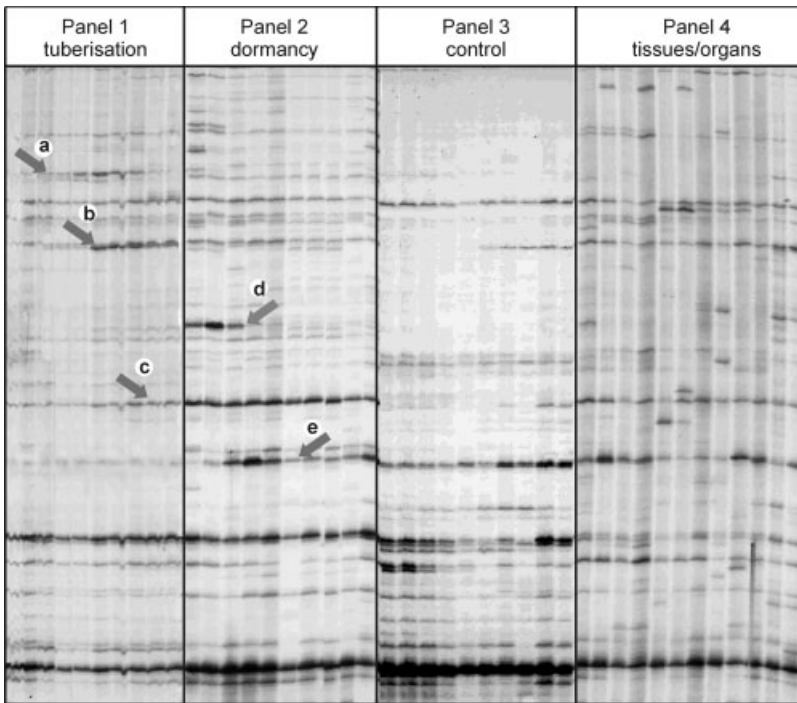


Figure 3.58 Example of cDNA-AFLP RNA fingerprinting applied to potato tuber development. Every lane of the two left-hand panels (tuberisation and dormancy) represents a time-point during the potato tuber life-cycle. Key events of tuber development such as stolon initiation (lanes 2 and 3, panel 1), tuber formation (lanes 4 and 5, panel 1), dormancy begin (lanes 1 and 2, panel 2), dormancy breakage (lanes 6 and 7, panel 2) and sprouting (lanes 7 and 8, panel 1) can be pinpointed on the gel. The intensity profiles of the bands show the expression profile of the corresponding gene during tuber development. Panel 3 (con-

trol) shows expression profiles of tissue explants grown on medium suppressing tuber formation (first ten days). Bands that differ in their expression profiles between panels 1 and 2 indicate a process specific expression pattern. Panel 4 shows an RNA fingerprint of a number of different tissues and/or organs from potato that give an indication of tissue specificity of expression. The arrows show bands that reveal time-points-specific expression: a) tuberisation-specific, b) tuber development-specific, d) dormancy initiation-specific, e) dormancy and sprouting-specific.

prints were displayed using all 256 possible +2 primer extensions. An example of one such fingerprint is shown in figure 3.58 (panel 1 and 2).

In general terms, all possible expression patterns were observed during tuber development (figure 3.58, see arrows). To test for expression specificity, a number of controls were incorporated. These include parallel RNA fingerprints of various tissues from the potato plant in order to show tissue specificity and fingerprints from axillary bud products produced on tuber-inducing medium containing the plant hormone gibberellic acid (GA₄) that suppresses the formation of tubers during first ten days (figure 3.58, panel 3). The screening programme has shown that a large number of genes change in expression during the first six days of tuber formation. Of these, the majority show an induced pattern (75 percent) with transient gene expression (ten percent) and transcriptional suppression (15 percent) being more rare. It is interesting to note that the majority of genes with temporal gene regulation also revealed a tissue-specific regulation pattern. Conversely, transcript-derived fragments that appeared to show constitutive expression were also expressed in all tested tissues (figure 3.58, panel 4). Gene expression dynamics are greatly reduced during dormancy (15 percent) and increase again towards sprouting (figure 3.58, panel 2) (Bachem *et al.*, 2000a).

To date, around 350 transcript-derived fragments have been identified, cloned and sequenced from the 20,000 visualised bands that show developmentally-specific expression patterns. Parallel screening of the controls has shown that the expression of a subset of these genes is highly process and tissue-specific. Complementary DNAs corresponding to seven transcript-derived fragments have been isolated and have been analysed in transgenic potato as antisense constructs for their phenotypic effects on potato plant and tuber development (Bachem *et al.*, 2000b and 2001).

3.3.3.7 Applications

The cDNA-AFLP method has been applied to diverse biological systems including bacteria (Dellagi *et al.*, 2000), fungi (Breyne & Zabeau, 2001; Money *et al.*, 1996), plants (Bachem *et al.*, 2000 and 2000b; Biezen *et al.*, 2000; Campalans *et al.*, 2001; Cooper, 2001; Daele *et al.*, 2000; Durrant *et al.*, 2000; Hirai *et al.*, 2000; Huang *et al.*, 2001), nematodes (Qin *et al.*, 2000) and mammals (Fukuda *et al.*, 1999). In these applications, changes of gene expression were visualised under a wide range of biotic and abiotic conditions. In all cases, process-specific alterations of gene expression were identified that led to the isolation of genes directly or indirectly involved in the process under investigation. The published data demonstrate the utility of applying cDNA-AFLP to biological questions involving induction or repression of gene expression. Due to the statistical nature of the method, it is ideally suited to systems where a moderate to large set of genes is likely modulated in expression.

In a further advancement of the technique, cDNA-AFLP was used in combination with traditional genetic analysis to develop a genome-wide transcript map (Brugmaus *et al.*, 2002). Using mapping populations of diploid potato and *Arabidopsis*

in combination with cDNA-AFLP, a large number of transcripts were mapped separately on both genomes. Polymorphic transcripts were shown to be useful as genetic markers and that they allow the construction of a linkage map. The resulting transcript map showed that the transcriptome-derived markers were evenly spread across all chromosomes, which, in contrast to genomic markers, are specifically targeted to transcriptionally active regions. The cDNA-AFLP markers used in the mapping are derived primarily from single nucleotide polymorphisms in transcripts, rather than from differences in expression regulation. This concludes that the higher potential of transcriptome markers as opposed to (anonymous) genomic markers will be useful for applications in genetic analyses, marker-assisted breeding and identification of candidate genes correlated with phenotypic markers.

3.3.3.8 Discussion

The cDNA-AFLP method has proven to be a robust and reliable method for monitoring the differential expression of genes in a wide range of biological systems. Along with other RNA fingerprinting methods, it has the great advantage of not requiring prior sequence information and affords the possibility of screening a large number of controls simultaneously with the treatments used for the experiments. Furthermore, utilising these features allows the unambiguous allocation of process specificity to the observed modulation of gene expression in virtually any organism under investigation. As opposed to other RNA fingerprinting methods, cDNA-AFLP allows the verification of band identity using the same technology adding selective base extensions on the primers after the sequence of a transcript-derived fragment is determined.

Due to the highly stringent conditions used during PCR (Bachem *et al.*, 1998), the rate of false-positives produced by mis-priming is very low. However, as with all analysis involving DNA retrieval from denaturing polyacrylamide gels, false-positives are generated by co-amplifying unrelated sequences that may either underlie the targeted band or co-migrate very closely with the chosen fragment. Unfortunately, there is no foolproof method of getting around this problem and detailed verification is required to confirm fragment identity and expression pattern.

From the authors' results and reconstruction experiments, it can be concluded that cDNA-AFLP is a very sensitive system – being able to detect gene expression of very rare messages. It has been suggested that this capacity may even be enhanced when using longer extensions on the primers by reducing competitive inhibition (Money *et al.*, 1996). The dynamic range of the system is limited by the detection system. Both radioactive and fluorescent labelling were used for detection. It was found that radioactive labelling together with the use of a PhosphorImager™ (Amersham Biosciences, Inc., Piscataway, New Jersey, USA) gives the highest dynamic response. The sensitivity to changes in expression depends largely on the number of repeats of treatments and controls – thus, a two-fold increase in expression should comfortably be within the range of the system as long as sufficient controls have been incorporated into the experiment.

The flexibility of cDNA-AFLP, in terms of the choice of enzyme combinations, anchor design and primer extension features, allows the system to be readily adapted to a wide variety of biological problems.

3.3.4

Competitive RT-PCR

3.3.4.1 Introduction

Becker-André & Hahlbrock (1989) were the first to propose the use of competitive RT-PCR assays employing a standard comprised of mutant RNA of the gene of interest. In this system, the RNA homologue shares sufficient sequence to be recognised by the reaction primers, but is modified either to alter the existence of a restriction site or to alter the length of the sequence intervening between the primer binding sites. Mutant internal standard RNA is added in a range of known quantities to multiple reactions each containing uniform amounts of the RNA preparation to be quantified. Gene expression is estimated by plotting the ratio of reaction products. The calculated concentration of competitor which would result in reaction products of equal abundance is equal to the unknown concentration of RNA. Numerous applications of this approach in varying modifications have been published (Volkenandt *et al.*, 1992).

As this technique gained wider application, there was a growing need to develop methods that provided accurate quantitative analysis of competitive RT-PCR reactions, but which limited the labour-intensive analysis of the products of multiple titration assays. The ideal approach would reduce post-reaction analysis to a single step and eliminate time-consuming activities such as digestion using restriction enzymes and reaction product gel preparation and analysis. It would also lend itself to automation and produce accurate and reproducible measurements of gene expression. For example, ion paired-reverse phase-high performance liquid chromatography (IP-RP-HPLC) was used to analyse these reactions. This technique satisfies all the requirements listed above and has revealed other aspects of competitive assays involving homologous templates which have been essential to assay validation, accuracy and simplification (Doris *et al.*, 1998; Hayward-Lester *et al.*, 1995 and 1996).

3.3.4.2 Principle and basics

PCR is a replicative technique in which DNA is detected because the original DNA in the sample is multiplied by numerous cycles of replication. With exponential increases occurring in repetitive cycles, small differences in the efficiency of amplification due to small variations in factors such as temperature, buffer composition, volume, or presence of contaminants means that comparisons of the amount of DNA amplified from one sample with that amplified from another are risky at best and may become meaningless if assay conditions vary by even a slight degree. As can be seen from the equation below, the effect of variations in PCR efficiency E can lead to dramatic variations in product yield.

The following equation describes the accumulation of reaction products in PCR. It predicts that small variations in PCR efficiency E can lead to dramatic variations in product yield.

$$N_x = N_0 [1 + E]^x$$

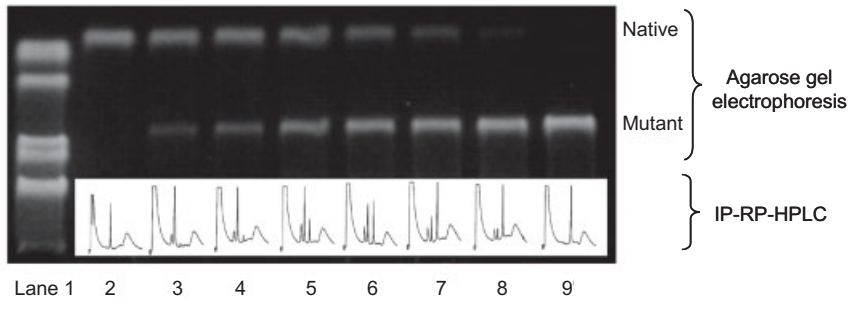
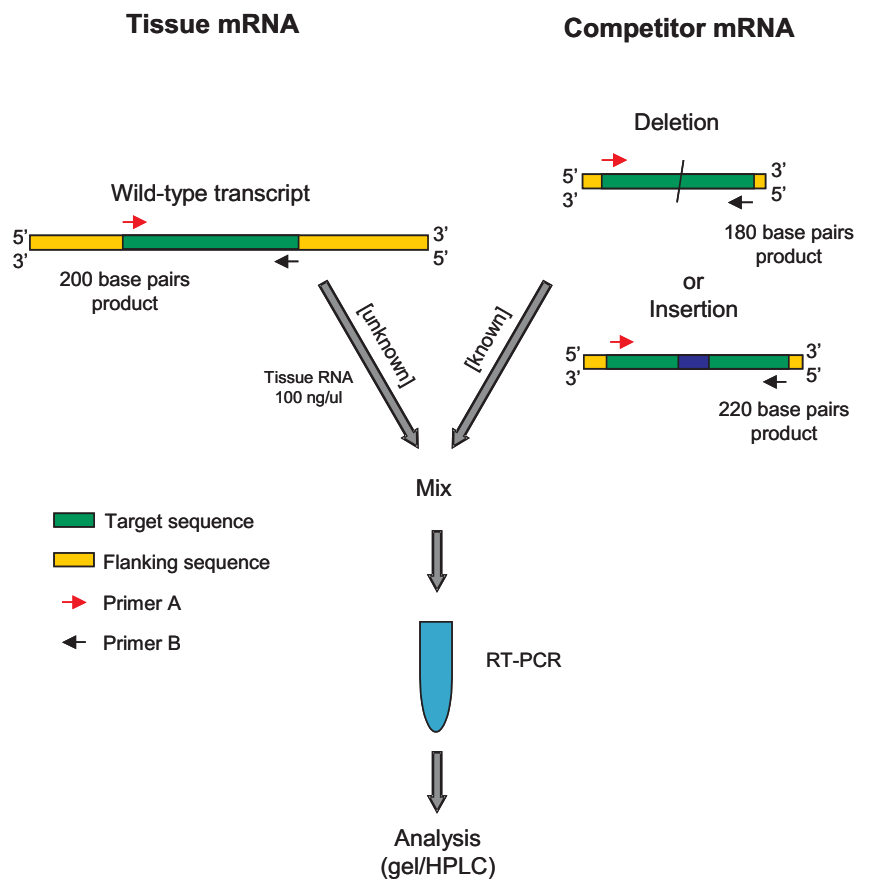
where N_x is the final amount of reaction product, N_0 is the initial amount of DNA in the reaction, x is the number of cycles and E is the efficiency of the polymerase chain reaction.

The competitor template must have the same PCR efficiency E as the native target template if it is to be a quantitatively accurate standard. This was tested using a variety of competitors constructed for several gene targets. In these tests, known quantities of both competitor and native DNA template are amplified and the reactions are used to estimate the known amount of native template. In all cases tested so far, identical PCR efficiency was observed. However, the reverse transcription efficiency (tested using known quantities of RNA templates) is not always equal between competitor and target mRNAs. Secondary structures in the form of hairpin loops may inhibit the processivity of the reverse transcriptase. Also, if the competitor mRNA has taken on a significantly different secondary structure than the native sequence due to the insertion or deletion of sequence then the reverse transcription may not be equally efficient in both transcripts. Up to now, however, this problem was encountered only once. However, it was shown that this altered reverse transcription efficiency was highly consistent and an accurate estimate of actual template copy number could be obtained by correcting for the consistent difference in RT efficiency.

The accuracy of this system was tested with RNA inputs reflecting RNA yields obtained from abundant tissue samples and also from microdissected samples comprising a few hundred cells. The sensitivity of the technique can be exploited, without costs of accuracy, down to very low target abundance.

Figure 3.59 illustrates the application of high-performance liquid chromatography (HPLC) to separation and analysis of RT-PCR reaction products produced in a competitive reaction titration. Chromatograms of reaction products are inset in the corresponding lanes of an agarose gel which was also used to examine the same reaction products. The expected competition of native template and competitor are visible on the gel as competitor input increases progressively from reactions 1 to 8 (lanes 2 to 9). An illustration of the analysis of such a titration is shown in figure 3.60. Densitometric analyses of such gels indicate that this titration meets the criteria expected, indicating that the titration should result in a straight line with a slope of one. Although linearity is observed, slopes of titrations obtained using gel densitometry have values different from one.

The reason for this discrepancy became apparent when the chromatograms were examined (figure 3.59). Each chromatogram has an initial peak in the void volume of non-retained reaction components (salts, water, free nucleotides, primers, proteins, *etc.*). Following this are the peaks corresponding to specific reaction products. Reactions which appear to produce two specific products on gel electrophoresis



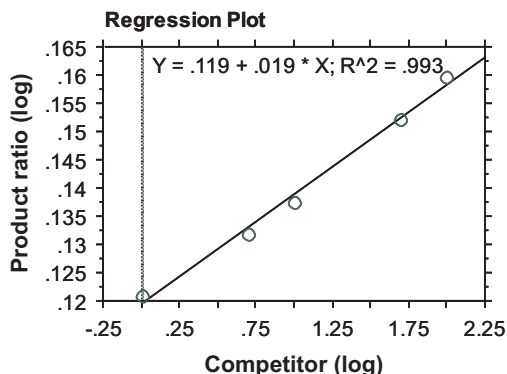
◀ **Figure 3.59 Schematic illustration of the competitive RT-PCR principle.** The purpose of single-tube competitive RT-PCR is to accurately quantify the number of transcripts of a specific mRNA present in a heterogeneous pool of tissue mRNA. This is performed by titration of a standard amount of native (tissue) mRNA against a known amount of synthetic competitor mRNA. This competitor is identical to the corresponding wild-type mRNA sequence present within the tissue mRNA except for the presence of an insertion sequence or a deletion. For this assay, the reverse transcription reaction is usually primed by random hexamers resulting in a heterogeneous mixture of cDNAs, a small subset of which is the wild-type cDNA of interest and the competitor. Specificity toward both the wild-type and competitor sequences are conferred by the use of primers that recognise

only those two species of cDNA. Therefore, these primers are used after the reverse transcription step to specifically amplify both wild-type and competitor cDNAs by PCR while excluding all other cDNAs. The reaction products are visualised here by gel electrophoresis and high-performance liquid chromatography (HPLC). Note the presence of a third (left most) peak in the chromatograms. This peak is the heteroduplex and is not visible in the gel. The results shown are from a multiple tube titration used to determine the range of native mRNA amount. Once this range is determined, single-tube titration can be performed from multiple mRNA preparations from the same tissue type. The following abbreviation is used in this figure: IP-RP-HPLC, ion paired-reverse phase-high performance liquid chromatography.

clearly produce three specific products on HPLC (for example, lanes 5 and 6). The third product (actually eluting first) is a heteroduplex made up of one strand (forward or reverse) of native and one strand of competitor DNA. This is known because if two reactions, each of which has produced only one (native or competitor) product, are mixed, heated and cooled to allow re-annealing, then three HPLC products are observed, though only two products are seen on gel electrophoresis.

Figure 3.60 Expression analysis results obtained by competitive RT-PCR.

Estimate of expression of the gene encoding the $\alpha 1$ subunit of sodium/potassium ATPase ($\alpha 1$ NKA) in RNA extracted from rat blood cells. Plot of logarithm of ratio of the amount of competitor to the logarithm of amount of native PCR reaction product determined as area under the curve in high-performance liquid chromatography (HPLC) against logarithm of the initial amount of competitor RNA added to the RT-PCR reaction. The amount of $\alpha 1$ NKA mRNA expression in this sample of rat blood total RNA was calculated as 263.7 femtograms per microgram total RNA. Note, however, that although the equation for this titration line is linear, the estimated slope of 0.019 differs from unity.



Analysis of mixtures which have not been heated to produce melting and re-annealing result in only the two homoduplex products being observed in both HPLC and gel electrophoresis. These observations are believed to be explained by the fact that the cross-sectional area of a heteroduplex (the attribute which principally determines gel migration) resembles closely that of a homoduplex made up of the two larger strands.

In IP-RP-HPLC, retention is determined by the number of matched base pairs. In a heteroduplex, this will always be slightly less than the number occurring in the smaller of the two homoduplex products. The inability of gel electrophoresis to detect heteroduplexes has led to confusion about if and when they are formed in competitive PCR reactions in which homologous internal standards are employed. Furthermore, heteroduplexes will disturb the 'apparent' ratio of products as seen on gel electrophoresis. In order to estimate the true ratio of reaction products the heteroduplexes must be quantified and the reaction products, of which they are comprised must be re-allocated back to their 'parent' products in order to accurately estimate the reaction product ratio. IP-RP-HPLC analysis of these reactions has clarified this issue so that not only are heteroduplexes detected, but their impact on the analysis can be fully incorporated into the data computation. The result is estimated in which the titrations are consistently linear with slopes indistinguishable from 1. A small computer program was written that automates the data analysis in order to incorporate the effect and size of the heteroduplexes detected (Doris *et al.*, 1997). The Q-RT-PCR program is available for download at <http://www.grad.ttuhs.edu/archive/index.html>.

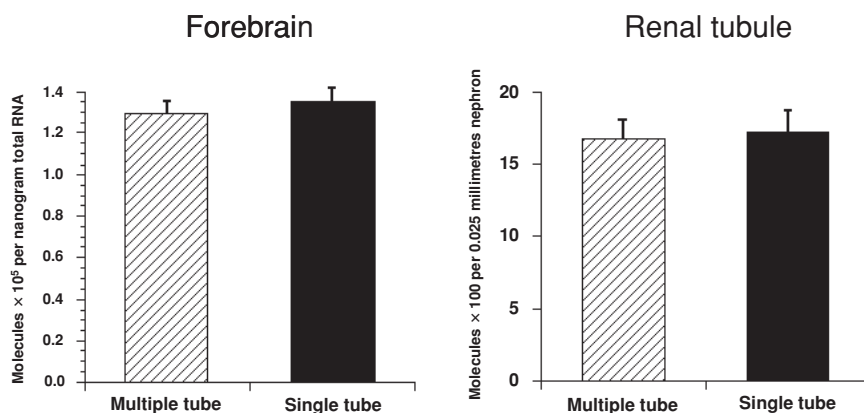


Figure 3.61 Estimation of precision of measurement of the RNA for the $\alpha 1$ subunit of sodium/potassium ATPase ($\alpha 1$ NKA) mRNA comparing multiple-tube titration with single-tube quantification for a high abundance (forebrain) and low abundance (renal tubule) samples. For each tissue, multiple measurements were made from one RNA sample and

the results are shown as mean $\pm$ standard error of the mean. This data confirms that, once the system is demonstrated to obey theoretical requirements for quantification, titration is not required and precise estimates of gene expression can be obtained from the analysis of products of a single tube reaction.

If the reverse transcription efficiency and PCR amplification efficiency are identical for competitor and native templates, then the final heteroduplex-corrected ratio of reaction products in a single reaction tube will always be identical to the initial ratio. Since one component of the initial ratio (amount of RNA competitor added) is known, the other component (the initial amount of unknown native RNA) can be estimated from the final product ratio. Having demonstrated the importance of heteroduplex detection and quantification, the quantification of transcript abundance was simplified to single-tube assays. The accuracy of single tube quantification was assessed in samples of brain and renal tubule RNA. The results are also shown in figure 3.61 (right bars) and indicate that similar precision can be obtained using single-tube quantification, even in low abundance samples such as RNA obtained from microdissected renal tubule segments.

3.3.4.3 Discussion

Single-tube competitive RT-PCR utilising IP-RP-HPLC for analysis provides a sensitive, precise and accurate assay system. The use of IP-RP-HPLC reduces post-reaction handling saving time and limiting the chance for crossover contamination of future competitive RT-PCR assays. Some investment of effort is required for initial assay set-up. This involves the *in vitro* synthesis of competitor mRNA. Competence in molecular biology techniques is required to generate the competitor which involves the cloning of the cDNA of interest into an expression vector, mutating the sequence, and use of the cDNA as a template for the RNA polymerase-directed synthesis of the competitor mRNA. Therefore, the set-up time is dependent on the researcher's experience with molecular biology techniques and the normal contingencies that occur when performing these techniques. Once the mRNA competitor is generated and the assay system validated, quantification can be performed using single reactions to determine the mRNA abundance of the target native mRNA in a heterogenous mRNA pool.

Other assays such as solution hybridisation, *in situ* hybridisation, and blotting have limitations. These limitations include the absence of precise quantification (blotting and *in situ* hybridisation) and the necessity of collecting relatively large amounts of RNA to perform the assay and generate sufficient signal for measurement (solution hybridisation and blotting). One assay offers similar sensitivity as competitive RT-PCR. This is the fluorogenic 5'-exonuclease assay utilising, for example, TaqMan® reagents and an ABI Prism® 7700 Detection System (Applied Biosystems GmbH, Weiterstadt, Germany) (Heid *et al.*, 1996). This assay allows for real-time quantification of mRNA expression. However, the system does not employ internal standards so that altered reaction efficiencies between samples which are attributable to sample contaminants may lead to erroneous quantification. The cost of the ABI Prism® Detection System and reagents is high and best suited to a specialised core facility.

Single-tube competitive RT-PCR, in particular in conjunction with IP-RP-HPLC, offers the researcher many advantages. Its sensitivity allows for the detection of

weakly expressed mRNAs or the detection of mRNAs from a small pool of tissue mRNA. Its accuracy and precision enables the researcher to detect small differences in mRNA abundance between samples. Lastly, the set-up time is offset by the minimal post reaction handling times of the samples.

3.3.5

GEF – Gene expression fingerprinting

3.3.5.1 Introduction

Gene expression fingerprinting (GEF) is a group of related techniques exploiting a common principle, namely visualisation of expression of cellular mRNAs using their representation in the form of specific cDNA restriction endonuclease fragments followed by the physical separation of the latter (Ivanova & Belyavsky, 1995 and 1997; Ivanova *et al.*, 1999; Shmelkov *et al.*, 2001). The concept of GEF involves three major steps: (i) creation of a highly representative primary set of 3'-terminal cDNA restriction fragments; (ii) subdivision of this primary set into simpler non or weakly overlapping secondary subsets; (iii) separation of the simplified subsets to visualise the individual or highly expressed sequences.

3.3.5.2 Principle and basics

The idea of the first step is to create a set of cDNA fragments in such a way as to represent (at least in theory) every mRNA by a single fragment of discrete size. Most commonly it is performed (figure 3.62) by introduction of the biotin label in the poly(A) proximal part (3'-end) of the cDNA during cDNA synthesis, followed by digestion with a frequently cutting restriction endonuclease such as *Nde* II (recognition site: 5'-GATC-3'; the enzyme generates a 5'-GATC overhang), whose recognition sequence occurs in 97 percent of cDNAs (Shmelkov *et al.*, 2001), and selection of the 3'-terminal fragments by streptavidin binding. This primary set of fragments is made amplifiable by the ligation of specific adapters to the 5'-end of the fragments.

The second step is necessitated by a very high degree of sequence complexity of the primary fragment set, which contains at least ten to twenty thousands of sequences, and therefore cannot be resolved by current separation methods. Several ways have been devised to split the primary fragment population into several simpler subsets. The common element in all these variants is a sequential liberation of non-overlapping secondary fragment subsets by the consecutive treatment of end-labelled fragments immobilised on streptavidin beads, with subsequent separation by high-resolution polyacrylamide gel electrophoresis (PAGE). In the original one-dimensional variant of the GEF procedure (figure 3.62, pathway A), there is no additional subdivision of the fragment sets. In this variant, each lane of the gel contains on average one to several thousands individual sequences, which substantially exceeds the resolving power of the gel. Therefore, only highly abundant mRNA species are visualised. This technique provides highly reproducible results and has worked well with substantially different mature cells types rich in abun-

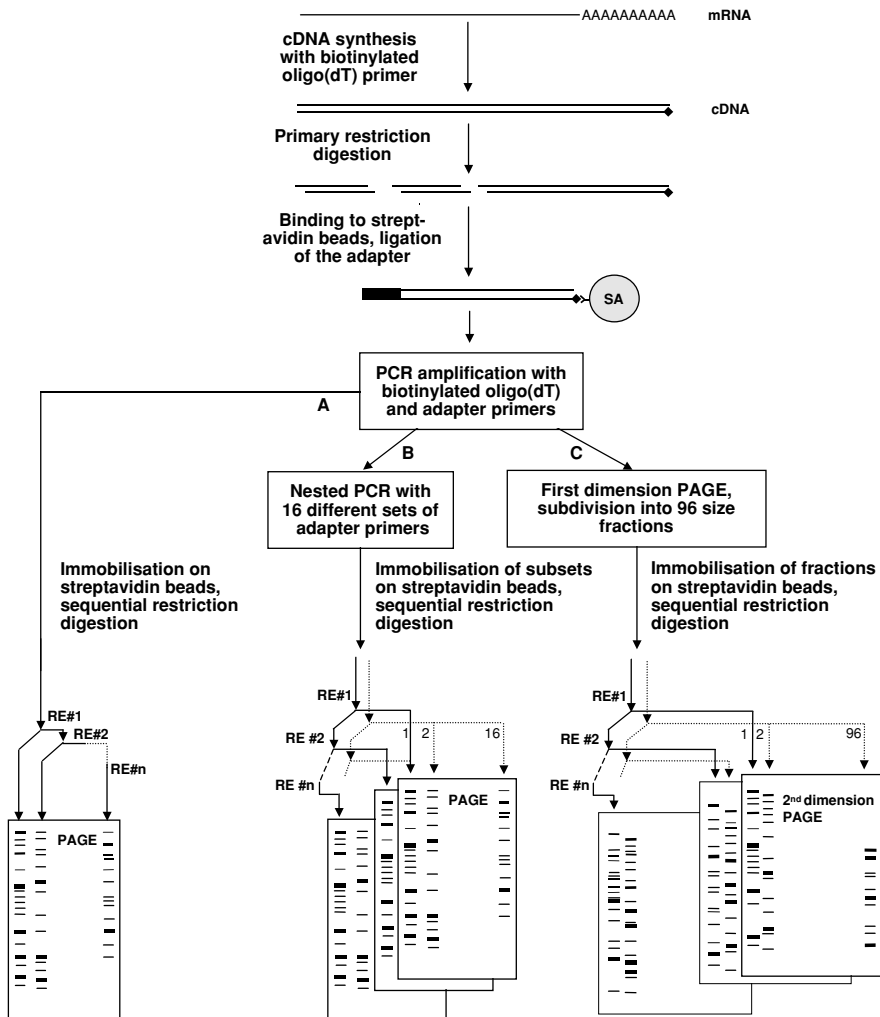


Figure 3.62 Scheme depicting three existing variants of the gene expression fingerprinting (GEF) protocol. In the original protocol **A**), the primary fragment set is directly subjected to sequential digestion with restriction enzymes. In other two variants, prior subdivision of the primary set into simpler subsets is achieved

either by additional PCR rounds with nested adapter primers **B**), or by separation according to size in the first dimension gel **C**). Rectangles indicate biotin groups. The following abbreviations are used in this figure: PAGE, polyacrylamide gel electrophoresis; RE, restriction enzyme; SA, streptavidin bead.

dantly expressed mRNAs, such as cells of spleen and thymus (Ivanova & Belyavsky, 1995); however it failed when highly-related embryonic cell populations were compared (Ivanova & Belyavsky, unpublished results). Therefore, more elaborated schemes for the subdivision of the primary fragment set were established. One of these exploits the use of nested PCR with adapter primers containing additional

3'-terminal bases (figure 3.62, pathway B). Under stringent PCR conditions, only the primers with perfectly matched 3'-ends are extended by *Taq* DNA polymerase. Therefore, nested PCR using four variants of adapter primer with a single additional 3'-base can subdivide the primary fragment set into four subsets. The adapter primers with two additional bases allow subdivision into 16 subsets, and so on (Ivanova & Belyavsky, 1997). The generated subsets are then immobilised on the streptavidin beads and subjected to the liberation of the secondary restriction fragment subsets as described above. This scheme was used to identify several mRNAs distributed differentially along the axes of developing *Xenopus laevis* embryos (Ivanova *et al.*, 1998). Although in theory this technique is very powerful, in practice it is difficult to avoid a certain degree of overlap between the fragment subsets. For minimisation of this overlap, the procedure must be performed under conditions of high stringency, which leads to the increased sensitivity of the PCR reaction to the minor uncontrollable variations of the experimental conditions. As a result, although the technique provides a much better resolution than the original variant, it also suffers from substantially higher irreproducibility, which necessitates running all samples in duplicate.

3.3.5.3 2D-GEF – Two-dimensional gene expression fingerprinting

A third variant of the GEF technique (figure 3.62, pathway C) utilises two-dimensional PAGE for efficient resolution of 3'-terminal cDNA restriction fragments produced by one primary (first dimension) and several sequential secondary restriction digestions. This variant combines the high reproducibility of the original version of GEF with very high resolving power, and obviates the need for a subdivision of the material into simpler subsets prior to electrophoresis. Many thousands individual sequences per cDNA sample can be visualised using this strategy. Owing to this high resolving power, most of the cellular mRNAs have a unique combination of the sizes of the primary and secondary cDNA restriction fragments. This allows one, using a computer analysis of known mRNA sequences, to identify the corresponding cDNA fragments on the basis of their two-dimensional coordinates.

A detailed scheme of this approach is shown in figure 3.63. The generation of the primary cDNA fragment set is performed essentially as for other variants. Double-stranded cDNA is synthesised by the method of Gubler & Hoffman (1983) using a one-base anchored 5'-biotinylated oligo(dT)-containing primer, and digested with a frequently cutting four-base recognition site restriction enzyme. 3'-terminal fragments of cDNA are selected and amplified as described in the previous section. This primary cDNA fragment population is resolved according to size in the first dimension of denaturing polyacrylamide gel electrophoresis. For transfer to the second dimension, the resolved fragments (usually in the 100 to 1,000 base range) are subdivided into 96 size fractions and eluted from the gel in separate wells of a 96-well plate. Eluted single-stranded fragments are bound to the streptavidin beads, rendered double-stranded using Sequenase® and an adapter primer that is heavily labelled with ³²phosphorus, and is then treated sequentially with a set of restriction enzymes. The set of restriction enzymes is se-

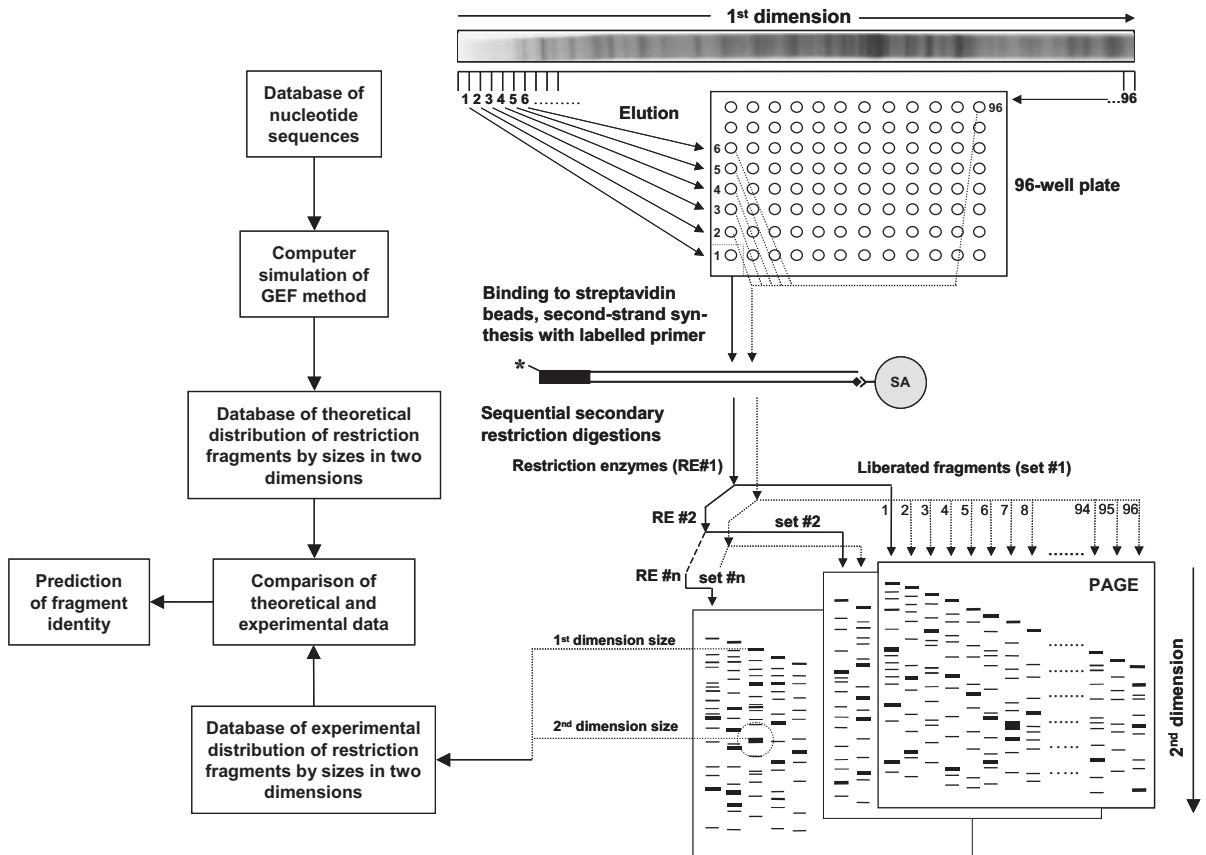


Figure 3.63 Detailed scheme of the two-dimensional gene expression fingerprinting (2D-GEF) procedure and computer-assisted analysis of the results of two-dimensional separations. Shown are the steps after preparation of the

primary cDNA fragment subset. See text for details. Rectangles indicate biotin groups. The following abbreviations are used in this figure: PAGE, polyacrylamide gel electrophoresis; RE, restriction enzyme; SA, streptavidin bead.

lected in such a way as to minimise variations in the number of cDNA fragments liberated after each round of digestion (Shmelkov *et al.*, 2001). Two different formats can be used for separation of the liberated fragments in the second dimension. In the first format, fragment subsets liberated by a given enzyme from the entire range of size fractions can be resolved on a single gel to produce a two-dimensional picture (figure 3.64). The second format is more appropriate if the identification of differentially expressed sequences is the major objective. In this case, fragments from equivalent size fractions corresponding to two or more different cell samples can be loaded side-by-side to facilitate the comparison of the expression patterns. If necessary, the cDNA fragments of interest can be recovered from the gel, oligo(dG)-tailed and amplified by PCR.

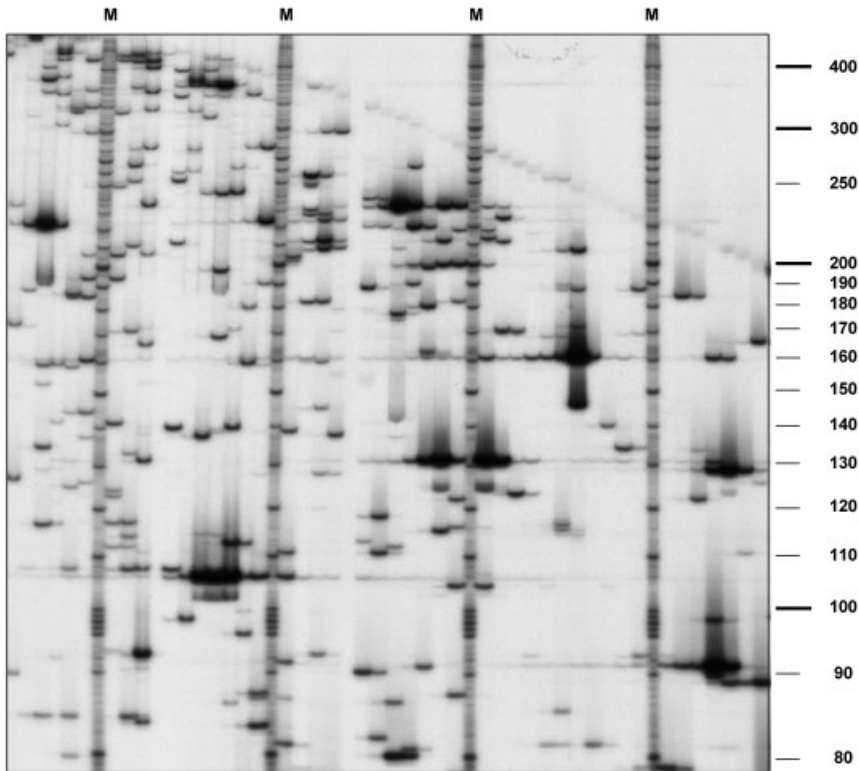


Figure 3.64 Example of the two-dimensional separation of the cDNA fragments on a six percent denaturing polyacrylamide gel. A cDNA sample is prepared from sorted murine bone marrow cells with the phenotype $\text{Rho}^-/\text{Rho}(\text{VP})^+$ (for details, see Belyavsky *et al.*, 2002; Zijlmans *et al.*, 1995). Shown are the fragments released from the streptavidin beads by *Pst* I

digestion. Prior to *Pst* I digestion, beads were treated with *Eco* RI. M indicates the SequaMark™ (ResGen™, Invitrogen GmbH, Karlsruhe, Germany) marker lanes. Although the marker provides a one-base resolution in the range of lengths of 40 to 400 bases, only selected band positions are indicated on the right side.

3.3.5.4 Computer simulation of the 2D-GEF procedure

To utilise the advantages provided by the 2D-GEF method, software for the analysis of the two-dimensional gel patterns was created (figure 3.63) (Shmelkov *et al.*, 2001). The UniGene database (National Center for Biotechnology Information, NCBI; Bethesda, Maryland, USA; available at <http://www.ncbi.nlm.nih.gov/UniGene/>) was used as a source of human cDNA sequences. The software simulates the primary and sequential secondary restriction digestions of cDNA and creates a sub-database of predicted two-dimensional coordinates of known human sequences. Two methods are available: (i) from database to the gel (to identify bands on two-dimensional gels which correspond to known genes), and (ii) from gel to the database (to identify in the database corresponding to fragments of interest on the second dimension gels). Experiments indicated that at least 70 percent of

the fragments can be identified correctly using this software. Most incorrect identifications were not caused by inherent flaws in the procedure, but were due to insufficient quality of the database (absence of complete 3'-terminal sequences, sequence errors, *etc.*) (Shmelkov *et al.*, 2001).

3.3.5.5 Advantages and disadvantages of the 2D-GEF procedure

One of the most significant advantages of the 2D-GEF procedure compared to the previously developed one-dimensional GEF variants is its high resolution based on additional subdivision of cDNA fragments in the first dimension. Accounting for partial overlap between size fractions, the reduction in sequence complexity of each lane of the second dimension is on average 60 to 80-fold. After subsequent ten to 15 secondary restriction digestions, each lane of the second dimension gel contains on average 600 to 1,200-fold fewer sequences than in the original cDNA population. The 2D-GEF procedure thus possesses one of the highest resolutions among techniques based on gel display of cDNA restriction fragments (Kato, 1995 and 1996; Prashar & Weissman, 1996; Suzuki *et al.*, 1996). Combined with computer analysis, this high resolution provides a way to identify expressed genes solely on a basis of the two-dimensional coordinates of the corresponding cDNA fragments. High resolution and clean separation of the bands on two-dimensional gels allow in many cases a direct sequence analysis of the cDNA fragments after PCR amplification without subsequent cloning.

2D-GEF is also characterised by high reproducibility, which reaches 99.5 to 100 percent in the parallel experiments on duplicate samples (Shmelkov *et al.*, 2001). 2D-GEF provides a sensitive technique for the detection of gene expression in small cell populations. As few as two thousand sorted cells can be used to produce high-quality 2D-GEF patterns (Belyavsky *et al.*, 2002), and even the low-abundance mRNA expressed at a level of several copies per cell can be detected (Shmelkov *et al.*, 2001).

Computer simulation of the procedure reveals some limitations in terms of percentage of cDNA available for the 2D-GEF analysis. Although about 97 percent of cDNA sequences are involved in the procedure after digestion with a primary restriction enzyme such as *Nde* II, 13 percent of these represent fragments that are either too short or too long for proper analysis. Of the remaining, 80 percent to 95 percent are liberated after ten to 15 cycles of restriction enzyme digestion. Taken together, the proportion of cDNA sequences analysed by the 2D-GEF procedure is about 70 to 85 percent of the total cDNA population. However, a nearly complete analysis of an mRNA population by the 2D-GEF (95 percent or more) can be achieved by performing an additional experiment with a different primary restriction enzyme.

Compared to the SAGE (serial analysis of gene expression) procedure (Velculescu *et al.*, 1995), 2D-GEF is likely to be a faster and less expensive technique. Moreover, due to a larger size of fragments analysed, downstream cloning of sequences of interest is more straightforward with the GEF. However, in contrast to SAGE, the GEF procedure does not directly provide sequence information or absolute frequencies of sequences in a cDNA population.

In comparison to the 2D-GEF analysis, cDNA or DNA microarray analysis (Schena *et al.*, 1995) represents a faster and less labour-intensive technique. Microarray assays depend heavily on a large body of pre-existing sequence information and cDNA/EST clone collections. Microarrays are therefore most efficiently used at late stages of genome projects. The 2D-GEF procedure does not depend on the sequence information of expressed mRNAs, but can use it for comparison of experimental and predicted patterns and for gene discovery. 2D-GEF analysis is therefore best suited for initial or intermediate stages of genome projects, and could probably most optimally be used in combination with EST cloning and DNA microarray analysis.

3.3.6

IBDD – Indexing-based differential display

3.3.6.1 Introduction

molecular indexing provides a simple and highly reproducible approach to profiling the expression status of a gene population, starting from small amounts of total RNA and enabling multiple side-by-side comparisons. Molecular indexing was initially utilised as a non-cloning method for isolating specific DNA fragments from complex genomic digests by selective ligation of defined adapters and PCR (Smith, 1992; Unrau & Deugau, 1994). The first application of this approach to cDNA populations demonstrated the power of selection by ligation to mixed adapter sets (Sibson, 1992; Sibson & Starkey, 1997). More recently, profiles of genes expressed in mouse tissues have been generated by PCR amplification of cDNA class IIs restriction enzyme fragments and recorded using an automated DNA sequencer (Kato, 1995 and 1996). Molecular indexing has also been applied to the isolation of differentially expressed genes (Mahadeva *et al.*, 1998; Ryan *et al.*, 2001; Shaw-Smith *et al.*, 2000). The potential of the approach to identify novel genes associated with a disease process was demonstrated in a rat model of pressure-induced left ventricular hypertrophy (Mahadeva *et al.*, 1998), and by a comparison of the expression profiles within prefrontal cortices derived from healthy individuals and patients with Alzheimer's disease (Ryan *et al.*, 2001).

3.3.6.2 Principle and basics

Total RNA is the template for the synthesis of double-stranded cDNA (figure 3.65). First-strand cDNA synthesis is performed using oligo(dT)₁₂₋₁₈ and/or random hexamers and a modified reverse transcriptase that contains no intrinsic RNase H activity. The population of cDNA molecules is digested with a type IIs restriction endonuclease, generating fragments with non-identical cohesive ends. The number of different end sequences is 4^n , where n is the length of the overhang. An enzyme such as *Bbv* I generates fragments with four-base 5'-overhangs. If two cohesive ends of a *Bbv* I fragment are considered, there are potentially $\frac{1}{2} \cdot 4^8$ (= 32,768) fragment classes, each with a different pair of cohesive ends.

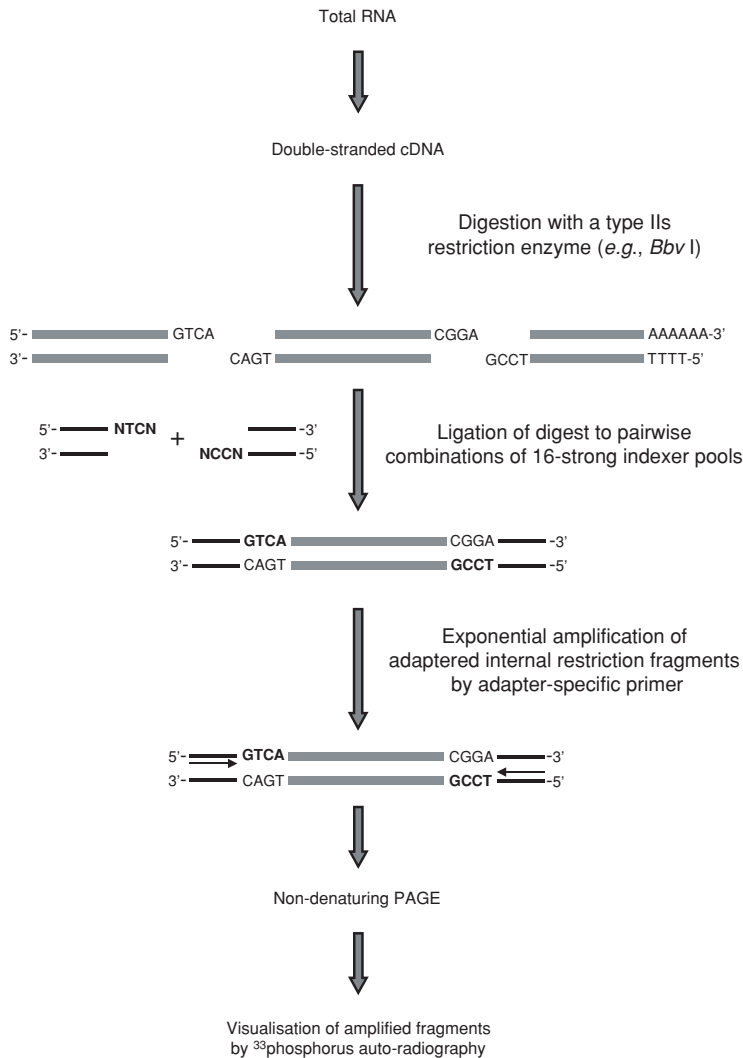


Figure 3.65 A schematic illustrating of the indexing-based differential display (IBDD) procedure. See text for detailed explanations. The following abbreviation is used in this figure: PAGE, polyacrylamide gel electrophoresis.

The T4 DNA ligase-mediated ligation of adapters with perfectly complementary overhangs enables ordered partitioning of the restriction fragment classes as non-overlapping subpopulations. For fragments with four-base overhangs, 4^4 (= 256) different adapters are required. In order to access every internal cDNA fragment from a restriction digest, it is necessary to provide each cDNA fragment with every adapter in combination with every other adapter. The use of single adapters in pairwise combinations means that 32,896 reactions are

required to isolate all the internal fragments in a *Bbv* I digest. However, division of the 256 adapters into 16 pools of 16 adapters reduces the number of ligation reactions to 136 (comprising 120 reactions featuring pairwise combinations of different adapter pools and 16 reactions effectively containing single adapter pools). Triplicate ligation reactions, each featuring a 0.25 micrograms of total RNA cDNA equivalent, are routinely performed using each adapter pool combination for each RNA sample.

Internal cDNA fragments, which have acquired an adapter at each end, are exponentially amplified by adapter primer PCR (incorporating [α -³³P]-dATP) and visualised by non-denaturing polyacrylamide gel electrophoresis (figure 3.65).

3.3.6.3 Examples and typical results

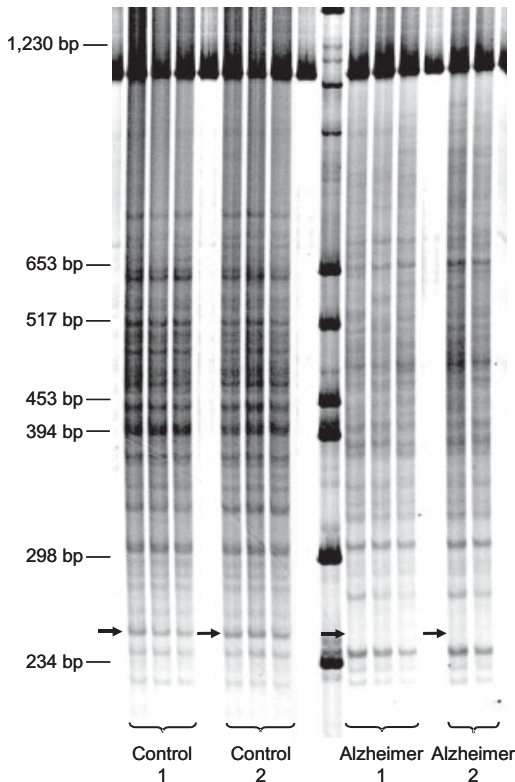
The power of IBDD for the identification of differentially expressed genes was first demonstrated (Mahadeva *et al.*, 1998) using a rat model of pressure-induced left-ventricular hypertrophy (LVH), a condition characterised by increased expression of certain oncogenes and structural proteins. Expression profiles were generated for LVH animals and age and weight matched control animals using eight of the potential 136 pairwise adapter pool combinations. Twenty-nine putatively differentially expressed fragments, from a total of 209 visualised, were excised and cloned. The differential expression and the direction of the differential expression of nineteen (66 percent) of the fragments isolated were confirmed by Northern blot analysis. Sequence analysis identified both genes known to be up-regulated in LVH, and novel genes.

The potential for IBDD to identify changes in gene expression in post-mortem brain tissue was subsequently evaluated by comparing the expression profiles of six Alzheimer's disease and five age, sex and post mortem interval-matched control prefrontal cortices, respectively (figure 3.66) (Ryan *et al.*, 2001). Consistent changes (confirmed by semi-quantitative RT-PCR on nine Alzheimer's Disease and nine control brains) were found in the levels of a number of mRNAs, some of which have a known association with the disease process in Alzheimer's Disease. This work demonstrated that valid and reproducible differences in gene expression could be detected by differential display using total RNA from human *post mortem* brain tissue.

3.3.6.4 Discussion

The partitioning of cDNA fragments generated by class II endonucleases enables the ordered subdivision of a transcript population (Kato, 1995 and 1996; Mahadeva *et al.*, 1998; Sibson, 1992; Ryan *et al.*, 2001; Sibson & Starkey, 1997), increasing the probability of both the identification of a differentially expressed transcript and its isolation as a single species. The subdivision of cohesive-ended cDNA restriction fragments occurs according to the sequence of their ends. Consequently, in contrast to approaches featuring subtractive hybridisation (*e.g.*, representational difference analysis, RDA; Hubank & Schatz, 1994), indexing is not prone to the elimination of low abundance cDNAs due to repetitive sequences shared by non-homologous cDNAs. Indexing provides a means of subdividing cDNA fragments into dis-

Figure 3.66 Comparison of gene expression profiles between Alzheimer's and control individuals using indexing-based differential display (IBDD). Non-denaturing polyacrylamide gel electrophoresis (PAGE) of fragments accessed by indexing-based differential display of *Bbv* I digests of cDNA prepared from prefrontal cortices of brains from two Alzheimer's and two control individuals, respectively. Depicted are profiles generated by duplicate or triplicate ligation reactions featuring a single adapter pool pair combination. An arrow indicates the dramatic reduction in the expression of GAP43 (growth associated protein 43) in the Alzheimer's brains.



tinct subpopulations, since in concept a given fragment can only be sorted into a single subset.

Fundamental to the effective isolation of differentially expressed genes is the fact that indexing generates display patterns that are highly consistent and reproducible (figure 3.66). The fidelity of T4 DNA ligase-catalysed adapter ligation is the sole basis for the sorting of cDNA fragments into distinct subsets. This is in contrast to similar procedures (Matz *et al.*, 1997; Prashar & Weissman, 1996; Vos *et al.*, 1995) in which cDNA population subdivision is achieved by single or two base-specific reverse transcription and/or PCR. It was sought to increase the fidelity of T4 DNA ligase-catalysed adapter ligation by 'hot start ligation' (Mahadeva *et al.*, 1998; Ryan *et al.*, 2001), an analogue of hot start PCR (Chou *et al.*, 1992). The components of the ligation reaction are heated to 65 °C prior to the addition of the T4 DNA ligase in order to reduce the frequency of mismatch ligation by precluding pre-ligation non-specific base pairing. Although the fidelity of adapter ligation was demonstrated through the isolation of known gene fragments by appropriate adapter combinations, evidence of mis-ligation was documented (Mahadeva *et al.*, 1998). Although mis-ligation is expected to occur at a rate at least 1,000-fold less than specific ligation under 'standard' conditions (Unrau & Deugau, 1994), the

likelihood of mis-ligation may be further reduced by reference to the ligase employed and the ligation temperature (Shaw-Smith *et al.*, 2000).

PCR amplification of indexed cDNA fragments is based upon a single 'long' PCR primer. This ensures that a fixed set of thermal cycling parameters is optimal. This is completely opposite to the low stringency annealing of arbitrary primers associated with the differential display RT-PCR (Liang & Pardee, 1992) and RNA arbitrarily primed PCR fingerprinting (Welsh *et al.*, 1992) protocols. Consequently, indexing is not prone to the consequences of stochastic events during the first few cycles of low stringency PCR, which contribute significantly to false-positives with use of conventional differential display RT-PCR.

Each individual subpopulation generated by indexing is of relatively low complexity as compared to the original population. Consequently, the concentration of any given cDNA will be higher than in the original population. PCR amplification of cDNA fragments separated into different subsets serves to further enrich the abundance of rare mRNAs to the extent that they become detectable. Assay of the current indexing procedure, by the spiking of mRNA populations with known amounts of a synthetic mRNA, suggests that it delivers a sensitivity equivalent to the detection of transcripts expressed at the level of between one in 10,000 and one in 100,000 molecules (Mahadeva *et al.*, 1998).

Quantitative differential hybridisation of DNA microarrays with complex cDNA probes offers a convenient high-throughput approach to simultaneously monitoring the expression of many previously isolated genes, and can provide a sensitivity of 0.0002 percent (Lashkari *et al.*, 1997; Schena *et al.*, 1996). However, from the perspective of gene discovery, the indexing procedure enables the visualisation of novel transcripts without the prerequisites for a gene to be represented in a cDNA library and selected for inclusion in an array.

One of the potential virtues of identifying differential gene expression using DNA restriction fragments is the immediate identification of the expressed genes through correlation of gel band sizes with cDNA sequence databases. The approach outlined here is directed at the isolation of internal cDNA fragments, as opposed to 3'-end fragments (Kato, 1995; Liang & Pardee, 1992; Matz *et al.*, 1997; Prashar & Weissman, 1996), which affords the cloning of protein coding regions.

3.3.7

Molecular indexing of 3'-cDNA ends

3.3.7.1 Introduction

The mRNA derived from mammalian cells and tissues is a complex mixture of more than 10,000 species. Several molecular biology techniques are designed to analyse individual genes, not entire populations. With advances in the human genome project, however, biological analyses now need to evaluate entire gene populations. Molecular indexing (Unrau & Deugau, 1994; Kato, 1995) is an approach facilitating such analysis. The unique concept behind this technique is to divide members of the mRNA population into several hundred groups, each containing an adequate number of genes for PCR co-amplification. Instead of the

analysis of each gene, groups of genes are handled. Consequently, the number of experiments is reduced. Each group is then used for differential display RT-PCR or differential hybridisation.

3.3.7.2 Principle and basics

3.3.7.2.1 Basic enzymatic process

Classic separation procedures are dependent on the physical characteristics of RNA molecules, such as molecular weight and intracellular concentration. These methods of separation can only divide these molecules into a few groups, inadequate for comprehensive analysis. In molecular indexing, grouping utilises terminal short sequences (three to four bases) of DNA fragments.

Class IIS restriction enzymes cleave DNA at a precise distance from the recognition site, producing an overhang with an unknown sequence. The basic technique of indexing ligates each DNA to an adapter set, which covers all possible nucleotide sequences (Kato, 1995). A total of 64 adapters covered all possible sequences. Each adapter had a 5'-four nucleotide-overhang, for which the outermost base was a mixture of adenine, cytosine, guanine and thymine and the inner three bases contained all possible sequences. The 5'-ends of the adapter common short fragment were labelled by biotin, allowing recovery with streptavidin-coated paramagnetic beads. DNA ligase isolated from *Escherichia coli* joined these adapters with matched sequences; ligation of each adapter results in the division of the digested fragments into 64 groups. Utilising the 3'-end of cDNA fragments is useful to analyse gene expression, as each gene contains a unique 3'-end cDNA fragment with a poly(A) tail. The 3'-end fragments can be subdivided into 576 groups by repeating the adapter ligation for each individual group (figure 3.67.A) and two additional procedures. After adapter ligation, and prior to recovery with paramagnetic beads, the cDNA was digested by the two class IIS enzymes not used in the previous digestion. Fragments cut by these enzymes were excluded from the PCR templates used in the next stage. Therefore, only fragments for which the first enzyme recognition site is near the poly(A) tail are amplified. Repeating the first digestion with all three enzymes divides the population into three groups (figure 3.67.B). Then, PCR amplification is performed using a primer specific for the adapter and another recognising the three anchored oligo(dT) primers, d(T)₂₅A, d(T)₂₅C and d(T)₂₅G (figure 3.67.C). Thus, the subdivision of the entire mRNA population into 576 groups is achieved.

The accuracy of the technique depends on the sequence specificity of the ligation. Therefore, the specificity of *Escherichia coli* DNA ligase was examined. Ten cDNA fragments selected from mouse liver mRNA sequences and containing a variety of adapter sequences were amplified with either the proper adapters or with adapters possessing a one-base replacement (Kato, 1995). Following separation by agarose gel electrophoresis and Southern transfer, each membrane was probed with a labelled oligonucleotide specific for each gene. Eight out of the ten genes examined were amplified to levels detectable by the DNA sequencer.

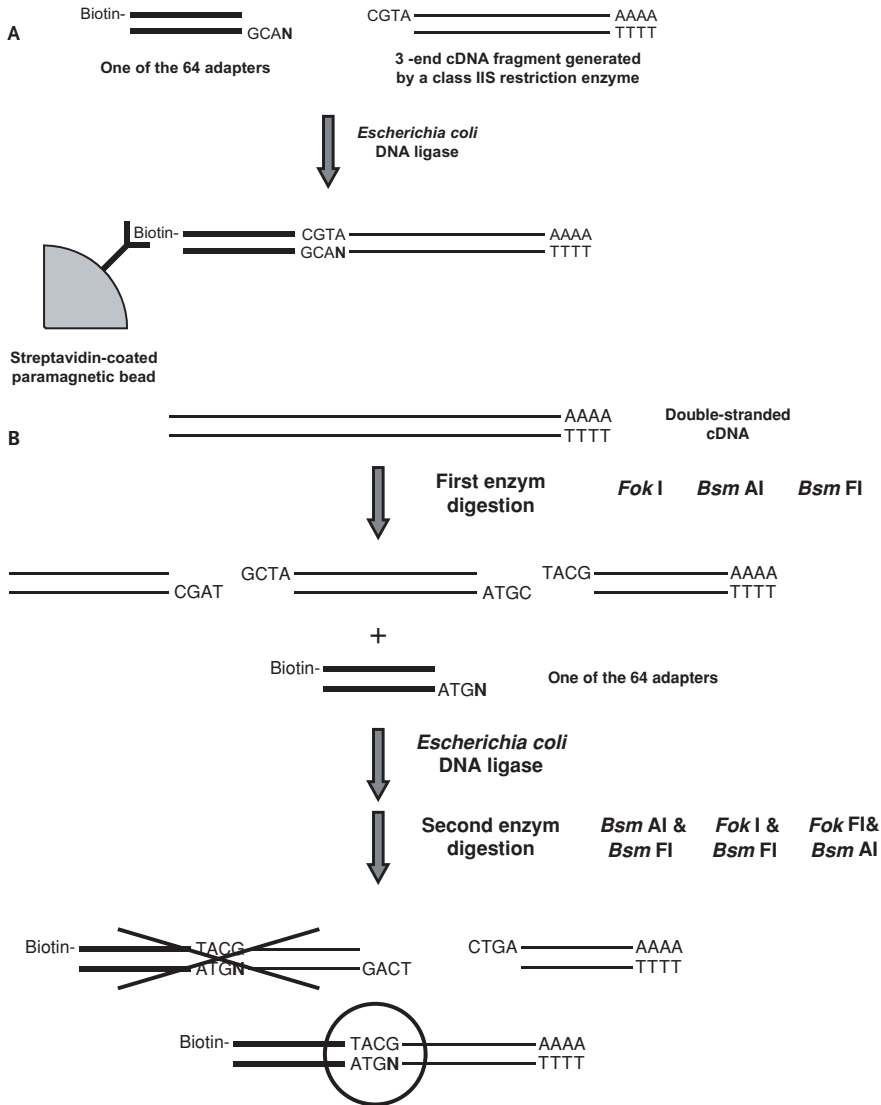
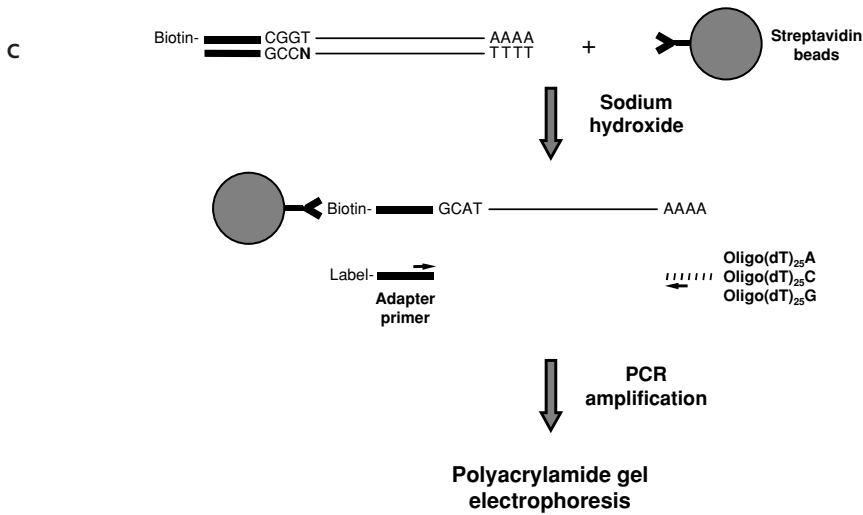


Figure 3.67 Schematic representation of the molecular indexing for 3'-ends of cDNA fragments method. See text for further details.

With the exception of prothrombin, amplification proceeded only in the presence of the correct adapters. The intensity of the prothrombin signal with the substituted 5'-NACC-3' sequence was seven percent of that obtained with the correct sequence, 5'-NCCC-3'. These results showed that *Escherichia coli* DNA ligase was capable of discriminating between the three nucleotides adjacent to the ligation site. Therefore, each 3'-cDNA fragment was amplified only in the presence of the correct adapter, except for abundant species.



In addition, DNA sequencing of the final amplified products was performed (Kato, unpublished results). As both the 5'-ends of the oligonucleotides are not phosphorylated and the longer fragments of adapters are not ligated, only overhang sequences of cDNA are amplified. By comparing the amplified sequences with those ligating adapters, the specificity of the ligation procedure was validated. Sixty out of 64 sequences from cloned cDNAs matched those of the adapter sequence. Thus, more than 90 percent of the final products were derived from the correct ligation molecules, suggesting that *Escherichia coli* DNA ligase is well suited to the indexing reaction. On contrary, no experimental evidence exists supporting the sequence specificity T4 DNA ligase used by Unrau & Deugau (1994).

Because this molecular indexing method described above uses 3'-cDNA fragments to achieve the full representation of all expressed genes, coding regions are often lacking from these fragments. In human or mouse tissues which have a huge resource of genomic and cDNA sequences, this does not prove to be a disadvantage. When the technique is applied to organisms with little sequencing information, however, this limitation may inhibit the identification of genes isolated by molecular indexing. Therefore, an alternative method to analyse cDNA-coding regions was designed (Kato, 1996). The principle of this method is schematically represented in figure 3.68.

First, the double-stranded cDNA is cut by enzyme '1', a class II restriction enzyme (e.g., *Eco* RI). The first adapter for this enzyme is ligated to the cohesive end. The ligated molecule is then cut by the second enzyme, a class IIS restriction enzyme (e.g., *Fok* I). Following ligation of one of the 64 indexing adapters to the end of the fragment, PCR amplification was performed with a primer specific for the indexing adapters and a primer recognising for the first adapter. Repeating these experiments with all of the indexing adapters divided the cDNA fragments' first and second enzyme recognition sites into 64 groups. Therefore, although it

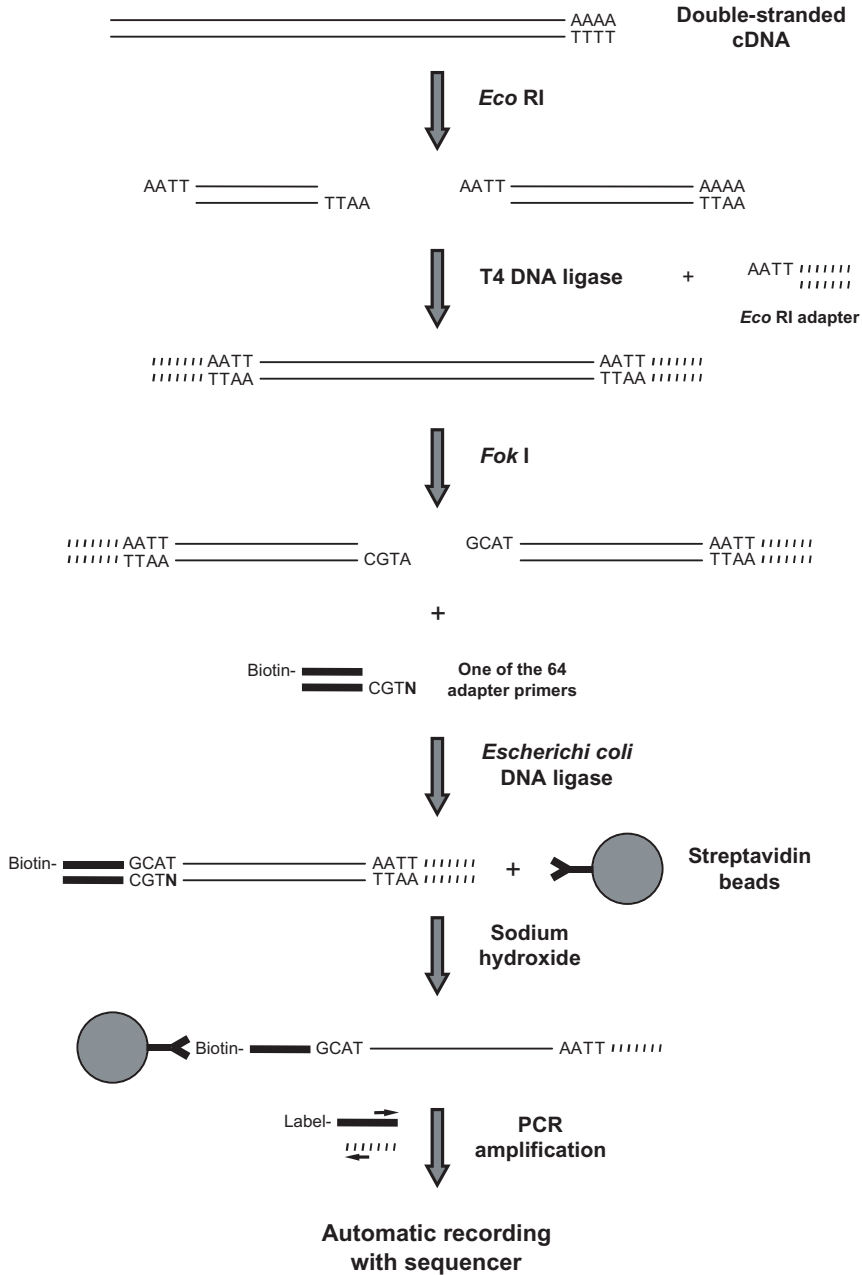


Figure 3.68 Schematic representation of the molecular indexing method for analysing upstream regions. Details are described in the text.

cannot represent the entire mRNA population, these amplified fragments are not biased to only include the 3'-ends.

3.3.7.2.2 Application to display and hybridisation

Molecular indexing was originally developed for use with differential display RT-PCR. Products amplified from two or more RNA samples are separated by polyacrylamide gel electrophoresis; the resulting fingerprinting patterns are then compared. Differentially expressed genes are easily detected through a comparison of the patterns (figure 3.69). Although three class IIS restriction enzymes were originally used, the use of only two enzymes proved to be more appropriate to separate the samples by polyacrylamide gel electrophoresis (Kato, unpublished results).

Currently, DNA microarrays are a popular tool for gene expression profiling. This method, however, utilises differential hybridisation, which suffers from many types of artefacts. These inaccuracies make experiments with high complexity mRNA very difficult. By reducing the complexity of the sample, these types of problems can be avoided.

Molecular indexing divides an mRNA population into 64 or more groups. Each group contains fewer species of mRNA, thereby reducing the sample complexity. The use of each fraction as a hybridisation target allows the identification of differentially expressed genes not detected by conventional differential hybridisation. Pilot experiments using a 16-group fractionation to isolate genes differentially expressed between the mouse cerebrum and cerebellum identified ten times more candidate genes than conventional hybridisation (Sakai *et al.*, 2000).

3.3.7.3 Examples

3.3.7.3.1 Comparison of gene expression between Alzheimer and normal brain

Alzheimer's disease is one of the most prevalent neurodegenerative diseases, causing progressive senile dementia. Many studies have suggested the involvements of amyloid- β protein and phosphorylated τ in neuronal cell death. The accumulation of these products may change the molecular environment of neurons, possibly leading to neuronal toxicity. A comprehensive analysis of gene expression in a brain affected by Alzheimer's disease may identify novel molecules involved in disease pathogenesis. Here, the work by Doyu *et al.* (2001) is presented.

Total RNA was purified from a brain affected with Alzheimer's disease and a control brain from a similarly aged patient. The molecular indexing reaction was performed with these RNAs. Two class IIS restriction enzymes, *Bsm* AI and *Bsm* FI, were used to amplify the samples into 384 groups. The separation of the products derived from each sample by an automatic DNA sequencer recorded the sizes and peak intensities automatically. By pairwise comparison, genes whose peaks exhibited greater than five-fold differences in abundance were selected as candidates for differential expression. The corresponding fragments were purified from these gels, then cloned and sequenced. Of the 70 genes identified, the gene expression patterns of 47 were confirmed by ATAC-PCR (Kato, 1997). Fifty-one demonstrated

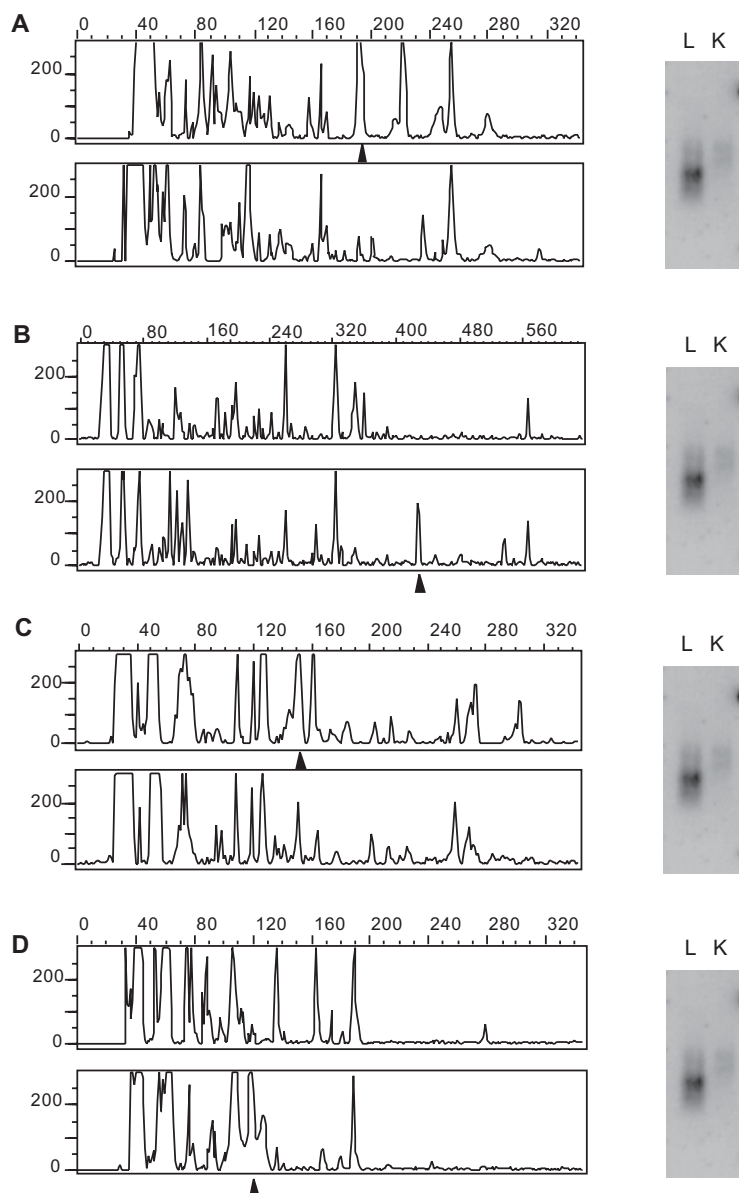


Figure 3.69 Electropherograms displaying gene expression differences identified by the molecular indexing method. *Left:* The top electropherograms display cDNA fragments derived from mouse liver RNA; those at the bottom detail cDNAs derived from mouse kidney RNA. cDNAs were cut with *Fok* I in the first digestion. Adapters and primers are as follows: **A)** 5'-NGAA-3' adapter, TAMRA-C1S and d(T)₂₅G;

B) 5'-NAGC-3' adapter, TAMRA-C1S and d(T)₂₅G; **C)** 5'-NGCG-3' adapter, TAMRA-C1S and d(T)₂₅G; **D)** 5'-NGAT-3' adapter, JOE-C1S and d(T)₂₅C. *Right:* Northern blot hybridisation to confirm differential mRNA expression. Each probe corresponds to the peak marked by an arrow head. Lane L, total mouse liver RNA; lane K, total mouse kidney RNA.

elevated expression in control brain, with high expression of the remaining 19 in Alzheimer's disease-affected brain. Glial fibrillary acidic protein (GFAP), a marker gene highly expressed in the brains of Alzheimer's patients, was identified in this analysis, reinforcing the validity of the technique. Fragments matched with known sequences are shown in table 3.7.

The above work indicates that molecular indexing can be applied to RNA isolated from clinical materials sampled under non-ideal conditions. Several previous studies have isolated genes differentially expressed between Alzheimer's disease-afflicted and normal brains (Santiard-Baron *et al.*, 1999; Doyu *et al.*, 2001; Ryan *et al.*, 2001). The advantages of molecular indexing in these studies, however, is not apparent. It should be emphasised, however, that the expression patterns of the candidate genes identified by molecular indexing were confirmed by adapter-tagged competitive PCR. The absence of pseudopositives provides a valuable advantage for this technique.

3.3.7.3.2 Comprehensive collection of mouse zinc finger motifs

Molecular indexing was designed to allow analysis of the entire mRNA population. This technique, however, can also be applied to limited populations of genes which share a common sequence. Protein families often share common amino acid sequences suitable for PCR primer design. The C₂H₂ zinc finger protein family contains more than 1,000 members. The comprehensive analysis of the family members and their expression patterns may be an interesting application of molecular indexing (Yamashita *et al.*, 2001).

In addition to a zinc finger motif, the C₂H₂ zinc finger proteins share a common sequence designated 'HC-link'. A primer based on this sequence was utilised for the indexing reaction, replacing the anchored oligo(dT) primers. As the number of amplified genes was estimated to be small, the templates were simultaneously digested by *Fok* I and *Bsm* AI. The amplified fragments were then classified into 64 groups by adapter ligation. Using total RNA isolated from eight independent mouse tissues, 1,471 fragments were identified. Sequence analysis of 1,086 fragments revealed that 259 encoded the C₂H₂ motif, including 66 known zinc finger genes, 28 dbEST-matched sequences, and 165 novel sequences. To the current knowledge, this is thought to be the largest collection of zinc finger motif sequences experimentally identified from transcripts.

3.3.7.4 Discussion

Molecular indexing has several advantages over conventional differential display RT-PCR. In differential display RT-PCR, sequences are recognised by arbitrary primers. Here, more accurate restriction enzyme digestion was utilised and the *Escherichia coli* DNA ligase to divide the entire mRNA population into 64 non-overlapping but comprehensive subpopulations. This technique is in clear contrast to differential display RT-PCR which has no mechanism to avoid redundancy. PCR then proceeds under standard conditions without sensitivity to the conditions of

Table 3.7 Selected cDNAs displaying altered expression in the Alzheimer’s disease-affected brain. All sequences are matched to GenBank entries. ‘Enzyme’, ‘Adapter’ and ‘Primer’ indicate the enzyme used for the first enzyme digestion, the sequence of the cohesive end of the adapter, and the anchor base of the oligo(dT) primer, respectively. ‘Source’ indicates RNA used for fragment cloning. Alzheimer’s disease/control ratio of each expressed cDNA was calculated based on the results derived from ATAC-PCR experiments performed with four Alzheimer’s disease (AD) brains and four control (C) brains. The following abbreviation is used: SD, standard deviation.

Number	Enzyme	Adapter	Primer	Size (base pairs)	Gene name	Source	AD/C mean	Ratio SD
1	<i>Bsm</i> AI	NATA	C	115	N-chimaerin	C	0.11	0.03
2	<i>Bsm</i> FI	NTCC	C	194	mRNA for OTK27	C	0.22	0.08
3	<i>Bsm</i> AI	NTTT	A	93	KIAA0099 (fruit fly pumilio gene-related)	C	0.23	0.11
4	<i>Bsm</i> FI	NATA	G	257	Mitochondrial consensus sequence	C	0.25	0.19
5	<i>Bsm</i> AI	NCGT	G	249	GTPase-activating protein-1 (GAP1)	C	0.28	0.15
6	<i>Bsm</i> AI	NCAT	G	209	G protein-coupled receptor	C	0.29	0.12
7	<i>Bsm</i> AI	NCAA	A	181	Phosphatidylethanolamine binding protein	C	0.29	0.03
8	<i>Bsm</i> AI	NACC	C	75	Lysosomal-associated membrane glycoprotein-1	C	0.32	0.06
9	<i>Bsm</i> AI	NCGT	G	248	Human homologue to yeast ribosomal protein L41	C	0.41	0.14
10	<i>Bsm</i> AI	NGTA	A	152	β-catenin	C	0.43	0.22
11	<i>Bsm</i> FI	NACT	G	225	Secreted cyclophilin-like protein	C	0.47	0.12
12	<i>Bsm</i> AI	NCGG	A	245	Mitochondrial trans and partial proteins 4 and 5	C	0.5	0.13
13	<i>Bsm</i> AI	NCAT	C	122	LLRep3	C	0.67	0.21
14	<i>Bsm</i> FI	NTAT	C	126	Arginine methyltransferase	AD	20.31	3.24
15	<i>Bsm</i> AI	NAGG	A	228	Apoferitin H chain	AD	10.5	2.94
16	<i>Bsm</i> AI	NCGC	A	249	Glial fibrillary acidic protein (GFAP)	AD	10.12	5.42
17	<i>Bsm</i> AI	NGCT	C	123	α-subunit of Gi2	AD	4.59	1.21
18	<i>Bsm</i> AI	NCCT	A	359	Water channel	AD	4.04	2.75
19	<i>Bsm</i> AI	NATA	C	268	Human BAC clone GS113H23 from 5p15.2	AD	3.36	0.38
20	<i>Bsm</i> AI	NCCT	G	146	Human <i>Cdc5</i> -related protein (PCDC5RP)	AD	3.03	0.26
21	<i>Bsm</i> AI	NATA	A	93	TI-227H	AD	2.98	0.11
22	<i>Bsm</i> FI	NAAT	C	148	Human uncoupling protein homologue (UCPH)	AD	2.75	0.23

thermal cycling. Reamplification is not necessary to recover the fragments, resulting in fewer pseudopositives.

However, multiple tedious experimental steps such as gel excision and sequencing are necessary to characterise genes. These types of techniques are useful with organisms possessing little known sequence information. For mammals, however, techniques based on such collections of sequenced genes (e.g., DNA microarrays, ATAC-PCR) may prove to be more efficient. Molecular indexing does provide the advantage of isolating previously unidentified genes.

Therefore, molecular indexing possesses several advantages over differential display RT-PCR, although it includes the complicated enzymatic process of adapter ligation with *Escherichia coli* DNA ligase. As this technique also shares some of the limitations of differential display RT-PCR, choice of the technique depends on consideration of these factors with each experimental system under examination.

3.3.8

MSP-PCR – Multiplex PCR using colour-tagged module-shuffling primers

3.3.8.1 Introduction

Multiplex PCR using colour-tagged module-shuffling primers (MSP-PCR) provides a simple means to compare with high fidelity mRNA expression patterns in different cells or tissues. The method is based on a polymerase chain reaction (PCR) that utilises specifically designed primer sets named module-shuffling primers.

3.3.8.2 Principle and basics

The amplification efficiency in PCR strongly depends on the nature of the DNA fragments in the sample. For this reason, comparison of reaction products requires the construction of internal standards or a calibration curve. In addition, the early stages of PCR are more suitable because the amplification reaction has not yet achieved saturation. Moreover, if the PCR of expressed genes derived from different sources is carried out in the same reaction tube, the PCR products can be directly compared with high reliability. However, distinguishing the original sources of the expressed genes is difficult because the sequences of targets are the same. The key to overcoming this difficulty is to design primers with the same melting temperature but with different sequences, thus permitting identification of the source of the expressed genes. The module-shuffled primers are designed to satisfy these requirements. When linked to multiplex PCR, this method can analyse several genes from various sources in a single reaction tube. Colour selection is used to determine the source of the gene, and size separation by gel electrophoresis is used to recognise the gene species.

The principle of the module-shuffled primers is illustrated in figure 3.70. Each module-shuffled primer is composed of six modules each of which in turn consist of three or four nucleotides. All the module-shuffled primers share the same modules, but in different order. Modules that have three nucleotides contain a cytosine at their 5' and 3'-termini; while modules with four nucleotides have thymine at

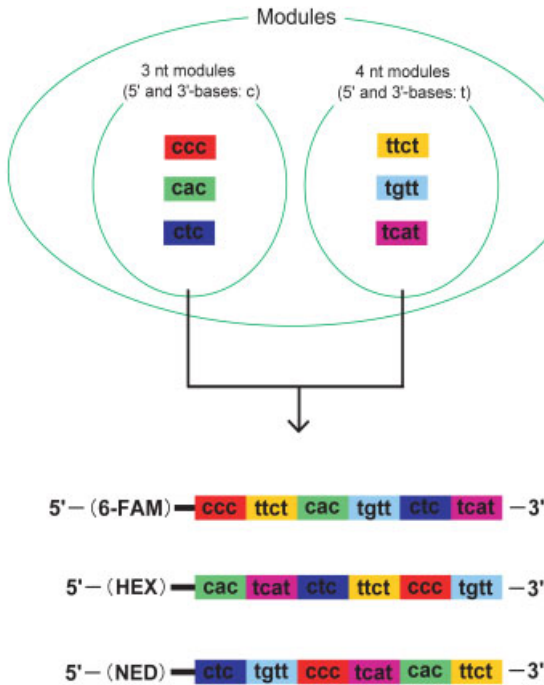


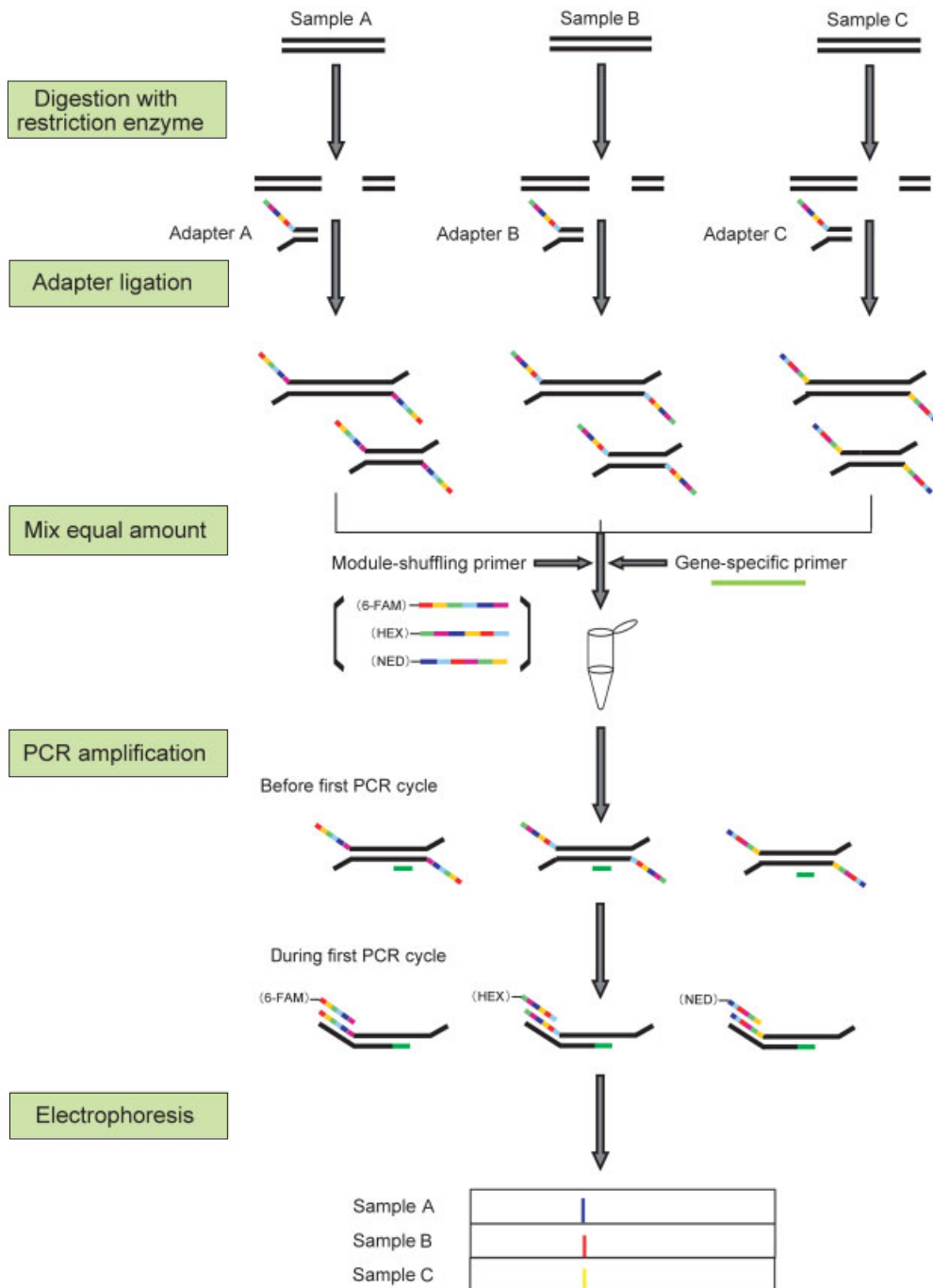
Figure 3.70 Structures of module-shuffled primers.

Module-shuffled primers are labelled with the fluorophores 6-FAM, HEX, or NED. The sequences of the modules are 5'-CNC-3' and 5'-TNNT-3' (where N is adenine, cytosine, guanine or thymine). Modules are arranged in a different order for each primer. Therefore, the sequence of each primer is unique, but their melting temperatures are identical. The source of each PCR product is determined by the fluorophore. The sequence of the primer in this example was MSP-A: 5'-(6-FAM)-**CCC** TTCT CAC TGTT CTC **TCA T**-3', MSP-B: 5'-(HEX)-**CAC** TCAT CTC TTCT CCC **TGT T**-3', and MSP-C: 5'-(NED)-**CTC** TGTT CCC TCAT CAC **TTC T**-3' (the modules are shown in bold letters). The following abbreviation is used in this figure: nt, nucleotides.

their 5' and 3'-termini. Since the module-shuffled primers assemble three-nucleotide modules and four-nucleotide modules in sequence, the sequences of each module-connecting part are either cytosine/thymine or thymine/cytosine. Because the module-connecting parts have this structure, the thermodynamic behaviour of the modules is identical because there is no difference in the neighbouring-base effect (Breslauer, 1986). The sequences of the module-shuffled primers differ due to the shuffling of the modules, but their melting temperatures are identical. For this reason, module-shuffling primers hybridise only with their complemen-

Figure 3.71 PCR using colour-tagged module-shuffled primers in the comparative analysis of expressed genes. mRNAs are converted to double-stranded cDNA, which are digested by a restriction enzyme with a four-base recognition site. Oligonucleotide adapters are ligated to the ends of the digested cDNA fragments. Priming by two module-shuffled primers is avoided by use of a Y-shaped adapter (for details, see figure 4.63; Prashar & Weissman, 1996). The sequences of the adapter oligonucleotides were A: 5'-CCC TTCT CAC

TGTT CTC TCAT CTG CGC TCA TAC CG-3', B: 5'-CAC TCAT CTC TTCT CCC TGTT CTG CGC TCA TAC CG-3', C: 5'-CTC TGTT CCC TCAT CAC TTCT CTG CGC TCA TAC CG-3', and L: 5'-(phosphate)-GAT CCG GTA TGA GCG CCA-(amino)-3'. A and L, B and L, and C and L adapters are able to form a Y-shaped duplex. Each module-shuffled primer reacts only with its respective adapter. Sizes, fluorescence wavelength, and fluorescent intensities of amplified products were monitored using a conventional DNA sequencer.



ary strands and have the same PCR amplification efficiency (Uematsu *et al.*, 2001). Labelling of the module-shuffled primers is by means of the fluorescent markers 6-FAM, HEX, and NED. A mixture of three module-shuffled primers is used for the PCR.

The procedure for comparing expressed genes from different sources using a PCR with module-shuffling primers (MSP-PCR) is shown in figure 3.71. The total RNA is converted to double-stranded cDNA using the conventional method (Gubler & Hoffman, 1983). Three kinds of cDNAs from different sources are separately prepared. The cDNAs are digested using a four-base recognition endonuclease, which produces DNA fragments averaging 256 base pairs. DNA fragments of this size can be detected using polyacrylamide gel electrophoresis (PAGE). Three kinds of oligonucleotide adapters corresponding to three module-shuffled primers are prepared. Each oligonucleotide adapter is introduced to the restriction fragments of cDNA using ligation. The procedures described above are carried out separately for each source. The adapter-ligated fragments from these different sources are then mixed and used as a template for the PCR. Each target cDNA fragment in the mixture is amplified using a primer pair consisting of one member of module-shuffled primer mix and a gene-specific primer. The module-shuffled primer discriminates between the sources of each amplified gene. The PCR products are analysed using a fluorescent DNA sequencer. Each fragment is identified on the basis of its electrophoretic mobility, each source is identified on the basis of the wavelength of fluorescent. A conventional DNA sequencer contains four colour-channels, one of which is used for detecting a DNA size marker, while three are used for the comparative analysis of the expressed genes.

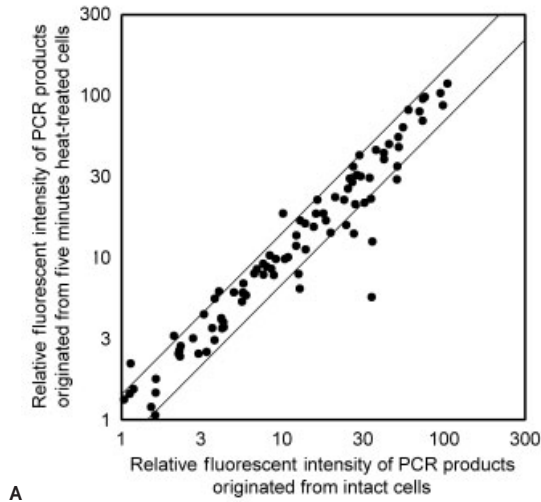
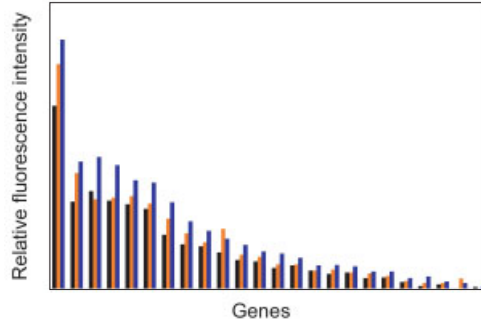
3.3.8.3 Typical results

Fluorescent signals from three fluorophores are measured with the appropriate compensation of fluorescence intensity for the fluorescence coefficient of each fluorophore. These compensated fluorescence intensities are proportional to the gene expression in the various sources. Typical results are shown in figure 3.72. The expression levels of 24 genes from three different sources were experimentally investigated using module-shuffled primer-PCR. The vertical axis shows relative gene expression, while the horizontal axis shows each gene individually. Comparison of relative expression levels is possible.

3.3.8.4 Example

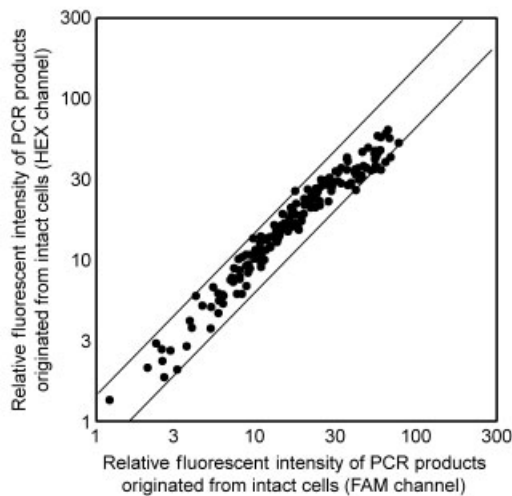
The expression levels of 100 genes in the FY1679 strain of *Saccharomyces cerevisiae* were measured. Intact cells were compared to cells heated to 42 °C for five minutes. The PCR products amplified from the intact cells were labelled with 6-FAM, and those from the heat-treated cells were labelled with HEX. Figure 3.73.A is a scatter plot of the relative fluorescence intensities obtained from the amplified genes expressed in the intact cells (*x*-axis) and the heat-treated cells (*y*-axis). The expression levels of ten genes were suppressed by more than 30 percent in the

Figure 3.72 Typical results of PCR using module-shuffled primers. The vertical axis shows the relative amount of expressed mRNA, and the horizontal axis shows each individual gene. Black, orange, and blue bars indicate the sources of the genes.



A

Figure 3.73 Comparison of mRNA expression. **A)** Scatter plot of relative fluorescent intensities of the amplified genes expressed in intact yeast (x-axis) and in yeast cells heated to 42 °C for five minutes (y-axis). **B)** Scatter plot of the relative fluorescent intensities of the same samples. The x-axis shows relative fluorescent intensities of the PCR products amplified with primer MSP-A and the y-axis shows relative fluorescent intensities of the PCR products amplified with MSP-B (see figure legend 3.70 for details on primers).



B

heat-treated cells compared to the intact cells, expression of five genes was increased in the heat-treated cells by more than 30 percent.

Figure 3.73.B is a scatter plot of the same samples in a control performed to test the accuracy of the plot shown in figure 3.73.A. The cDNA was aliquoted into two tubes. Each was ligated with a different adapter and the two cDNA aliquots were mixed together. Randomly selected genes were then amplified in one reaction tube using MSP-PCR. The difference in the fluorescent intensities of the amplified fragments with different ligated adapters was within 30 percent. Therefore, genes whose expression levels varied by more than 30 percent were detected by PCR using module-shuffled primers.

3.3.8.5 Discussion

PCR using module-shuffled primers eliminates many of the troublesome aspects of quantitative PCR such as the need for internal standards and a calibration curve. Genes of interest in different sources can be amplified and measured simultaneously, allowing simple and direct comparison. Because the amplification is carried out in one tube, the technique is reliable. It is of particular use in measuring rare mRNA species, because of its use of PCR. The method has a useful dynamic range of between 0.1 to 10.0 attomole in a 30-cycle reaction. Furthermore, combining it with a multiplex PCR is possible, because products of the same size from different sources can be distinguished.

However, PCR using module-shuffled primers has some weakness. Designing the primer that distinguishes between genes is more difficult than with conventional PCR. This is because only one gene-specific primer is used to distinguish between gene species, as opposed to the two primers in conventional PCR. Nevertheless, PCR using module-shuffled primers is a simple and reliable method for comparing differences in gene expression.

3.3.9

ODD – Ordered differential display

3.3.9.1 Introduction

Ordered differential display (ODD) (Matz *et al.*, 1997) belongs to a diverse group of methods that can be generally called ‘systematic differential display approaches’ (Matz & Lukyanov, 1998), many of which are also explained in this book. All these techniques, similarly to the very popular ‘classical’ differential display RT-PCR, or DDRT-PCR, approach (see chapter 4), rely on comparison of band patterns on gels to pinpoint cDNA fragments representing differentially expressed genes, but utilise radically different principles of pattern production. Rather than randomly picking cDNAs by means of PCR with short arbitrarily matching primers, systematic techniques attempt to accurately subdivide the whole cDNA population into a known number of non-overlapping subsets, which are of sufficiently low complexity to allow separation on polyacrylamide gels. The major advantage of this is that one can be completely sure that essentially all cDNA species have been screened to the limit of

the method's sensitivity. There are several more advantages of systematic techniques, and of ODD in particular, which for certain tasks compensate for the higher labour requirements compared to differential display RT-PCR.

3.3.9.2 Principle and basics

The principle of ODD is depicted in figure 3.74. The detailed protocol of the method is available in the internet at <http://www.ibch.ru/~lgr/protocols.htm>. Briefly, after synthesising double-stranded cDNA by any conventional technique (using only the ODD-cDNA synthesis primer for initiating first-strand synthesis), the cDNA is digested by a four-base-recognising restriction endonuclease. Two different endonucleases can be used with the standard ODD set-up described in the protocol, *Rsa* I (recognition site: GT↓AC) and *Hae* III (recognition site: GG↓CC). The choice of enzyme depends on the average GC-content of the cDNA of interest. Alternatively, both enzymes can be used together to produce 3'-cDNA fragments of less than one kilobase in length. A pseudo double-stranded adapter (Chenchik *et*

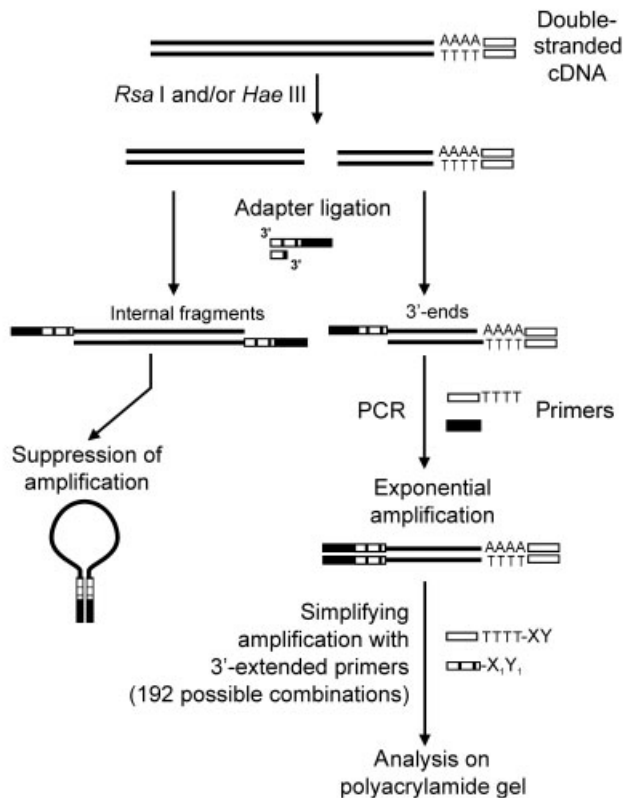


Figure 3.74 Schematic representation of the ordered differential display (ODD) protocol. See text for details.

al., 1996; Lukyanov *et al.*, 1997) is then ligated to the cDNA fragments, and PCR is performed using the cDNA synthesis primer and a primer corresponding to the distal half of the adapter. This evokes a PCR suppression effect (Siebert *et al.*, 1995), which prohibits amplification of the molecules that do not contain annealing sites for the cDNA synthesis primer. As a result, an amplified sample containing only 3'-restriction fragments of cDNAs is obtained. This amplification technique replaces the laborious and relatively inefficient procedures of physical separation of cDNA fragments, and obviates the problem of genomic DNA contamination (Lukyanov *et al.*, 1997). Most importantly, its use effectively lowers the requirement for initial amount of total RNA to 50 to 100 nanograms (Lukyanov *et al.*, 1997; Matz, 2002). Using conventional cDNA synthesis methods this amount is sufficient to produce a representative pool of 10^6 or more individual cDNA molecules as starting material for the ODD amplification. If fewer than 20 PCR cycles are required to produce a DNA concentration of ten to fifteen nanograms per microlitre, this indicates that at least 10^6 molecules were initially present (Lukyanov *et al.*, 1997; Matz & Lukyanov, 1998). It is very important to produce a representative sample of cDNA, since band intensity may fluctuate widely when underrepresented samples are used (figure 3.75.A). Providing 3'-ends are adequately represented in the cDNA sample, ODD is reproducible (figure 3.75.B).

To produce simplified subsets for comparison, adapter-specific primers extended by two bases are used for amplification. There are 16 possible extensions for the ligated adapter-specific primer and twelve for the oligo(dT) containing primer (four fewer because in this case thymine cannot be used in proximal position),

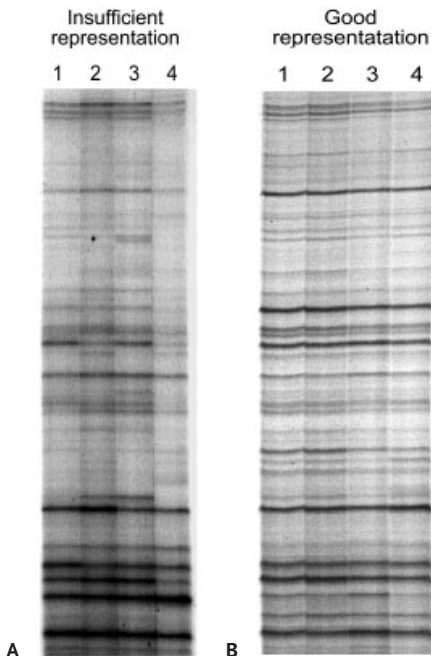
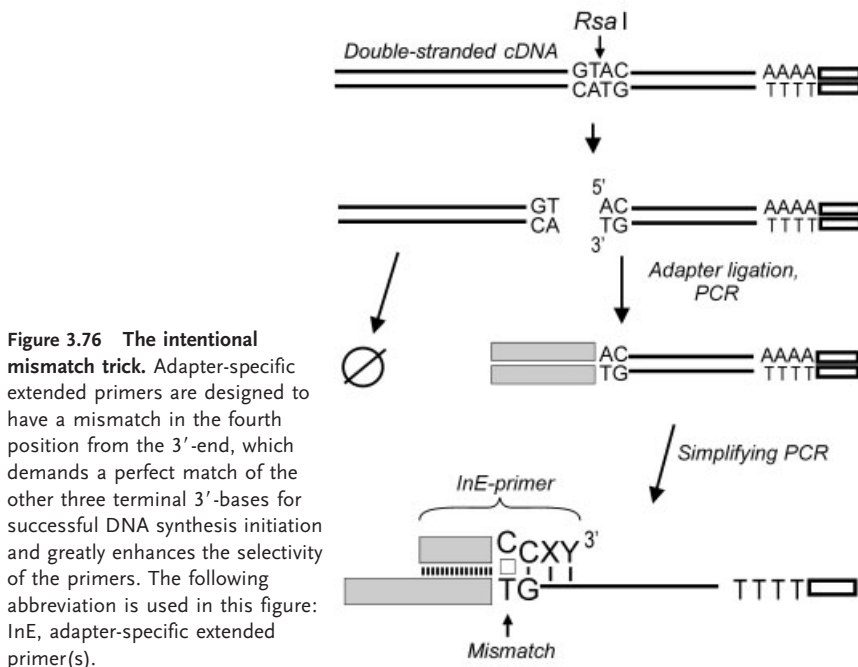


Figure 3.75 Effect of initial cDNA sample representation on ordered differential display (ODD) reproducibility. In both cases, the pattern of the same subset is shown for two sets of analogous cDNA samples. **A)** Complete quadruplicate (starting from RNA isolation) of growing blastema (regenerating tissue) of freshwater planarian *Dugesia (Girardia) tigrina*. A single blastema was used for each sample, which yielded cDNA amplified in more than 20 PCR cycles (underrepresented). **B)** Same as **A)**, but 20 blastemata were used to prepare each sample, yielding a representative cDNA sample amplified in 18 PCR cycles. Note that the pattern in **A)** is significantly more 'noisy' and therefore prone to produce false-positives. The increased sharpness of the bands in **B)** is actually because ^{33} phosphorus isotope was used instead of ^{32} phosphorus as in **A)**.

which gives a total of 192 combinations. Theoretically, a particular cDNA should be represented by a single 3'-end restriction fragment and be amplified with a single primer combination. Under such conditions, a cDNA pool of average complexity containing about 10^4 species should be subdivided into subsets containing about 50 fragments each, which is perfect for polyacrylamide gel analysis. However, in practice, ideal conditions are not met for two reasons.

First, first-strand cDNA synthesis is often primed by a oligo(dT) primer not only at the poly(A) region, but also at other adenine-rich stretches within the mRNA. For this reason, more than one 3'-end fragment may appear for such transcripts, increasing the complexity of amplified fragments by a factor of two or three compared to the original mRNA complexity. This is not a big problem, since overall complexity remains relatively low, and this situation also provides a further opportunity to encounter a differentially expressed sequence which might otherwise be missed due to masking by a redundant band in a particular subset pattern.

The second problem is that the adapter-specific extended primers are not absolutely specific. The selectivity of subset production is a critical factor in ODD, as in any differential display-like technique. If it is not high enough, the same abundant cDNAs are amplified in many subsets, leading to crowded band patterns, higher background and lower overall sensitivity. In ODD, the selectivity of adapter-specific extended primers (InE-primers) is increased by introducing intentional mismatch into the fourth position from the 3'-end (figure 3.76) (Matz *et al.*, 1997), which demands a perfect match of the three terminal 3'-bases for successful prim-



ing. This modification contributes greatly to primer selectivity (Okano *et al.*, 1998). Although no systematic study has yet been done, practical experience with ODD indicates that the 'sure detection' limit of ODD is at least as low as 10^{-4} (that means that each differential cDNA species should comprises one ten-thousandth of the total pool). The detection limit may be even lower since rarer cDNAs comprising 10^{-5} of total or less are often detected.

3.3.9.3 Examples and typical results

In addition to the author's works (Matz *et al.*, 1997 and 1998), five reports describing successful application of ODD have been published (Choi *et al.*, 2000 and 2001; Hirate *et al.*, 2001; Kim *et al.*, 1998 and 2001b). Table 3.8 summarises the data from these articles that are relevant in the technical respect. Unfortunately, these details are not provided in three of the five papers. On the other hand, only in one case the method was used in its full power, that is, for exhaustive search (Hirate *et al.*, 2001). However, in this work, the number of sequences confirmed to be truly differentially expressed (19) looks frustratingly small in comparison to the huge number of differential bands extracted and analysed (296). The reason for this most probably lies in the nature of the biological system that was studied (developing zebrafish brain). Brain, albeit embryonic, harbours the most diverse mRNA populations of all the tissues in an organism, which demands higher representation of the original cDNA samples for reproducible pattern generation and leads to crowded band patterns on ODD gels. The latter might be the actual cause why of the 296 isolated differential bands only 19 made it to the differential *in situ* hybridisation signals. The authors state that in their case each lane contained about 130 bands (this figure is about 80 in most other tissues; Mikhail Matz, unpublished observations) and there were persistent problems with many products found within a single band (Hirate *et al.*, 2001). The additional procedure of selecting between co-migrating products would naturally lower the success rate. In contrast, in a simpler system such as dorsal root ganglia (Kim *et al.*, 2001b), the bands represented unique products and the false-positives' rate was much lower – four false-positives out of 14 PCR fragments.

3.3.9.4 Conclusions

In conclusion, the advantages and disadvantages of ODD compared to other differential display-related techniques are briefly discussed (table 3.9). Apart from the absence of commercial ODD kits (a large practical disadvantage), the greatest drawback of ODD in comparison to differential display RT-PCR is that it requires more work. On the other hand, ODD is one of the easiest systematic display approaches. Also, the benefits of ensuring truly representative cDNA samples should not be underestimated. Although preparation of such samples might take some effort in the beginning, this high reproducibility of analysis ensures low false-positives rate, which may save a lot of time at the stage of selecting the true differentially distributed sequences among the ones isolated from the display gels. ODD may be re-

Table 3.8 Ordered differential display (ODD)-relevant technical data from the works in which ODD was applied.

References	Model (system/source of difference)	Number of primer combinations	Bands per gel/ total number of bands observed	Number of bands extracted	Confirmed to be differen- tially expressed (method)
Matz <i>et al.</i> (1997)	Planarian <i>Dugesia (Girardia) tigrina</i> /	4	80/320	6	5 (RT-PCR)
Matz <i>et al.</i> (1998)	Position along the antero-posterior axis				
Choi <i>et al.</i> (2000)	<i>Xenopus laevis</i> /early development	Single gene description, no technical details given			
Choi <i>et al.</i> (2001)	Same as above	Same as above			
Hirate <i>et al.</i> (2001)	Zebrafish embryonic brain/ LIM ^{1st-3} overexpression	192	130/26,000	296	19 (<i>in situ</i> hybridisation)
Kim <i>et al.</i> (1998)	Zebrafish/early development	Single gene description, no technical details given			
Kim <i>et al.</i> (2001b)	Rat dorsal root ganglia/static nerve injury	92	NA	14	10 (RT-PCR)

Table 3.9 Comparison of ordered differential display (ODD) with other differential display techniques.

<i>Pros</i>	<i>Contras</i>
Guarantees that the whole cDNA pool will be analysed	Relatively laborious (in comparison to differential display RT-PCR)
Designed for reliable use with microscopic initial amounts of RNA	3'-restriction fragments may not contain coding regions, which may prohibit gene identification
Improved sensitivity due to 'one cDNA/one fragment' principle and 1/192 simplification ratio	No commercial kits available
Requires little tune-up for non-standard models	

commended for cases in which the initial amount of RNA is low. Another disadvantage of ODD is that the 3'-restriction fragments displayed seldom contain open reading frames, making it impossible to identify the gene from the sequence of the isolated band. However, in the post-genomic era, this issue is becoming irrelevant for many models, as any cDNA fragment can be matched to the gene using genomic or expressed sequence tags information. In addition, most invertebrates do not have such long 3'-untranslated regions as vertebrates. Finally, ODD is easy to apply to non-standard models. The only modification that is necessary is choosing an appropriate restriction enzyme to produce fragments that are short enough. In contrast, in differential display RT-PCR a redesign of the whole set of arbitrary primers may be required in such circumstances.

3.3.10

PACS – Preferential amplification of coding sequences

3.3.10.1 Introduction

Identification of cell-type and phenotype-specific genes is essential to understanding the mechanism and role of a given cell in the context of a biological process. Indeed, comparison of gene expression between phenotypically different cells can provides important information in comprehending the phenotypes at the genetic level as well as the basis of many disease processes. Identification of genes (coding for mRNAs), however, is not an easy task. Since all genes coded by a genome do not contain universal identifiers, computer programs designed to identify them can only have limited success. Until very recently, the most favoured and widely used experimental methods for identifying genes in general, and identifying differentially expressed genes between two (or more) cell or tissue types in particular, always resulted in the acquisition of expressed sequence tags (ESTs) from the extreme 3'-noncoding region of mRNAs. This is because such techniques are critically dependent upon the 3'-poly(A) sequence (Pardee, 1989). Such ESTs provide little or no information about the coding region of a given gene and have other limitations regarding their utility in downstream applications. It can be speculated

that methods that can acquire sequence information from the coding region of mRNAs will be invaluable in subsequently identifying the cognate genes. Indeed, some methods have been developed that allow identification of both coding and noncoding regions of mRNA sequences, but these methods have not been widely used, presumably due to inherent technical complexity or a lack of evidence of much advantage over the standard methods, such as differential display RT-PCR (Caetano-Anolles, 1996; Matz *et al.*, 1997; Money *et al.*, 1996; Ralph *et al.*, 1993).

To accomplish the objective of rapidly identifying tumour-specific genes, a method was developed termed preferential amplification of coding sequences (PACS) that predominantly identifies coding region of mRNAs. PACS is achieved by performing PCR with a set of forward primers anchoring at sequences complementary to AUGs in mRNAs (targeting the methionine codon), and another set of primers based on restriction site sequences (much as described in the section on SuRF of this book) to anchor at a PCR-amplifiable distance from these AUG sequences (Fuchs *et al.*, 2000b; Weber *et al.*, 1998). PACS was initially applied to identify differentially expressed coding sequence tags (dCSTs) between cultured osteoblasts and osteogenic sarcoma-derived cells (Fuchs *et al.*, 2000b). Subsequently, the method was applied to identify dCSTs between healthy human breast tissue and a breast carcinoma cell line, T47D. PACS has been found to identify both up and down-regulated genes, but significantly, it identifies differentially expressed cDNA tags predominantly from the coding regions of mRNAs. Results suggest that PACS targets coding over noncoding sequences at a high ratio, approximately 13:1.

3.3.10.2 Principle and basics

3.3.10.2.1 Isolation of RNA

To develop PACS, mRNA expression was compared between FOB, an immortalised normal human fetal osteoblast cell line (Harris *et al.*, 1995), and MG63, an osteogenic sarcoma-derived cell line (Heremans *et al.*, 1978). Total RNA is extracted from these cell lines as described in the section presenting SuRF and quantified by measuring absorption in a ultraviolet spectrophotometer. The integrity of the RNA samples is evaluated and contaminating DNA is then removed by DNase I treatment using a procedure described previously to avoid unwanted amplification in the subsequent PCR step (Sanyal *et al.*, 1997).

3.3.10.2.2 PACS methodology

A scheme illustrating the principle of the PACS procedure is shown in figure 3.77. In a first step, PCRs are conducted in duplicate tubes using appropriate cDNA that is synthesised from total RNA using random primers (any standard protocol may be used). Routinely, two micrograms of total RNA can be used to synthesise cDNA with random hexamers in a final volume of 50 microlitres, one microlitre of which is used as template for PCR in PACS. The PCR is subjected to 45 cycles using a special set of primers described in detail in the next section. At the end of cycling, a stop buffer containing formamide, EDTA, bromophenol blue and xylene cyanol is

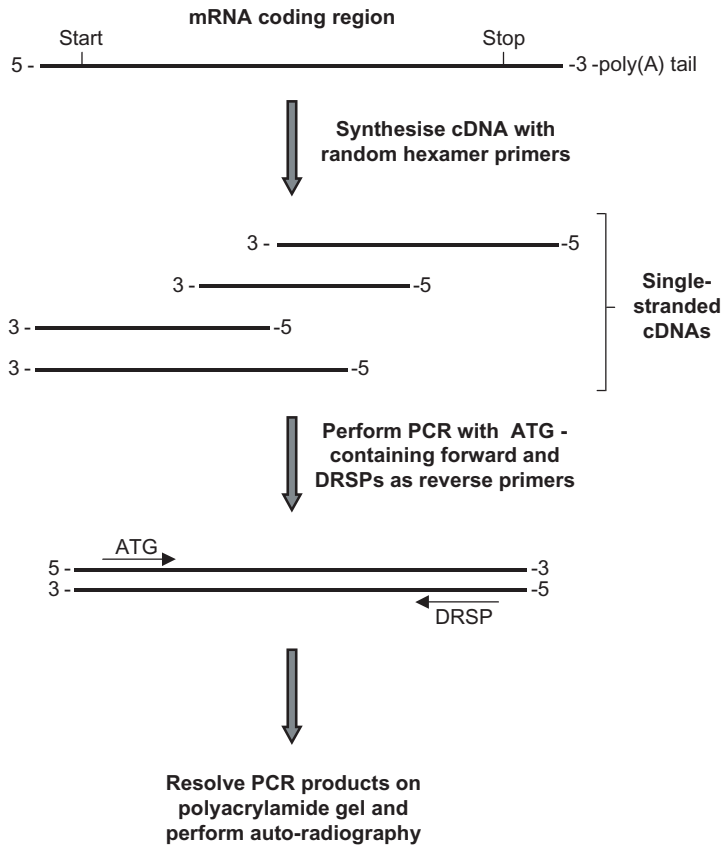


Figure 3.77 Schematic illustration of the preferential amplification of coding sequences (PACS) method. The method is described in detail in the text. The following abbreviation is used in this figure: DRSP, double restriction site primer.

added to the reaction mix and the whole sample is denatured by boiling. Three microlitres of this denatured PCR product are then electrophoresed on a six percent polyacrylamide gel containing urea. After electrophoresis, the gel is dried and subjected to auto-radiography.

3.3.10.2.3 Extraction, cloning and sequencing of gel-embedded DNA

These steps are described in detail under the section about subtracted RNA fingerprinting (SuRF) in this book.

3.3.10.2.4 Rationale for PACS

Considering that the poly(A) tail is part of the 3'-noncoding or 3'-untranslated region (UTR) of mRNAs, this is the only characteristic which distinguishes coding from noncoding regions of an mRNA. Thus, it can be assumed that a primer not based upon the poly(A) sequence will, at random, recognise coding and non-

coding sequences with similar frequency, exhibiting no preference for the coding region. If the length ratio of coding to noncoding sequences in mRNAs is about one, then with a random distribution of 'ATG'-seeking primers, 50 percent of acquired sequences would be expected to originate from the coding region by a method such as PACS as opposed to zero percent with the standard differential display RT-PCR method (Frohman *et al.*, 1988). Any shift in the ratio would either elevate or lower the frequency of sequences representing the coding region. Therefore, a key element of PACS was to design primers to anchor at ATG-containing sequences.

To assess the ratio of coding *versus* noncoding sequence of mRNAs, ten human mRNAs, with easily identifiable coding and 3'-noncoding sequences (just upstream of the poly(A) tail), were randomly selected. This ratio was determined to be about 1.5 coding *versus* noncoding sequence (Fuchs *et al.*, 2000b). If this ratio reflects the distribution of coding and 3'-noncoding sequences in general, then preferential acquisition of coding sequences with a non-poly(A)-based anchoring primer should be possible. In practice, not only was it possible to skew the ratio toward preferential acquisition of coding sequences, but it was also found that experimental observation greatly exceeded the theoretical expectation. A schematic illustrating the steps required for PACS is presented in figure 3.77.

3.3.10.2.5 Primer design

Although any triplet sequence can be used as a target for the anchoring primer for PACS, it was decided to anchor an oligonucleotide primer at the methionine codon for a number of reasons. First, since 'ATG' is the initiation codon on mRNAs in virtually every organism, anchoring at the triplet should enable acquisition of coding sequences from the extreme 5'-terminus of an open reading frame (ORF) more often. Second, since the translation initiation codon is encoded by the same triplet nucleotides in all organisms, there is no need to design multiple primers to anchor at the codon; excluding exceptions such as non-AUG initiation codons used by some viral (Corcelette *et al.*, 2000), chloroplast (Hirose *et al.*, 1999), plant mitochondrial (Sakamoto *et al.*, 1997) and bacterial (Yang *et al.*, 1999) mRNAs. Finally, as this particular codon is not highly repetitive in ORFs, the probability of generating multiple sequences from the same region of an mRNA or, alternatively, generation of very small sequences, should be low. For PACS, four anchoring primers were generated whose general sequence is 5'-*Bam* HI-NNN ATG X-3' (where 'N' represents a completely degenerate position and 'X' is either adenine, cytosine, guanine, or thymine). Sixteen different double restriction site primers (DRSPs) were also generated to serve as reverse primers. The general structure of a DRSP is: 5'-*Eco* RI-NNN-four or six-base restriction site sequence-3' (see table 3.10).

In the forward primers (for example, primer 5'-GGA TCC NNN ATG A-3'), the last nucleotide (either adenine, cytosine, guanine, or thymine) serves to increase specificity of the primer by allowing DNA synthesis to proceed only from appropriate sequence contexts in an mRNA. Thus, all forward primers have seven contiguous 3'-nucleotides (such as ...NNN ATG A-3'). that should completely match

Table 3.10 Sequences of double restriction site primers (DRSPs) used for preferential amplification of coding sequences (PACS). N is adenine, cytosine, guanine, or thymine.

<i>Eco</i> RI-NNN- <i>Alu</i> I	<i>Eco</i> RI-NNN- <i>Mse</i> I	<i>Eco</i> RI-NNN- <i>Aat</i> I	<i>Eco</i> RI-NNN- <i>Nco</i> I
<i>Eco</i> RI-NNN- <i>Dpn</i> I	<i>Eco</i> RI-NNN- <i>Msp</i> I	<i>Eco</i> RI-NNN- <i>Bam</i> HI	<i>Eco</i> RI-NNN- <i>Pst</i> I
<i>Eco</i> RI-NNN- <i>Hae</i> III	<i>Eco</i> RI-NNN- <i>Rsa</i> I	<i>Eco</i> RI-NNN- <i>Bst</i> UI	<i>Eco</i> RI-NNN- <i>Sac</i> I
<i>Eco</i> RI-NNN- <i>Hha</i> I	<i>Eco</i> RI-NNN- <i>Taq</i> I	<i>Eco</i> RI-NNN- <i>Hin</i> dIII	<i>Eco</i> RI-NNN- <i>Xba</i> I

with a template. The restriction enzyme sequence appended at the 5'-end can facilitate cloning of the amplified DNA in a desired vector, if necessary (Sarkar *et al.*, 1993; Weber *et al.*, 1998). Whether this restriction sequence is required for PACS has not been investigated. However, it can be speculated that the six-nucleotide *Bam* HI recognition sequence can be replaced by virtually any hexanucleotide sequence without affecting the performance of PACS.

3.3.10.3 Typical results, examples and discussion

PACS was carried out to identify dCSTs associated with osteogenic sarcoma and breast cancer. dCSTs associated with a disease fulfil the criteria for candidate molecular markers. To identify candidate molecular markers for breast cancer, PACS was performed to compare RNA from normal breast tissue with RNA from the breast cancer cell line T47D. A representative differential display pattern between these two samples obtained after performing PACS is presented in figure 3.78. In this case, the forward primer (anchoring primer) 5'-*Bam* HI NNN ATG A-3' and a DRSP with *Msp* I sequence at the 3'-end were used for PCR in duplicate tubes. Amplified material from each tube was electrophoresed in duplicate lanes on a sequencing gel. Potential differentially expressing bands are marked with upward triangles in the figure. Routinely, more than 200 amplified bands were obtained with one primer pair. Assuming that each of these amplified bands represents an mRNA, then one primer pair screens 200 mRNAs. So far, 16 different combina-

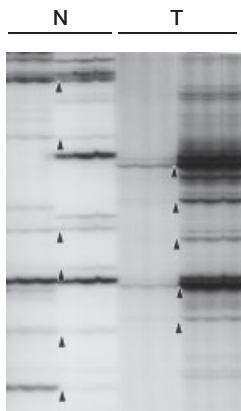


Figure 3.78 Identification of differentially expressed mRNAs by preferential amplification of coding sequences (PACS). Upward triangles indicate potentially differentially expressed bands. PCR was carried out in duplicate and a sample from each tube was loaded in triplicate in the sequencing gel. After electrophoresis, the gel was dried and processed for auto-radiography. Exposure time was about 24 hours. The following abbreviations are used in this figure: N, RNA from normal breast tissue; T, RNA from cells of the human breast cancer cell line T47D.

tions of primers were used covering an estimated pool of 3,000 mRNAs (this is about 20 to 25 percent of all mRNAs expressed by a cell assuming a cell type expresses between 10,000 to 15,000 mRNAs).

To identify candidate molecular markers for osteogenic sarcoma, PACS was performed between FOB and MG63 cell lines. For this part of the investigation, 64 different combinations of primers were used for PACS, virtually accomplishing 100 percent screening of the two cell lines for identifying dCSTs. This estimate of screening is realistic since redundant sequences were found at a very low (smaller than one percent) frequency (Fuchs *et al.* 2000b).

While acquisition of breast cancer-associated candidate molecular markers are ongoing, identification and sequence determination of virtually all osteogenic sarcoma-candidate molecular markers by PACS was completed (a detailed analysis of these sequences will be presented elsewhere). In general, inserts from two to five clones corresponding to a differentially expressed band (as seen on a sequencing gel) were sequenced. In the majority of the cases, all sequences produced from a differentially expressed amplified DNA band were as expected. In some cases, different sequences were obtained from a given band indicating that the amplified DNA corresponding to the band actually contained more than one species of DNAs. One very satisfying observation was that more than 95 percent of the sequences show the presence of the 'ATG' triplet sequence at the expected location suggesting that the anchoring primers lived up to expectation. At least five percent of the sequences obtained represent mitochondrial sequences. This is expected since a large proportion of total cytoplasmic RNA contains RNA from mitochondria. A few sequences each for α -satellites, cloning vector and ribosomal RNAs were also obtained. The sequences remained after excluding repeated, vector and rRNA sequences could be classified in three general categories: sequences representing known genes (category 1), sequences not yet known to represent genes but a completely matched sequence is present in the GenBank database (category 2), and sequences for which no matched sequence exists in GenBank (category 3).

It was observed that more than 95 percent of the category 1 sequences originated from the coding region of mRNAs strongly suggesting that coding sequences are indeed more abundant than 3'-noncoding sequences (in human mRNAs) and that, given such a distribution pattern, PACS will predominantly identify coding *versus* noncoding sequences. Thus, the frequency of identification of coding sequence tags by PACS was found to be much higher than expected. It is not immediately clear why this occurred, although there are several plausible explanations. First, even if an anchoring primer anneals to 3'-noncoding sequences, the DRSP might not find a suitable location to produce a PCR-amplifiable template. For example, if an anchoring primer binds to a location 100 bases upstream from the end of the 3'-UTR, a DRSP will have to anneal to a sequence within the 100 bases to initiate PCR. Since, no differentially amplified segment as small as about 100 base pairs was processed, they are not represented in the database of sequences. Second, 'ATG' might occur with greater frequency within the coding regions compared with the 3'-UTRs of mRNAs. Finally, because of small sample size, it is estimated

that coding regions are about 1.5 times longer than the 3'-UTRs and might not reflect the actual distribution of these segments (Fuchs *et al.*, 2000b).

To date, sixty differentially expressed breast cancer-associated sequences identified by PACS are analysed. Analysis of these sequences reveal a pattern closely comparable to osteogenic sarcoma-associated sequences described above.

3.3.10.4 Multi-PACS and coupled PACS and SuRF (PACS-SuRF).

In many situations, multiple samples need to be examined to identify desired gene(s). Osteogenic sarcoma, for example, is clinically presented in multiple forms based on morphology and/or anatomical location of a tumour. Each of such forms may have unique genetic characteristic. To identify genes potentially associated with different forms of osteogenic sarcoma, each of the forms may be compared with a reference/control sample by PACS (or a related method). However, a pairwise comparison of samples implies requirement of considerable amount of labour and resources. To offer a solution to this problem, PACS was modified in such a way as to allow simultaneous comparison of multiple samples. This modification is referred to as multi-PACS (Fuchs *et al.*, 2001). Apart from providing speed and ease of comparing multiple samples in the search for dCSTs, the strategy greatly improves the frequency of identification of true dCSTs. This is because in multi-PACS, every other sample acts as a control for a test sample, thus minimising scoring of false-positives. Multi-PACS is now routinely used for identifying dCSTs.

Coupling of SuRF (subtracted RNA fingerprinting, see the corresponding section in this book) and PACS can offer yet another approach to increase the frequency of identification of true dCSTs between samples. The most important aspect of SuRF is that it eliminates common mRNAs from samples, thus enriching representation of low-abundant genes. Application of PACS to subtracted mRNAs (as in SuRF) should, in theory, result in very efficient identification of true dCSTs and should specifically be considered for identification of all genes expressed in a given sample. Preliminary results clearly indicate that the application of coupled SuRF and PACS meets the authors' expectations and appears to be a very powerful approach for identifying and retrieving differentially expressed genes from either cell lines or tissues.

3.3.10.5 Concluding remarks

By designing specific primers, to date, differential expression of about 100 dCSTs associated with osteogenic sarcoma were confirmed by RT-PCR or Northern blotting (Fuchs *et al.*, 2001, and unpublished observations). These results suggest that identification of true differentially expressed dCSTs by PACS has been more than 80 percent. It appears that there is no other method that identifies differentially expressed genes/dCSTs at such a frequency. Therefore, PACS could be considered a robust arsenal in the hunt for cell or tissue-specific genes.

PACS is not the only method that can obtain sequence tags from the coding region of mRNAs. Several methods such as RNA arbitrarily primed PCR fingerprinting (RAP-PCR) (Ralph *et al.*, 1993) and restriction fragment differential display (RFDD) (Money *et al.*, 1996) have been developed with the aim of acquiring coding sequence tags. These methods, however, have some limitations. First, none provides a clear rationale underlying selection of their 'random' primers. Second, they do not provide appropriate information on the total number of random primers necessary to target all expressed genes. Third, these reports lack information on the frequency of identification of sequence tags from the coding region *versus* the noncoding region of mRNAs. Moreover, for RAP-PCR, there is no information on the number of nested RAP-primers ('circularly permitted nested primers') needed to screen all the cDNAs synthesised by a cognate random primer. Requirement of separate cDNA synthesis with separate random primers is also a disadvantage of this method. In addition, RFDD and similar methods have other complex requirements such as double-stranded cDNA synthesis, restriction enzyme digestion of the double-stranded cDNA, purification of the digested double-stranded cDNA and ligation of linkers. Since PACS does not have any of these limitations, it can be considered the simplest method for identifying mRNAs and obtaining coding sequence tags.

The use of differential display RT-PCR and similar methods for identification of differentially expressed genes has been increasingly replaced by DNA array-based expression screening (Brazma & Vilo, 2000; Duggen *et al.*, 1999; Heller *et al.*, 1997; Weis, 1998), a practical and powerful approach for generating extensive databases of candidate molecular markers for a given cancer or tissue type. However, the potential success of DNA microarrays is directly associated with the number of known genes being used as probes. Clearly, the more comprehensive the database of candidate molecular markers, the more effective this technique will be in identifying 'informative' genes. Until all open reading frames encoding proteins are identified in the human genome, DNA array-based methods will have moderate to severe limitations. It is reported that, at present, the best gene-finding programs can identify protein-coding exons with about 80 percent accuracy (Brazma & Vilo, 2000). Therefore, it is difficult to predict when all the open reading frames in the human genome will be identified for use as probes in DNA array-based screening to identify differentially expressed genes in the most comprehensive fashion. In addition to identifying differentially expressed genes, PACS can be easily applied to identify virtually all genes expressed in a cell or a tissue. This can simply be done by cloning PCR-amplified DNA after performing PACS with a given set of primers and sequencing the inserts. Thus, use of PACS can speed up the task of identification of genes coded by the human genome and facilitate construction of human a 'whole transcriptome array' as a valuable tool to realise the fruits of functional genomics.

3.3.11

RAP-PCR – RNA arbitrarily primed PCR fingerprinting**3.3.11.1 Introduction**

RNA fingerprinting using arbitrary primers is a modification of the polymerase chain reaction (PCR). The method relies on the selective amplification of sequences that, by chance, are flanked by adequate matches to a primer whose nucleotide sequence is arbitrarily chosen (AP-PCR). It was used first for generation an information-rich and unbiased fingerprint of genomic DNA (Welsh & McClelland, 1990; Williams *et al.*, 1990). An adaption of AP-PCR was used to generate an RNA fingerprint for the detection and cloning of transcripts that are differentially expressed between cells that have been subjected to different environments or developmental programs (McClelland *et al.*, 1994). The RNA fingerprints of two or more different RNA populations are compared by electrophoresis, differentially expressed genes are indicated by different patterns of bands.

3.3.11.2 Principle and basics

When using RNA as template for AP-PCR, the first step in fingerprinting is reverse transcription. Liang & Pardee (1992) were the first to describe an RNA fingerprinting protocol, using a primer for reverse transcription based on oligo(dT) but with an 'anchor' of one or two bases at the 5'-end (this is the first step of differential display RT-PCR).

Alternatively, an arbitrary primer can be used for reverse transcription (RAP-PCR), selecting those regions internal to the RNA that have six to eight base matches with the 3'-end of the primer. In contrast to differential display RT-PCR, in which an oligo (dT) primer is used for first-strand synthesis, this particular protocol samples anywhere in the RNA, including open reading frames (Welsh *et al.*, 1992). Second-strand synthesis is achieved by adding *Taq* DNA polymerase and the appropriate buffer to the reaction mixture. Once again, priming occurs at the sites where the primer finds the best matches. Therefore, second-strand synthesis is performed with low stringency (long annealing time and/or low annealing temperature), followed by high stringency PCR (see figure 3.79). The consequence of these enzymatic steps is the construction of a collection of molecules that are flanked at their 3' and 5'-ends by the exact sequence (and complement) of the arbitrary primer. Finally, the samples are directed to electrophoresis on a four or six percent polyacrylamide sequencing-type gel. For detection, samples can be labelled by [α -³⁵S]-dATP, [α -³³P]-dATP, or [α -³²P]-dATP, or fluorescence primers may be used for an automated comparison between lanes.

The arbitrary primers are chosen with several criteria in mind. First, the primers should not have stable secondary structure. Second, the sequence should be chosen such that the 3'-end is not complementary to any other sequence in the primer. In particular, palindromes should be avoided. Third, primers of ten to 20 nucleotides in length can be used. Longer primers can contain more sequence information designed to aid in subsequent steps in the experiment, such as cloning and sequencing.

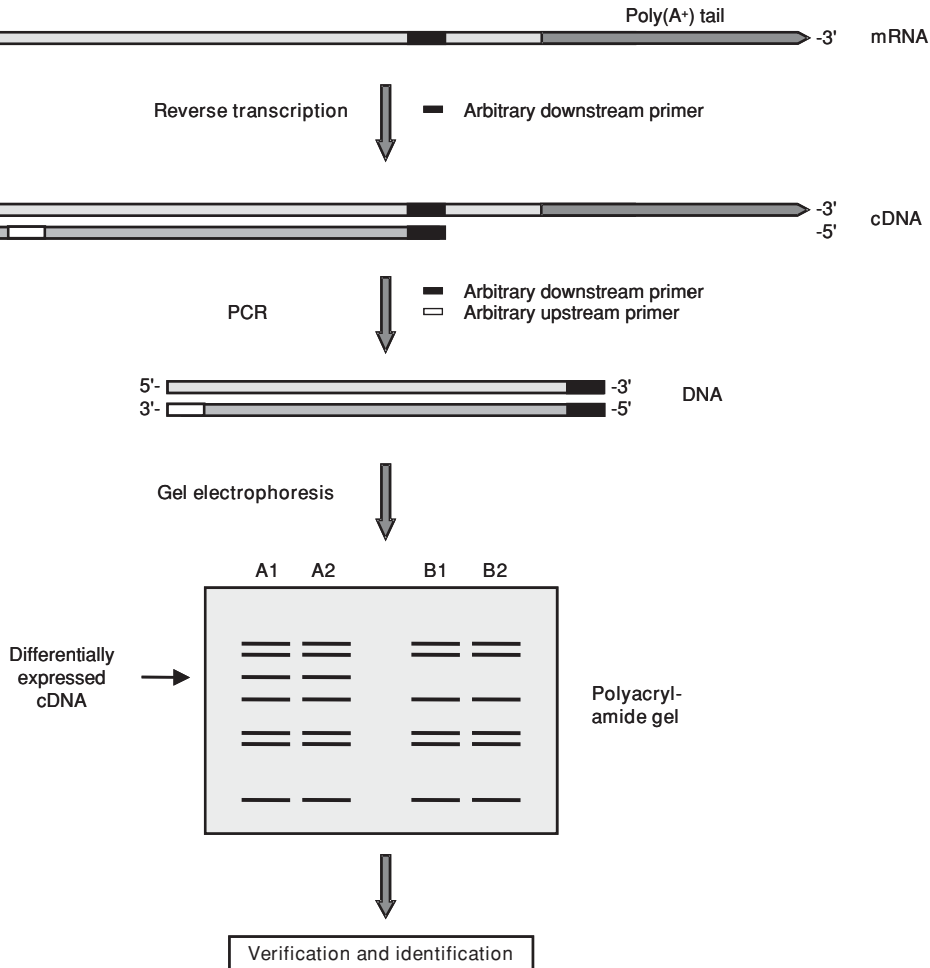


Figure 3.79 Schematic illustration of the RNA arbitrarily primed PCR fingerprinting (RAP-PCR) principle. An arbitrary primer is used to generate the first-strand cDNA. A second arbitrary primer is then used to synthesise the second-strand of DNA. Due to the low affinity of the arbitrary primers these steps are performed with low stringency (long annealing time, low

annealing temperature). There is a statistical chance that both arbitrary primers will fit to the same RNA in correct orientation. Therefore, a random pool of cDNA is generated out of an RNA population. This cDNA pool is then amplified by normal (high stringency) PCR using the original arbitrary primers.

3.3.11.3 Examples

RAP-PCR is a powerful tool for studying differential gene expression in, for example, cancer cells. Systematic analysis of human tumour samples may provide a list of markers with potential application to the diagnosis, prognostic assessment, and treatment of the disease. Applications of RNA fingerprinting by RAP-PCR to tu-

mours have included the identification of genes differentially expressed between normal and tumour cells in mammary epithelium (Liang *et al.*, 1992; Watson & Fleming, 1994) and ovarian epithelium (Mok *et al.*, 1994). Tortola *et al.* (1998) performed RAP-PCR under different conditions in human cell lines and in a series of 129 paired human normal colonic mucosa-colorectal carcinoma samples. They showed that RAP-PCR provides reliable fingerprints in a relatively wide spectrum of circumstances, including variations in RNA concentration and contamination by DNA. Serial analysis of paired normal/tumour cases yielded a number of bands that were recurrently either underexpressed or overexpressed in tumour relative to normal mucosa. These differentially expressed bands are prime targets of research because they represent candidate tumour-specific up or down-regulated genes with a relevant role in carcinogenesis (Tortola *et al.*, 1998). A further example is the detection of differentially expressed genes in renal cell carcinoma as shown in figure 3.80 (Vila *et al.*, 2000). If non-isogenic materials are being compared, such as clinical samples from different individuals, it is very important to fingerprint samples from a number of individuals. Any sequence polymorphisms between individuals are thereby eliminated as candidate differentially expressed genes.

Other genes cloned by RNA fingerprinting methods include mouse mammary tumour markers (Zhang *et al.*, 1993), a vitamin-induced gene in osteosarcoma (Kumar *et al.*, 1994), genes induced by exposure to radiation in a squamous carcinoma cell line (Jung *et al.*, 1994), genes induced by retinoic acid in myeloid cells (Burn *et al.*, 1994), and adipocytes (Dimaculangan *et al.*, 1994), and genes regulated by transforming growth factor β (TGF- β) and cycloheximide in epithelial cells (McClelland *et al.*, 1994; Ralph *et al.*, 1993).

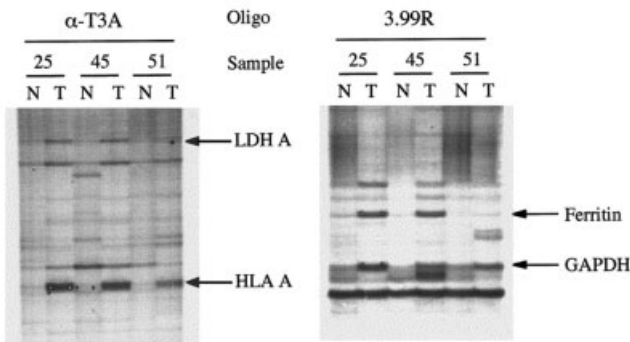


Figure 3.80 Results obtained by RNA arbitrarily primed PCR fingerprinting (RAP-PCR). RAP-PCR analysis of normal (N) and tumour (T) kidney tissues is shown. One microgram of total RNA from tissues (samples 25 and 45 corresponded to a clear cell carcinoma, whereas sample 51 was a chromophobe cell carcinoma) was used for initiate first-strand cDNA synthesis (reverse transcription) with two different primers (α -T3A and 3.99R). Second-strand cDNA synthesis was done by arbitrary priming PCR with the same oligonucleotides by using five microlitres of the first-strand cDNA mix. Arrows indicate differentially expressed PCR products analysed. The following abbreviations are used in this figure: LDH A, lactate dehydrogenase A; HLA A, human leukocyte antigen A; GAPDH: glyceraldehyde-3-phosphate dehydrogenase. The figure is reproduced from Vila *et al.* (2000) with permission from *Cancer*.

RAP-PCR was also used for analysing gene expression of differently stimulated bovine vascular smooth muscle cells (VSMC). These cells were stimulated either by platelet-derived growth factors (PDGF-AA and PDGF-BB), basic fibroblast growth factor (bFGF) or a nitric oxide donating drug (sodium nitroprusside) and RNA fingerprinting was compared to non-stimulated vascular smooth muscle cells using non-radioactive RNA fingerprinting and standard gel electrophoresis (Alfke *et al.*, 1999).

A very interesting application of the RAP-PCR method is RNA fingerprinting by using 'motif' primers for particular sequences that are conserved among a number of genes, such as the sequence conserved among serine/threonine protein kinases. Therefore, it is possible to focus the RNA fingerprinting to a specific class of genes with a common sequence motive. For example, new members of the zinc finger family were isolated by this variation of RAP-PCR (Stone *et al.*, 1994).

3.3.11.4 Discussion

The difference between differential display RT-PCR and RAP-PCR is the use of an arbitrary primer instead of an oligo(dT) primer in case of RAP-PCR. Therefore, RAP-PCR samples anywhere in the RNA, including open reading frames (Welsh *et al.*, 1992). In addition, it can be used on RNAs that are not polyadenylated, such as many bacterial RNAs (Wong *et al.*, 1994).

It is important to keep in mind that arbitrarily primed PCR methods including differential display RT-PCR are sensitive to template quality and concentration. Therefore, it is necessary to fingerprint a titration of the RNA. Only products that occur at two or more concentrations in one sample and not at all in the other sample need to be considered. Although RAP-PCR is a powerful tool to detect differentially expressed genes, it is necessary to perform control experiments to prove the results obtained by this method, for example, Northern blotting with the amplified and cloned cDNA used as a probe (Utans *et al.*, 1994). Alternatively, quantitative RT-PCR can be used. Other methods to confirm the authenticity of differentially amplified products have been reported (Callard *et al.*, 1994; Mou *et al.*, 1994).

PCR-based screening techniques show a high incidence of false-positive results (about 40 to 90 percent). In order to increase the efficiency and feasibility of RAP-PCR, the original protocol was modified by Bongrazio *et al.* (2001). The authors optimised the amplification cycle number, template amount, and concentration of primers, dNTPs and Mg^{2+} . Detection of fingerprints was performed by silver staining and samples were directed to sequencing using RAP-PCR primers. Confirmation of results by RT-PCR showed that the rate of false-positives attributable to this modified screening method was less than 20 percent (Bongrazio *et al.*, 2001).

A further limitation of the RAP-PCR is the complexity of a probe or the rarity of an RNA target. Some differentially expressed genes may not be detected. The vast majority of an RNA population of a given cell belongs to very few strongly expressed genes (*e.g.*, housekeeping genes). Due to this high background it is difficult to detect differences in RNA levels of weakly expressed genes. Technologies

based on subtractive hybridisation may be used alternatively for the identification of rare RNAs.

An advantage of RNA fingerprinting compared to subtractive hybridisation is that it can provide an estimate of the proportion of transcripts that respond to various developmental or environmental cues. Additionally, fingerprinting many RNA samples in parallel allows overlaps between the responses to different experimental treatments to be observed, indicating shared or antagonistic regulatory pathways. For example, lung epithelial cells were treated with transforming growth factor β (TGF- β). Over a period of many hours, various genes were repressed or depressed and RNAs were sampled at various times to observe the kinetics of regulation (McClelland *et al.*, 1994; Ralph *et al.*, 1993). Some cells were treated with cycloheximide and TGF- β . Cycloheximide blocks protein synthesis and therefore it was possible to classify responsive genes into primary and secondary response categories.

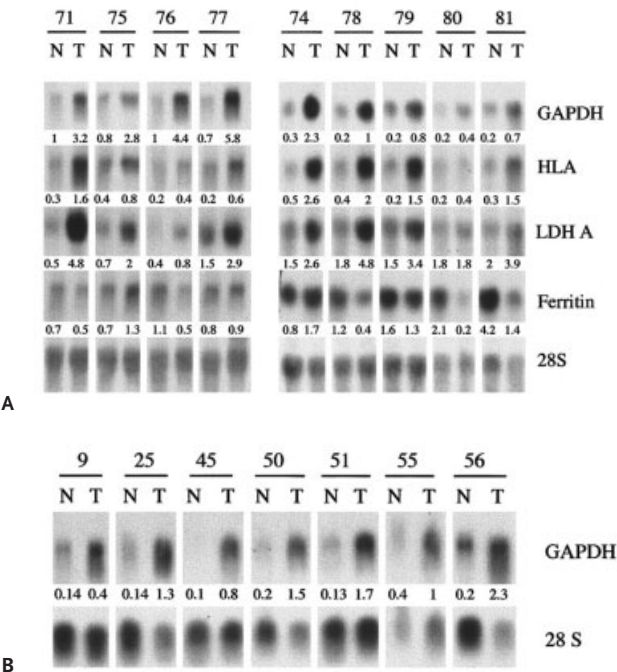


Figure 3.81 Verification of RNA arbitrarily primed PCR fingerprinting (RAP-PCR) results shown in figure 3.80. Northern blot analysis of differential genes on normal and tumour renal tissues on two different blots **A**) and **B**) is shown. Fifteen micrograms of total RNA were loaded onto each lane. RNA integrity and equal loading were determined by hybridisation with a 28 S ribosomal RNA probe. Numbers indicate the tissue samples used whose characteristics

are described in detail in Vila *et al.* (2000). Ratios of densitometric analysis are indicated below each lane. The following abbreviations are used in this figure: GAPDH, glyceraldehyde-3-phosphate dehydrogenase; HLA A, human leukocyte antigen A; LDH A, lactate dehydrogenase A; N, normal; T, tumour. The figure is reproduced from Vila *et al.* (2000) with permission from *Cancer*.

Such RNA fingerprints can yield much information even before the identities of responsive genes are established by cloning and sequencing.

3.3.12

Real-time reverse transcription polymerase chain reaction

3.3.12.1 Introduction

The real-time fluorescence-based reverse transcription polymerase chain reaction (RT-PCR) integrates the amplification and analysis steps of the PCR reaction, thereby eliminating the need for post-PCR processing. It does this by monitoring the amount of DNA produced during each PCR cycle and its sensitivity, specificity and wide dynamic range have revolutionised the approach to PCR-based quantification of RNA, making it the method of choice for quantifying steady-state mRNA levels (Bustin, 2000). Three key innovations make this technology possible. First, a process had to be developed that permitted the simultaneous amplification and detection of specific DNA sequences. The original system involved the addition of an intercalator, ethidium bromide, to the PCR reaction (Higuchi *et al.*, 1992) and the use of a thermal cycler with an ultraviolet light source and a computer-controlled charge-coupled device (CCD) camera (Higuchi *et al.*, 1993). A plot of increased fluorescence, caused by ethidium bromide binding to newly synthesised DNA, against the cycle number allowed the calculation of the quantity of starting material. The drawback of this approach was that this measurement could not easily distinguish between specific and non-specific products. The second innovation provided that specificity by adding a separate oligonucleotide probe, labelled with ^{32} phosphorus at its 5'-end and designed to hybridise within the target amplicon during the amplification. This generates a substrate suitable for the double-strand-specific 5'-3' exonuclease activity of *Taq* DNA polymerase (Holland *et al.*, 1991) and cleavage of the probe can occur only if the target sequence is being amplified. However, measurement of the ^{32} phosphorus-labelled cleavage products was not straightforward and required their separation by thin layer chromatography. The introduction of fluorogenic probes (Lee *et al.*, 1993) made it possible to amalgamate the two methodologies in the 5'-nuclease assay, the first practical real-time fluorescence-based quantitative PCR method (Heid *et al.*, 1996).

3.3.12.2 Principle and basics

There are currently five techniques that use fluorescent dyes and combine the processes of amplification and detection of an RNA target to permit the monitoring of PCR reactions in real-time during the PCR. The simplest method uses fluorescent dyes, for example, SYBR® Green, that binds specifically to double-stranded DNA. The other four rely on the hybridisation of fluorescence-labelled oligonucleotides to the correct amplicon, which obviates the need for post-PCR melting curves, Southern blotting or sequencing to confirm the identity of the amplicon.

3.3.12.2.1 DNA-binding dyes

This method (figure 3.82) involves detection of the binding of SYBR® Green to DNA (Morrison *et al.*, 1998). The unbound dye exhibits little fluorescence in solution, but during elongation increasing amounts of dye bind to the nascent double-

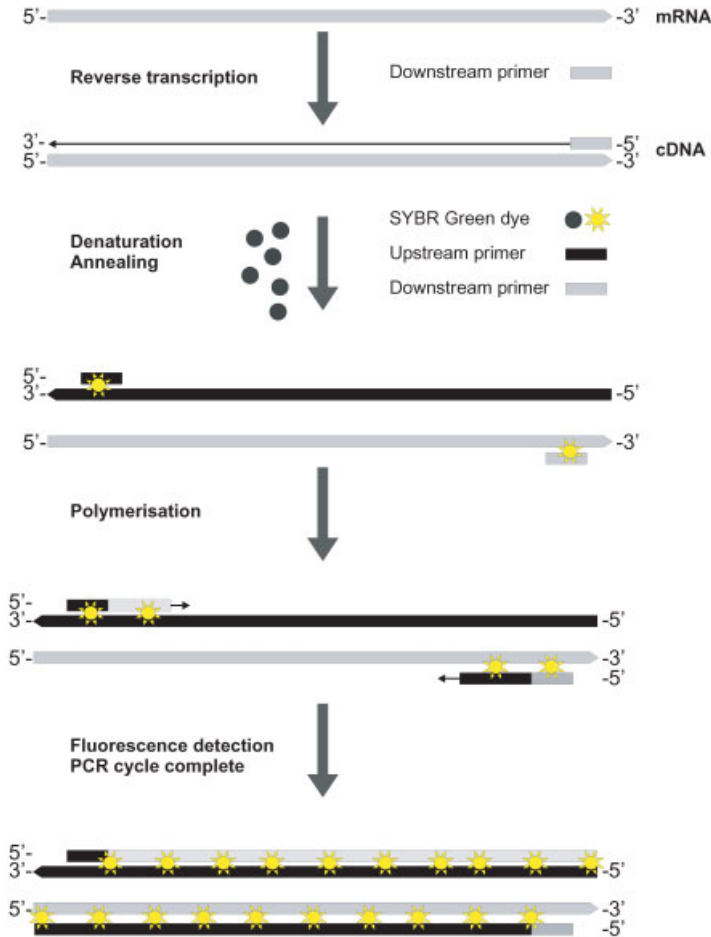


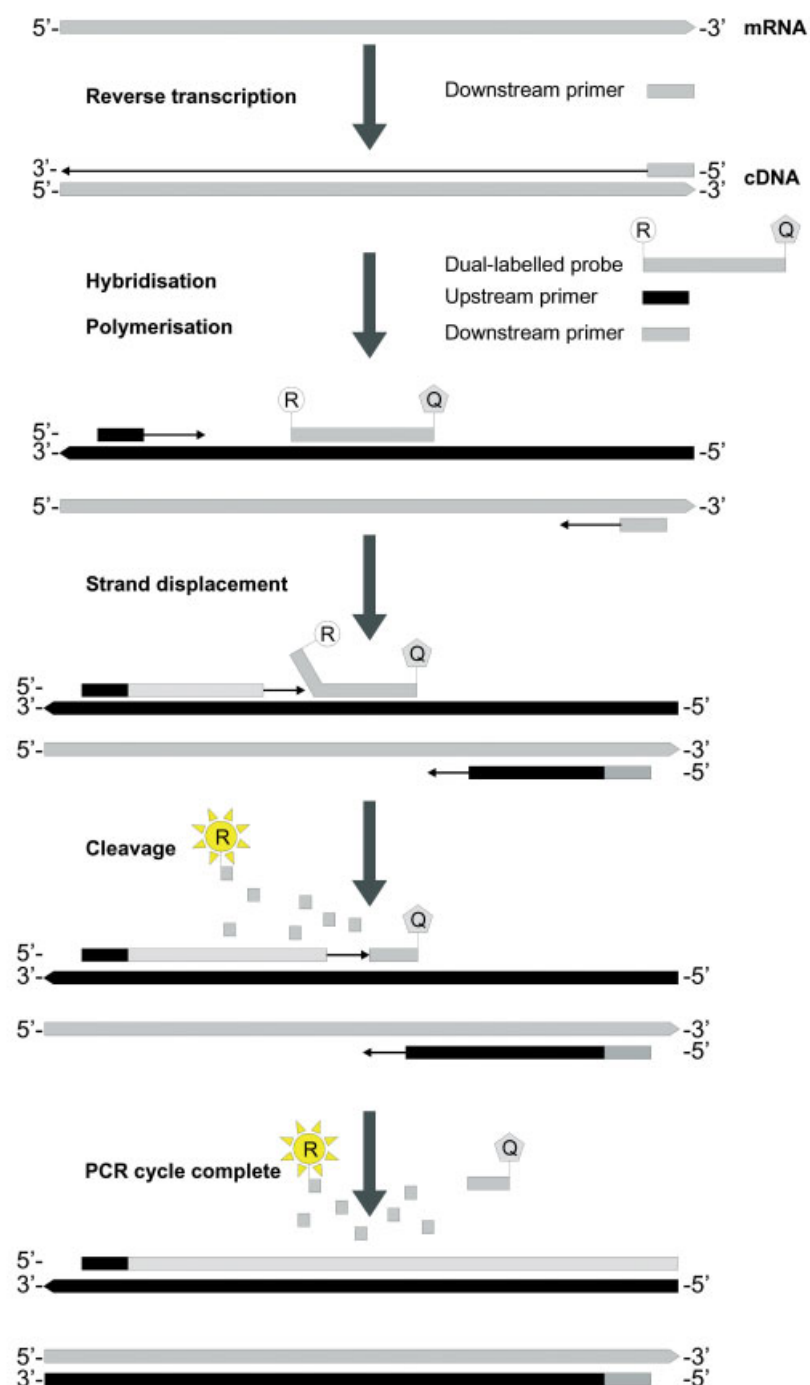
Figure 3.82 SYBR® Green real-time RT-PCR assay. A standard reverse transcription step is carried out, preferably using a specific downstream primer. During the denaturation step of the subsequent PCR, unbound SYBR® Green I dye exhibits little fluorescence. At the annealing temperature, a few dye molecules bind to the double-stranded primer/target hybrid, resulting in light emission upon excitation. During the polymerisation step, more and more dye molecules bind to the newly synthesised DNA, and the increase in fluorescence can be monitored

in real-time. Upon denaturation, the dye molecules are released and the fluorescence signal returns to background level. Monitoring during the early cycles, when the PCR amplification is exponential, provides precise data for accurate quantification. The amount of PCR product is measured directly in real-time using a fluorimeter and quantification is based on the threshold cycle, the first PCR cycle where the fluorescence signal first exceeded the level of background fluorescence.

stranded DNA. When monitored in real-time, this results in an increase in the fluorescence signal that can be observed during the polymerisation step, and that falls off when the DNA is denatured. Consequently, fluorescence measurements are performed at the end of the elongation step of every PCR cycle to monitor the increasing amount of amplified DNA. There are two problems inherent in this approach. First, since the specificity is determined entirely by the primers, this approach provides no more specificity than conventional RT-PCR. However, the PCR product can be verified by plotting fluorescence as a function of temperature to generate a melting curve of the amplicon (Ririe *et al.*, 1997). This is done by slowly elevating the temperature above the amplicon's melting temperature (T_m) and measuring the fluorescence. Since the melting temperature of the amplicon depends markedly on its nucleotide composition, it is possible to identify the signal obtained from the correct product. A characteristic melting peak at the amplicon's melting temperature will distinguish it from amplification artefacts that melt at lower temperatures in broader peaks. A second drawback is that multiple dye molecules bind to a single amplified molecule and consequently the amount of signal generated following irradiation is dependent on the mass of double-stranded DNA produced in the reaction. Assuming the same amplification efficiencies, amplification of a longer product will generate more signal than a shorter one. If amplification efficiencies are different, quantification will be even more inaccurate. This is in contrast to the use of fluorogenic probes, in which the signal from a single fluorophore is observed for each amplified molecule synthesised, regardless of its length.

3.3.12.2.2 The 5'-nuclease assay

This assay, popularly known as 'TaqMan™' (Applied Biosystems GmbH, Weiterstadt, Germany), has become the most widely used among the five chemistries currently available for real-time PCR (figure 3.83). Reactions are characterised by the point in time during cycling, the threshold cycle (C_t), when amplification of a PCR product is first detected rather than the amount of PCR product accumulated after a fixed number of cycles. Therefore, the threshold cycle is the key to understanding quantification by the 5'-nuclease assay. To determine the threshold cycle, data are collected from the first ten to thirteen PCR cycles and average fluorescence and its standard deviation are calculated. This establishes the level of background fluorescence for that particular run. Next, the algorithm multiplies the standard deviation of the background fluorescence by a default factor of ten to define a threshold. Finally, the algorithm searches the data for a point that exceeds the baseline by the value of the threshold. The cycle at which this point occurs is defined as the threshold cycle. Hence the threshold cycle represents a detection threshold for the instrument and it is dependent on the starting template copy number, and the efficiency of both PCR amplification and cleavage of the fluorogenic probe. The fewer cycles it takes to reach a detectable level of fluorescence, the greater the initial copy number. However, note that the choice of threshold is made by the operator and introduces a subjective element into real-time quantification.



The correct choice of DNA polymerase is important as not all possess the 5'-nuclease activity required to hydrolyse the hybridisation probe bound to its target amplicon (Kreuzer *et al.*, 2000). Most protocols utilise either *Taq* or *Tth* polymerase, but other polymerases, for example, *Tfl*, have been used successfully. Since the DNA polymerase will cleave the probe only while it remains hybridised to its complementary strand, the temperature conditions of the polymerisation phase of the PCR must be adjusted to ensure probe binding. Most probes have a melting temperatures of around 70 °C; therefore, the TaqMan™ system uses a combined annealing and polymerisation step at 60 °C to 62 °C, thus ensuring that the probe remains bound to its target during the primer extension step. It also ensures maximum 5'-3' exonuclease activity of the DNA polymerases. However, it reduces the processivity of the polymerases, which are required to function at sub-optimal temperatures. For longer amplicons, this may require a longer annealing/polymerisation step and/or increased Mn^{2+} or Mg^{2+} concentrations to stabilise probe binding. It also makes this system potentially less efficient and flexible than other real-time systems, especially when attempting mutation-specific RT-PCR.

The TaqMan™ probes carries a fluorescein donor at its 5'-end, which emits green fluorescent light when excited by a light source. Its emission spectrum overlaps the excitation spectrum of an acceptor fluorophore (quencher) that is attached to the 3'-end of the same oligonucleotide. Excitation of the 5-donor molecule results in fluorescence resonance energy transfer to the quencher and the emission of red fluorescent light without emission of green light. The TaqMan™ probe hybridises to the denatured DNA strand. During elongation of the PCR primers the TaqMan™ probe is cleaved by the DNA polymerase due to its 5'-3' exonuclease activity. The cleaved oligonucleotides, the first carrying the fluorophore and the last carrying the quencher, dissociate from the DNA strand and in solution, the two dyes are apart. Since the energy transfer depends on the spacing between the two dye molecules, the acceptor stops emitting red light due to the loss of the fluorescence resonance energy transfer. The emitted green light of the fluorophore now becomes detectable.

- ◀ **Figure 3.83 The 5'-nuclease assay.** Following the reverse transcription step, best carried out using the specific downstream primer, successful quantification requires the annealing of three oligonucleotides to the DNA. Two template-specific primers define the endpoints of the amplicon and provide the first level of specificity. Problems associated with DNA contamination are minimised by choosing intron-spanning primers. Additional specificity is provided by the use of a third oligonucleotide probe that hybridises to the amplicon during the annealing/extension phase of the PCR. This probe contains a reporter fluorescent dye (R) attached to the 5'-end and a quencher dye (Q) attached to the 3'-end nucleotide. When the intact probe is irradiated by a laser, the excited fluorescent dye transfers energy by Förster resonance energy transfer (FRET) to the nearby quenching dye molecule rather than fluorescing, resulting in a non-fluorescent substrate. During the polymerisation step of the PCR, when the polymerase replicates a template on which a probe is hybridised, the polymerase displaces and its 5'-3' exonuclease activity cleaves the probe. This separates the fluorescent and quenching dyes and FRET no longer occurs. Additional probes are hydrolysed during each PCR cycle and fluorescence increases in each cycle, proportional to the rate of probe cleavage, *i.e.*, the amount of amplicon produced.

3.3.12.2.3 Hybridisation probes

This method uses two hybridisation probes to maximise specificity (figure 3.84) (Wittwer *et al.*, 1997). One of the probes carries a fluorescein donor at its 3'-end, which emits green fluorescent light when excited by a light source. Its emission spectrum overlaps the excitation spectrum of an acceptor fluorophore (quencher) that is attached to the 5'-end of the second probe. This probe must be blocked at its 3'-end to prevent its extension during the annealing step. Excitation of the donor results in fluorescence resonance energy transfer to the acceptor and the emission of red fluorescent light. In solution, the two dyes are apart, and since the energy transfer depends on the spacing between the two dye molecules, only background fluorescence is emitted by the donor. Following the denaturation step, both probes hybridise to their target sequence in a head-to-tail arrangement during the annealing step. This brings the two dyes in close proximity to one another and the fluorescein can transfer its energy at high efficiency. The intensity of the light of longer wavelength emitted by the second dye is measured, with increasing amounts of measured fluorescence proportional to the amount of DNA synthesised during the PCR reaction. A fluorescent signal is detected only as a result of two independent probes hybridising to their correct target sequence. This increases specificity and generates additional flexibility for probe design. Furthermore, since the probes are not hydrolysed, fluorescence is reversible and allows the generation of melting curves.

3.3.12.2.4 Molecular beacons

Molecular beacons are DNA hybridisation probes that form a stem-and-loop structure; the loop portion of the molecule is complementary to the target nucleic acid molecule and the stem is formed by the annealing of complementary arm sequences on the ends of the probe sequence (figure 3.85) (Tyagi & Kramer, 1996). A fluorescent marker is attached to the end of one arm and a quencher is attached to the end of the other arm. The quencher is a non-fluorescent chromophore that dissipates the energy that it receives from the fluorophore as heat. In solution, free molecular beacons adopt a hairpin structure and the stem keeps the arms in close proximity, resulting in efficient quenching of the fluorophore. When molecular beacons encounter a complementary target at the annealing temperature, they undergo a conformational transition that forces the stem apart and results in the formation of a probe/target hybrid that is longer and more stable than the stem (Bonnet *et al.*, 1999). This separates the fluorophore and the quencher, leading to the restoration of fluorescence, which can be detected, whereas any free molecular beacons remain closed and non-fluorescent. If the target DNA sequence does not exactly match the molecular beacon sequence, hybridisation and fluorescence will not occur. This is because the thermodynamic properties of the molecular beacons favour the formation of a hairpin form rather than continued hybridisation to a less than perfectly matched target sequence.

The main drawback with molecular beacons is associated with the design of the hybridisation probe. Optimal design of the molecular beacon stem is crucial since the molecular beacon may fold into alternate conformations that do not place the

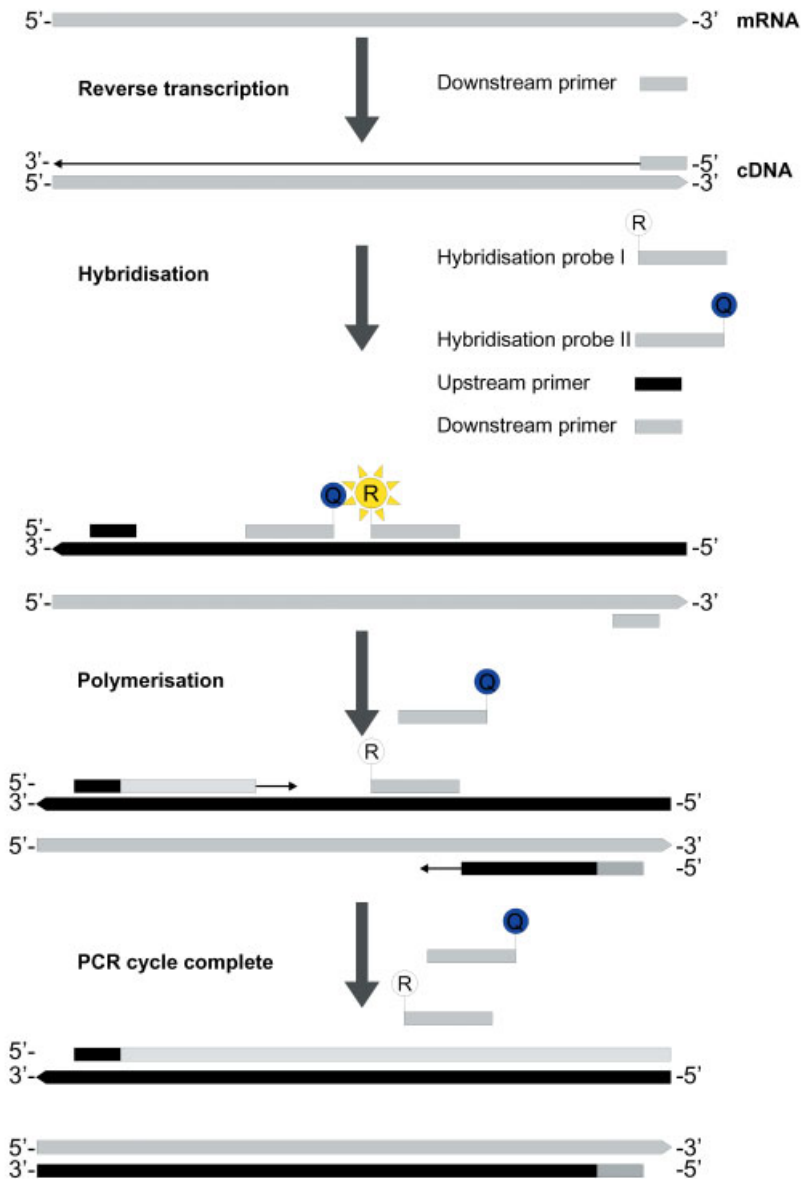


Figure 3.84 Hybridisation probes. A standard reverse transcription step is carried out, preferably using a specific downstream primer. During the denaturation step of the subsequent PCR, both hybridisation probes remain in solution and separate. Any emission from fluorescein is at 530 nanometres, and is disregarded by the detector. During the annealing step, the

probes hybridise in a head to tail arrangement, the two dyes come in close proximity and the emitted energy excites the second dye, which emits red fluorescent light at a longer wavelength. At the polymerisation temperature, both probes return into solution and any emissions from fluorescein are ignored.

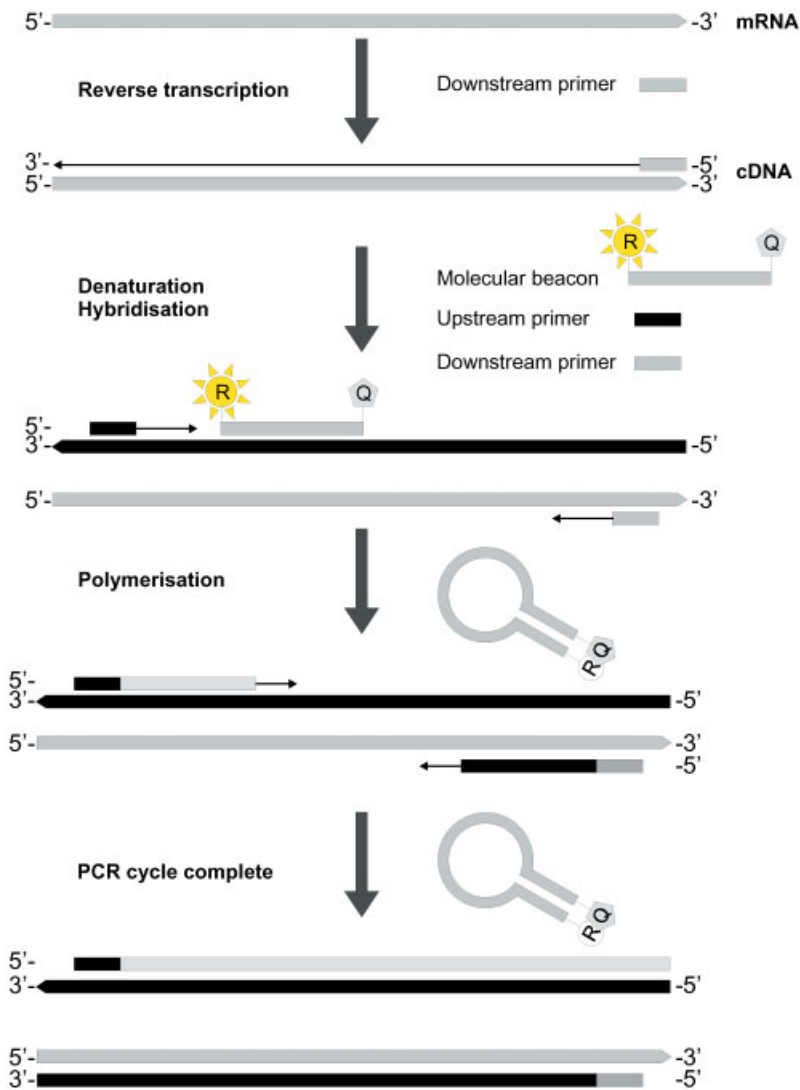


Figure 3.85 Molecular beacons. Molecular beacons are designed to be complementary to a sequence in the middle of the targeted amplicon. A standard reverse transcription step is carried out, preferably using a specific downstream primer. During the denaturation step of the following PCR, the molecular beacons assume a random coil configuration and fluoresce. As the temperature is lowered to allow annealing of the primers, stem hybrids form rapidly, preventing fluorescence. However, in

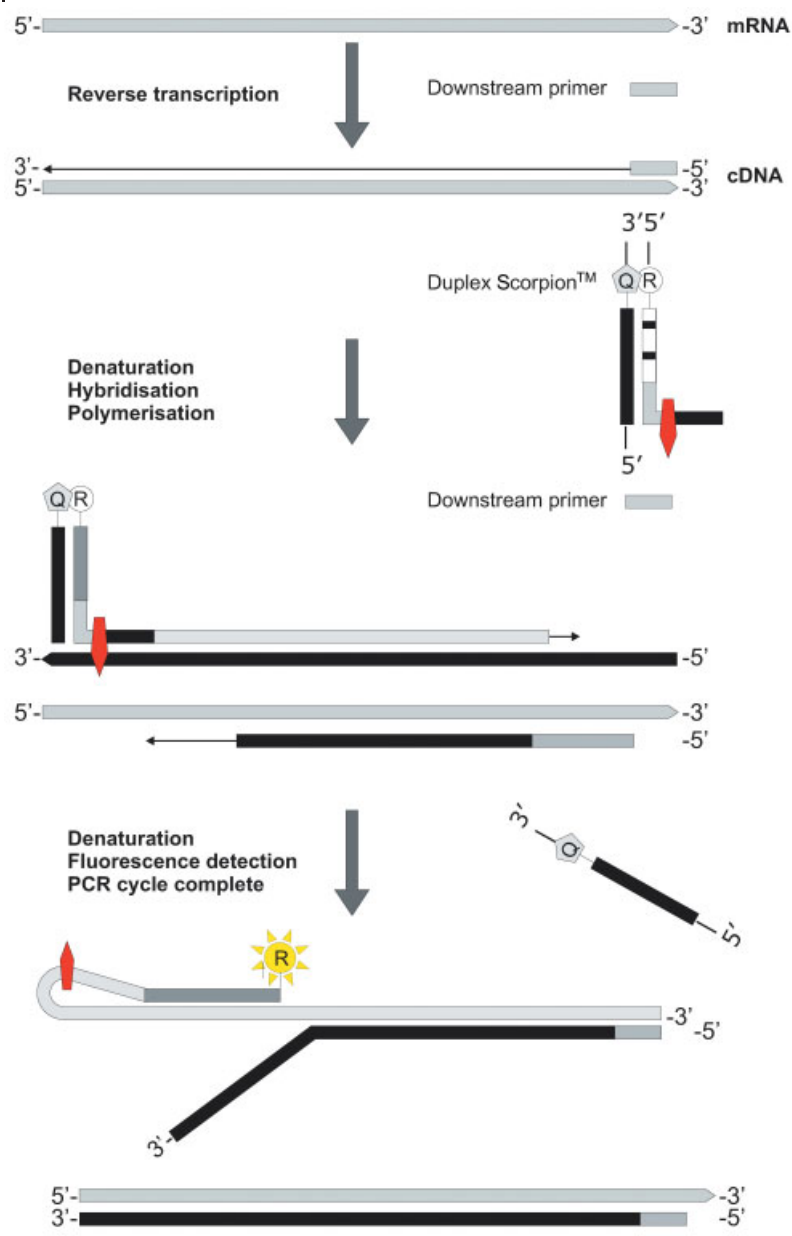
the presence of target, molecular beacons also bind to the amplicons and generate fluorescence. When the temperature is raised for the primer extension step, the molecular beacons dissociate from their targets, and fluorescence is again quenched. A new hybridisation takes place in the annealing step of every cycle, and the intensity of the resulting fluorescence indicates the amount of accumulated amplicon at the end of the previous cycle.

fluorophore in the immediate vicinity of the quencher, resulting in a sub-population that is not quenched well and in large background signals. Alternatively, if the stem of a molecular beacon is too strong, it can interfere with the hybridisation and the fluorescence of the beacon annealed to its target may be incompletely restored. Therefore, accurate thermal denaturation profiles to determine their melting characteristics have to be established for each molecular beacon and different target sequences will have to be matched with different stem sequences.

3.3.12.2.5 **Scorpions™**

Scorpions™ describe a novel chemistry that converts sequence-specific priming and probing into a unimolecular event, making the signalling reaction extremely fast (Whitcombe *et al.*, 1999). The most recent modification, termed duplex Scorpion™, contains the fluorophore and the quencher separated onto different, complementary oligonucleotides (figure 3.86). The fluorophore is attached to the 5'-end of an oligonucleotide that is made up of two separate sequences: (i) a probe sequence complementary to the extension product of the primer, (ii) a target-specific primer. Both are linked *via* a non-amplifiable monomer, typically hexaethylene glycol. This prevents copying of the probe sequences during the polymerisation step of the PCR reaction. The quencher is linked to the 3'-end of a second oligonucleotide that is fully complementary to the probe sequence. Following extension of the Scorpion™ primer, and after denaturation and reannealing, the specific probe sequence is able to hybridise to its complement within the same strand of DNA. Fluorescence is no longer quenched and a fluorescent signal is observed. This unimolecular hybridisation is kinetically more favourable than the quencher oligonucleotide reannealing with the Scorpion™ as it does not depend on a chance meeting between amplicon and probe, present at relatively low concentration. This allows the introduction of more rapid cycling conditions together with a significantly stronger signal strength compares with both TaqMan™ and molecular beacons (Thelwell *et al.*, 2000). Another advantage over TaqMan™ assays is that the PCR reaction is carried out at the optimal temperature for the DNA polymerase, rather than at the reduced temperature required for the 5'-nuclease assay to displace and cleave the probe. It also compares favourably with the high signal/high background ratio of the TaqMan™ and low signal/low background ratio of molecular beacons.

As with molecular beacons, the design of Scorpions™ requires consideration of secondary structure as well as primary sequence since it is crucial that another secondary event will not out compete the correct probing event. The use of secondary structure programs (*e.g.*, DNA mfold, Rensselaer Polytechnic Institute, 331 Amos Eaton Hall, Troy, New York, USA; available at <http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi>) helps, and it is likely that programs tailored to optimise Scorpion™ design will appear in the near future.



3.3.12.3 Typical results

The quantitative results shown in figures 3.87 and 3.88 were obtained using standard reaction conditions and a one enzyme/one tube assay design. This has the advantage of minimising any variation arising from a separate reverse transcrip-

- ◀ **Figure 3.86 Scorpions™.** A standard reverse transcription step is followed by denaturation of the double-stranded template during PCR amplification. Fluorophore and quencher are separated onto different oligonucleotides to improve signal intensity. The quencher oligonucleotide has the quencher at its 3'-end and is complementary to the probe sequence. The probe oligonucleotide is linked to the 5'-end of a specific primer through a PCR stopper that prevents read-through. During PCR, the Scorpion™ primers are extended to become part of the amplicon. During the annealing/extension phase of the PCR reaction, the probe sequence in the Scorpion™ hybridises to the newly formed complementary target sequence within the same strand of DNA and results in a fluorescent signal. As the tail of the Scorpion™ and the PCR product are now part of the same strand of DNA, the interaction is intra-molecular. The target sequence is typically chosen to be within three bases of the 3'-end of the Scorpion™ primer. Following denaturation and polymerisation, intra-molecular interaction of probe and newly synthesised product is more favourable than intermolecular binding between quencher oligonucleotide and the probe.

tion step and, from a practical point of view, requires the addition of reagents and template only once. Figure 3.87 shows typical results obtained when constructing a standard curve for absolute quantification of mRNA copy numbers and demonstrates the intra-assay reproducibility as well as the dynamic range of the assay. However, significant run-to-run variation can occur, especially when different operators, instruments or reagents are used. Repeated runs of the same standard curve give variations of two to three percent in the slope and ten percent in the y-intercept. Therefore, maximum reproducibility and precision is achieved only when running a standard curve with every set of unknown samples.

Figure 3.88 shows typical amplification plots obtained when quantifying numerous samples (in this case human colonic biopsies). Whilst most of the amplification plots conform to the expected shape, there are some that show reduced amplification efficiency during later cycles. Again this highlights the power of this assay and the advantage of using threshold cycles rather than endpoint assays.

3.3.12.4 Examples

There are now hundreds of publications that demonstrate the usefulness of real-time RT-PCR assays for quantifying steady-state mRNA levels. One of the most exciting applications concerns the ability to quantify mRNA from single cells, and the 5'-nuclease assay has been used successfully to quantify mRNA levels from individual tissue culture cells (Al Taher *et al.*, 2000) as well as from individual cells obtained from paraffin-embedded tissue (Sheils & Sweeney, 1999), laser-assisted cell picking (Fink *et al.*, 1998) and even immunostained tissue sections (Fink *et al.*, 2000). Its applications in pharmacogenomics are wide-ranging and the rapid and accurate detection of genetic polymorphisms has great potential for application to drug development, improvement of human clinical trials and monitoring surveillance for drug efficacy and toxicity (Shi *et al.*, 1999). It may have a role in medical diagnostics (Bustin & Dorudi, 1998) and has been used to detect micro-metastases (Bustin *et al.*, 1999) as well as viral (Hennig *et al.*, 2001) and bacterial (Norton & Batt, 1999) pathogens in patients' blood. Finally, it offers an alternative approach for detecting the presence of pathogens in the food industry (Kimura *et al.*, 1999).

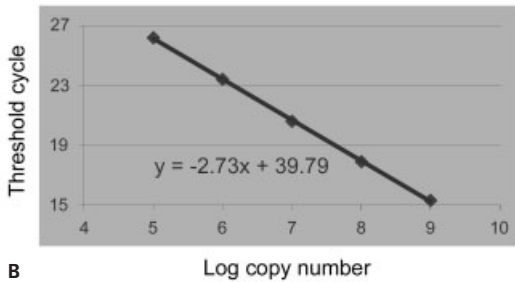
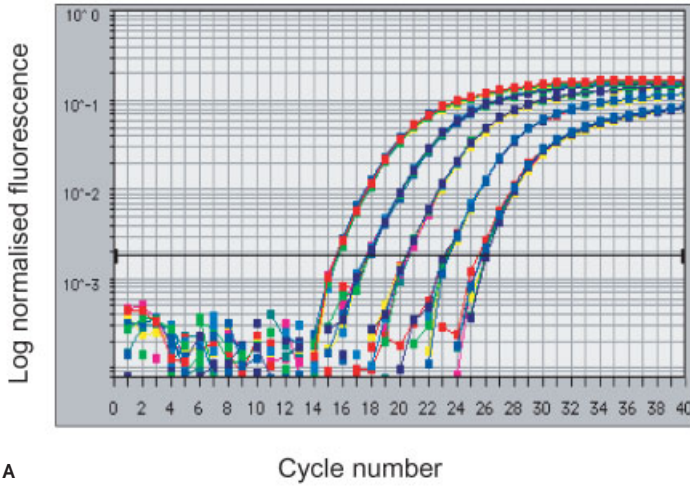


Figure 3.87 Examples of real-time RT-PCR assays. A) Amplification plot for a standard curve. Sense-strand oligodeoxynucleotides were ten-fold serially diluted from 1×10^9 down to 1×10^5 molecules and subjected, in triplicate, to real-time RT-PCR assays on an ABI® Prism 7700 (Applied Biosystems GmbH, Weiterstadt, Germany). The software calculates the threshold cycle (C_t) where the amplification plot crosses a defined fluorescence threshold (bold horizontal line). This provides several orders of linear dynamic range. The graph of normalised fluorescence *versus* cycle number during PCR has three stages. Initially, the fluorescent signal is below the detection limit of the sequence

detector. In the second stage, the signal can be detected as it continues to increase in direct proportion to the increase in the products of PCR. As PCR product continues to increase, the ratio of polymerase to PCR product decreases, product ceases to grow exponentially and the signal is roughly linear and finally reaches a plateau. **B)** A standard curve generated from the threshold cycles reveals the amplification efficiency of the reaction (slope) and gives some indication of its sensitivity (y-intercept). Given the threshold cycle of any unknown sample, its initial concentration can be interpolated from the standard.

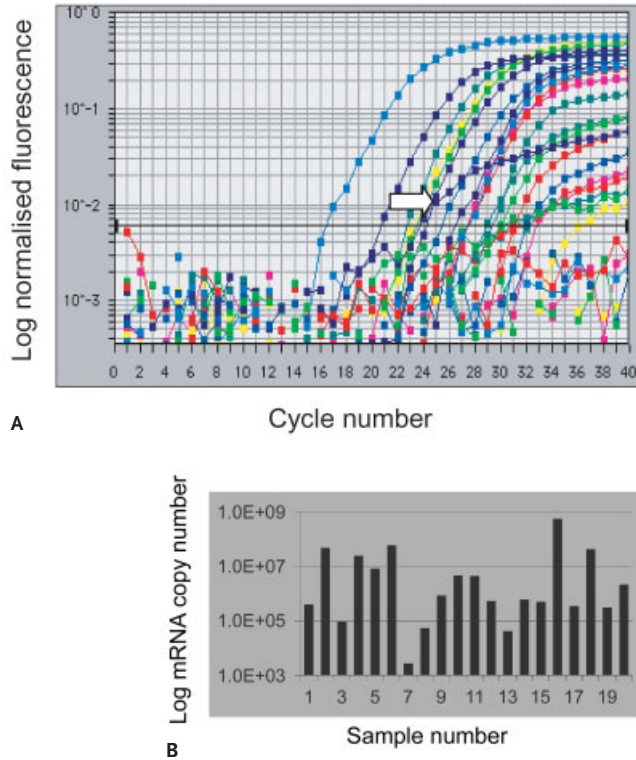


Figure 3.88 Real-time RT-PCR assays results. **A)** Amplification plots for insulin-like growth factor I (IGF1) mRNA levels obtained from 50 nanograms of colorectal cancer biopsies. The arrow points to a plot that becomes sub-optimal after cycle 24 and demonstrates

the enhanced reliability of measuring threshold cycles rather than end points. **B)** The absolute copy numbers of IGF1 mRNA resulting from the above threshold cycles reveal the wide range of mRNA levels in tissues from different individuals.

3.3.12.5 Discussion

Real-time RT-PCR is the most versatile, rapid and easy to perform of all the methods used to quantify steady-state mRNA levels. It offers streamlined assay development, reproducible results and a large dynamic range. However, several problems remain concerning the interpretation of the quantitative data, particularly when handling *in vivo* samples obtained from different individuals. The first relates to the reporting of target mRNA copy numbers and how this is normalised (see chapter 1). A meaningful comparison requires accurate and relevant normalisation to some standard. Clearly, the quality of quantitative data cannot be better than the quality of the normaliser and any variation in the normaliser will obscure real changes and produce artifactual changes. The most straightforward approach is to normalise mRNA levels to copies per cell number (when using tissue culture cells or nucleated blood cells) or amounts of total RNA (solid tissue biopsies). Com-

bined with the use of a target-specific standard curve as external standard, this is as close as it is possible to achieve absolute quantification and allows relatively easy comparison between different laboratories. In some instances, it might be more relevant to normalise the target gene relative to a second mRNA. For example, an experiment designed to compare the degree of activated T-cell infiltration in colorectal cancers might usefully measure interleukin-2 receptor (IL-2R) mRNA levels relative to those of CD8 (cluster of differentiation 8). If interleukin-2 receptor mRNA levels are not normalised to some T-cell-specific markers, different expression levels might be due simply to the sample containing significantly more or fewer T-cells (Bustin *et al.*, 2001). Where no such relationship exists, normalisation to rRNA may be appropriate (Schmittgen & Zakrajsek, 2000). Certainly, the use of single housekeeping genes has been shown to be inappropriate, since the most commonly used ones show significant variation between individuals and their mRNA levels vary with treatment (Bustin, 2000).

A second problem is associated with the heterogeneity of *in vivo* samples and the lack of positional information provided by real-time RT-PCR methods. The development of laser capture microdissection, together with the ability to extract amplifiable RNA from paraffin-embedded tissue sections promises to overcome this problem and may allow the substitution of user-unfriendly *in situ* RT-PCR protocols with real-time RT-PCR technology. This technology is also ideally placed to complement studies that use microarrays to investigate gene expression and disease-related changes on a genome-wide scale. It can provide independent verification of the variable quality information obtained from microarrays and its use will become increasingly crucial for a correct and biologically relevant interpretation of the gene expression profiling data.

In conclusion, the exquisite sensitivity of all real-time RT-PCR methodologies has removed the need to develop ever more sensitive and specific quantification assays. Instead, the emphasis must shift towards developing experimental protocols and designs that are rigorously controlled to allow meaningful global comparisons and biologically relevant interpretations of the results.

3.3.13

RD-PCR – Restriction display polymerase chain reaction

3.3.13.1 Introduction

If the expression level of all or a set of genes in certain cells is compared with the expression level of cells under different physiological conditions or with the expression level of different cell types, an expression pattern can be identified that can then be used to identify the 'status' of a cell type as a kind of RNA fingerprinting (McClelland *et al.*, 1995). For example, tumour cells can be differentiated from normal cells on the basis of the differential expression of oncogenes.

Differential expression can be examined using a method called differential display RT-PCR (DDRT-PCR) (Liang & Pardee, 1992). Several modifications have been made to the standard DDRT-PCR protocol (Liang & Pardee, 1995; Zhang *et al.*, 1998), but these have not solved all the problems of this powerful technique.

For this reason, the restriction display PCR (RD-PCR) technique was invented (Weinstein & Boulamwini, 1997) to circumvent remaining problems. Subsequently, a similar technique called cDNA-based amplified fragment length polymorphism fingerprinting (cDNA-AFLP) (Bachem *et al.*, 1996), and two variations called restriction fragment length polymorphism-coupled domain-directed differential display (RC4D) (Fischer *et al.*, 1995; Theissen & Fischer, 1997) and restriction fragment differential display (RFDD) (Gravesen *et al.*, 2000) were also reported. The details of the RD-PCR technique and its successors are described in the next sections.

3.3.13.2 Principle and basics

The RD-PCR method includes the following essential steps: (i) synthesis of double-stranded cDNA, (ii) digestion of the cDNA with one or more restriction enzymes, (iii) ligation of adapter primers, and (iv) PCR amplification of the cDNA fragment using sets of primers based on the adapter sequences and adjacent bases (see figure 3.89).

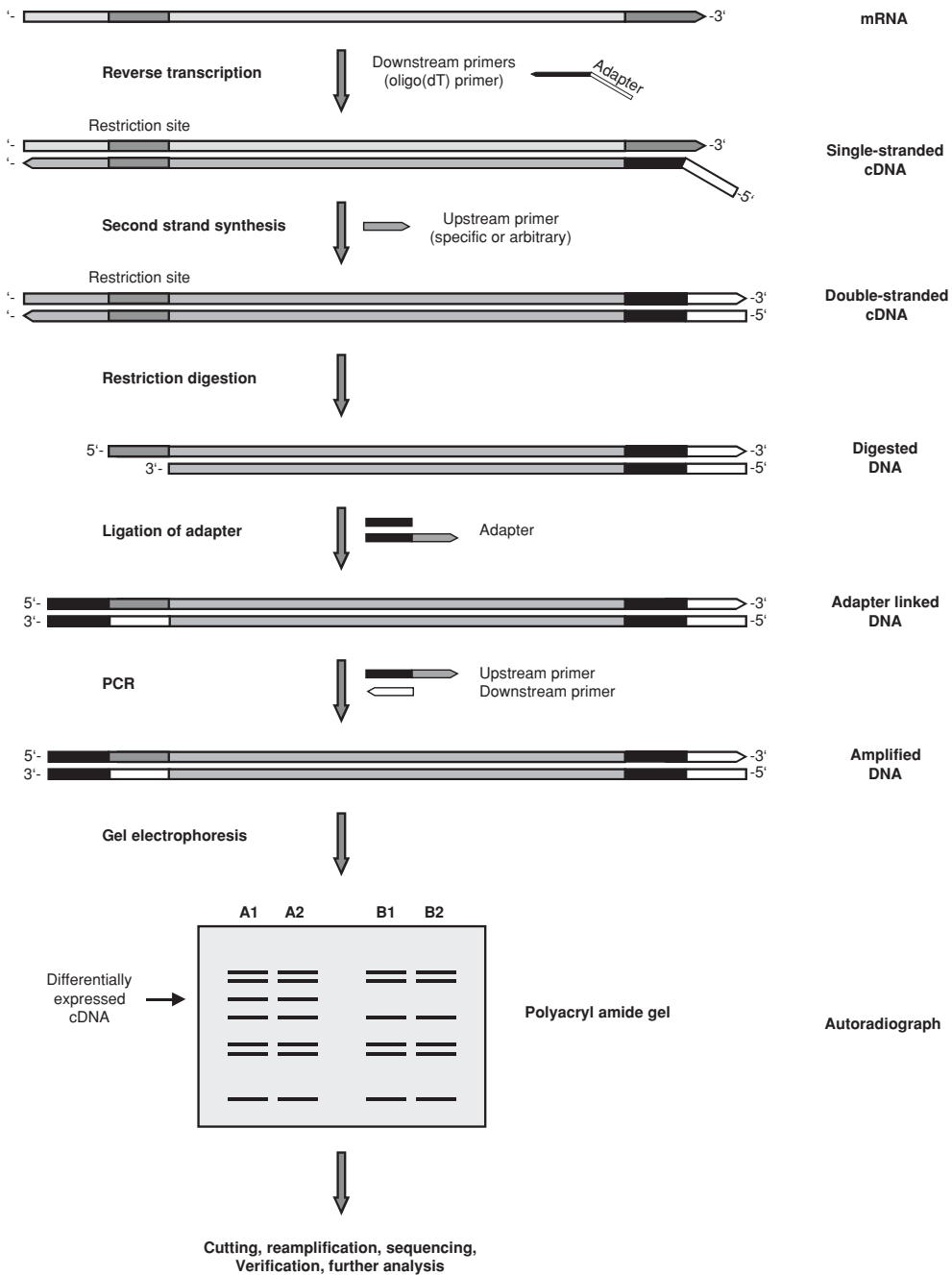
The method begins with providing purified and isolated mRNA. This can be achieved by using the standard method of hybridisation of mRNA to oligo(dT) primers linked to cellulose and removal of residual DNA by RNase-free DNase I. Purified mRNA is reverse transcribed using an oligo(dT)-containing primer. Double-stranded cDNA is then synthesised using standard methods.

The cDNA is digested with a restriction endonuclease to produce cDNA fragments with at least one sticky end capable of hybridising to an adapter primer. The ligation of adapters enables PCR at high-stringency conditions in contrast to the low-stringency conditions necessary in the original DDRT-PCR protocol. Finally, using a primer complementary to the adapter and a set of arbitrary primers, several PCR products are amplified. These PCR product mixtures are displayed by gel electrophoresis. The patterns of PCR products can be used to identify the type or the physiological status of cells. Furthermore, PCR products of interest can be cloned, sequenced and identified.

3.3.13.3 Variations of RD-PCR

Refinement of the RD-PCR technique led to the RC4D technique (Fischer *et al.*, 1995; Theissen & Fischer, 1997) which was specific as opposed to arbitrary primers the cDNA second-strand synthesis step (figure 3.89). This reduces the problem of weak bands and low reproducibility of the display patterns. However, the RC4D method is restricted to the analysis of members of gene families because the primers used are complementary to family-specific regions. The principle of RC4D is shown in figure 3.90.

Recently, another variation of RD-PCR, called RFDD, was developed (Gravesen *et al.*, 2000). This method is specialised for prokaryotic RNA, which lack poly(A) tails, and uses arbitrary primers in the cDNA synthesis step. Additionally, since oligo(dT) primers cannot be used, the method utilises for cDNA restriction fragments



- ◀ **Figure 3.89 Schematic illustration of the restriction display polymerase chain reaction (RD-PCR) technique.** First, eukaryotic mRNA is transcribed into cDNA by a reverse transcriptase using an oligo(dT) anchor primer. Second-strand synthesis of the cDNA is performed using arbitrary or specific upstream primers. Double-stranded cDNA is then digested using a restriction endonuclease. After the ligation of at least one double-stranded adapter primer (introducing a second anchor region) cDNA fragments are amplified by PCR. The two different anchor primers are used for the PCR reactions. In the case of restriction fragments digested at both ends, the second anchor primer (shown in black) can be used as upstream and downstream primer (not shown in this figure, see also figure 3.91 instead).

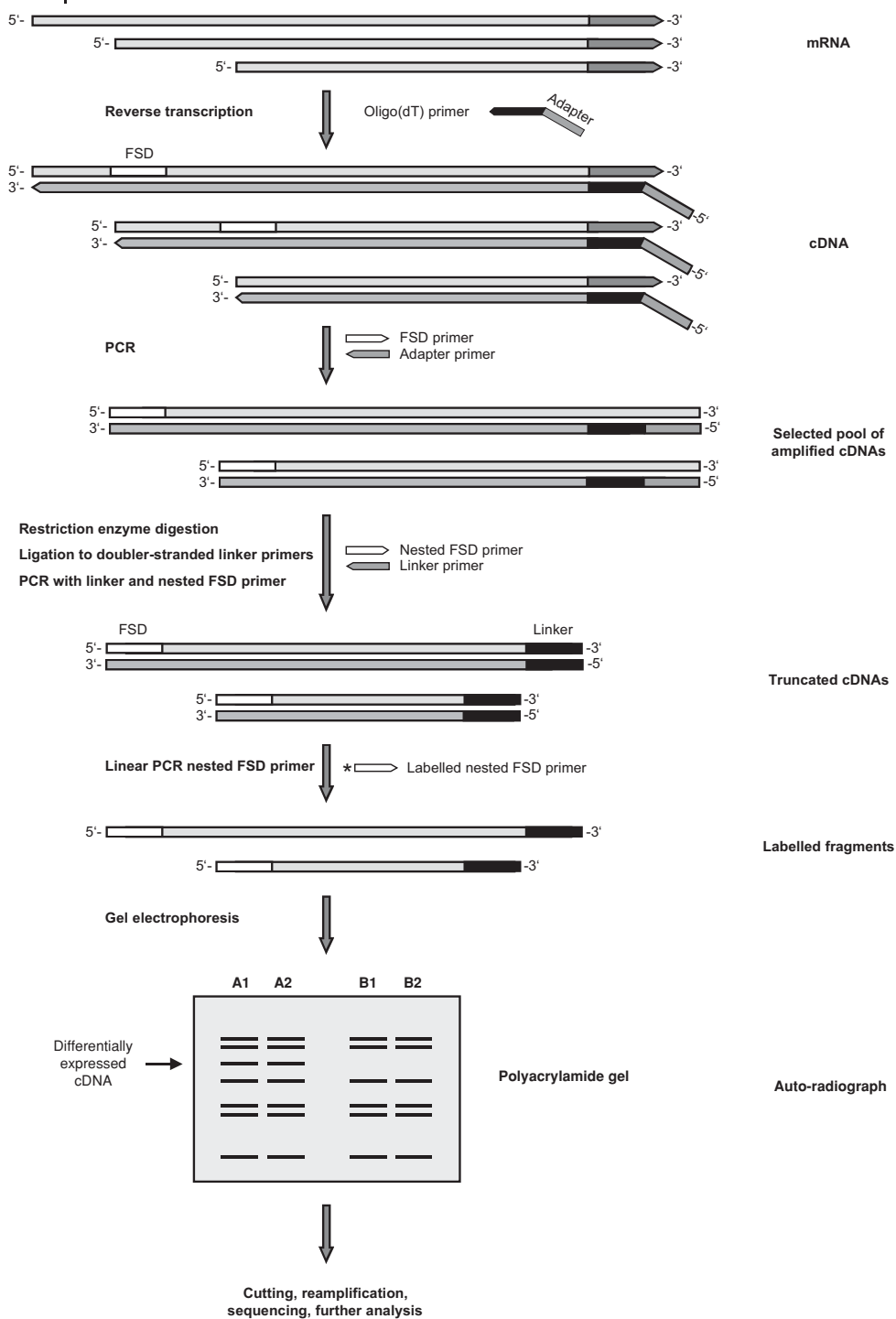
that have been cut at both ends. RFDD uses two types of anchor primers: a conventional, non-labelled anchor primer and a second anchor primer that lacks part of the complementary strand (see figure 3.91). An extension protection group of the second adapter (the EP adapter) prevents amplification of cDNA fragments with EP adapters on both sides (left arrow in figure 3.91). Fragments with a conventional anchor primer ligated on either end can be amplified but show no display signal as the usual anchors are not labelled (right arrow in figure 3.91). Therefore only fragments with one standard adapter and one EP adapter are amplified by PCR and (centred arrow in figure 3.91). The number of displayed bands is therefore reduced and the analysis of the restriction fragments is simplified.

Overall, RFDD has been used to analyse gene expression patterns in bovine oocytes matured in the presence or absence of fetal calf serum (Rzucidlo *et al.*, 2001), to identify changes in gene expression following induction of ischaemic tolerance in rat brain (Wrang *et al.*, 2001), and to display differences in mRNA expression profiles in nisin-resistant and pediocin-resistant *Listeria monocytogenes* 412 mutants (Gravesen *et al.*, 2000 and 2001).

3.3.13.4 Discussion

The DDRT-PCR method is an important technique for studying gene expression (Zhang *et al.*, 1998). Nevertheless, although widely used, it suffers from reproducibility problems due to the low stringency PCR conditions that are necessary for the unspecific annealing of the arbitrary primers (Debouck, 1995). This has been improved by the use of special or anchored primers and optimised PCR protocols (Bauer *et al.*, 1993; Mou *et al.*, 1994; Zhao *et al.*, 1995). However, Northern blotting is still required to confirm differential gene expression revealed by DDRT-PCR (Utans *et al.*, 1994).

The RD-PCR technique was invented to introduce anchor primer regions to the cDNA restriction fragment and, thus, to increase the stringency of the PCR conditions as a step towards improving the reproducibility of the method. But as arbitrary primers are still used in the standard RD-PCR technique the problems of low stringency in the first step and of numerous PCR reactions remain. Two further disadvantages of DDRT-PCR also apply to RD-PCR: often fragments containing 3'-untranslated regions (*i.e.*, regions of limited scientific value) are obtain-



◀ **Figure 3.90 Schematic illustration of the restriction fragment length polymorphism-coupled domain-directed differential display (RC4D) procedure.** In the first step, cDNA is reverse transcribed from mRNA using an oligo(dT) primer containing a 5'-adapter stretch. Single-stranded cDNA is amplified using the adapter primer and a primer directed to a protein family-specific domain (FSD) of the mRNA.

Following amplification, double-stranded cDNA is digested using a frequent cutting restriction enzyme, double-stranded linkers are ligated to the digested cDNA and a further round of PCR is performed using a linker primer and a nested FSD primer. Resulting PCR fragments are labelled, separated on a polyacrylamide gel and detected by auto-radiography. Bands of interest are excised, re-amplified and sequenced.

ed, and measurements are limited to eukaryotic organisms due to the dependence on the poly(A) tail of the mRNA (Gravesen *et al.*, 2000).

Finally, due to the possibility that the anchor primer may ligate to the digested fragments the display pattern may be difficult to analyse due to the high number of displayed bands. These problems may be the reason why RD-PCR has not entered routine laboratory practice. The variations described in this section, RC4D and RFDD, may change this in the future. One indication for this is the commercial production of a RFDD kit by Azign Bioscience A/S (Copenhagen, Denmark), a Danish company formerly known as Display Systems Biotech (Rzucidlo *et al.*, 2001).

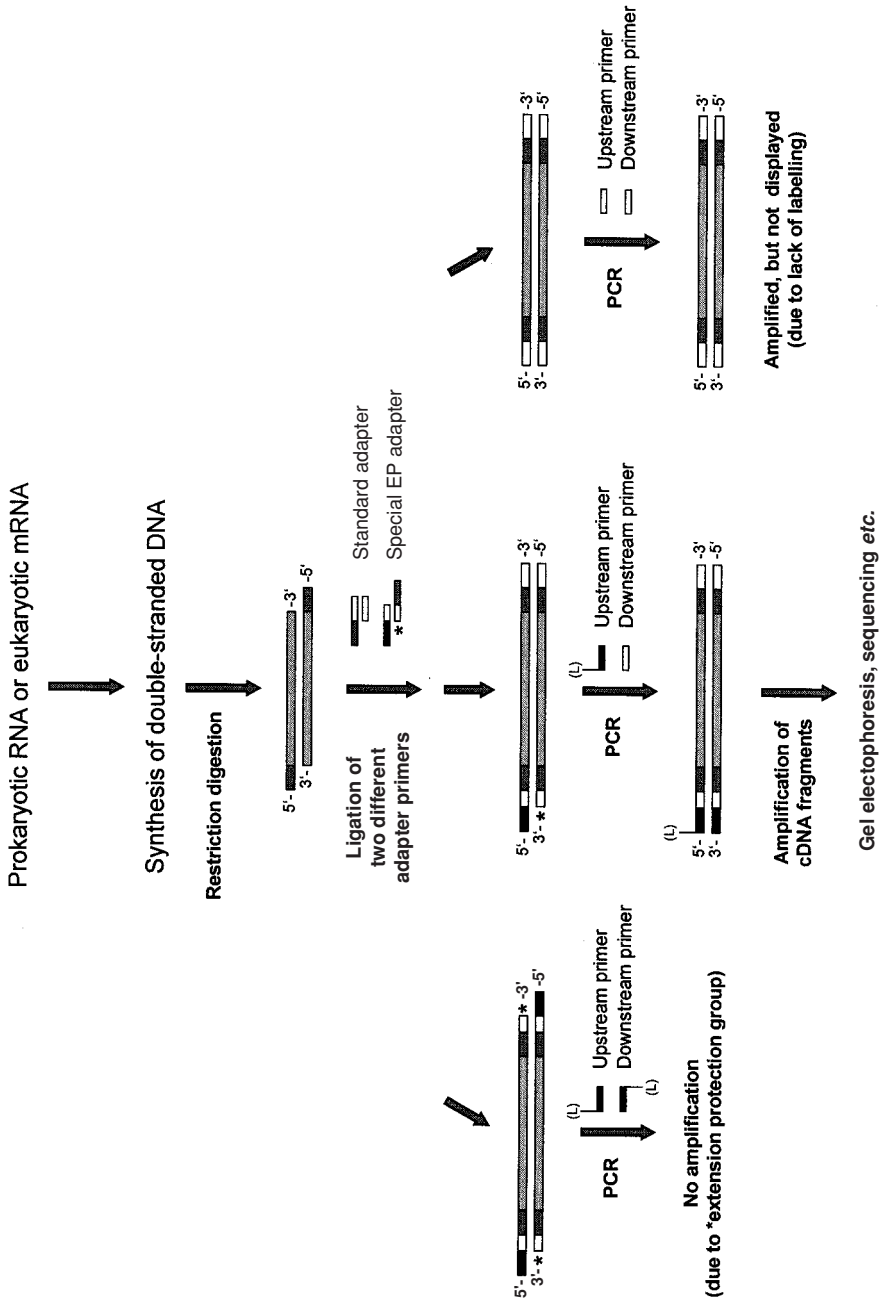
3.3.14

SIP RT-PCR

3.3.14.1 Introduction

One of the challenges of gene expression profiling is to assess cellular diversity by mRNA phenotyping of small homogeneous cell populations and single cells. However, the quantities of mRNA accessible for such analysis are often too low to apply techniques routinely used for gene expression profiling. To overcome this limitation mRNA can be globally amplified prior to the analysis of specific gene expression by standard techniques. So far, five methods for global mRNA amplification have been reported (for a review, see Brady, 2000). One protocol invented by Eberwine *et al.* (1992) generates approximately 100 to 1,000 copies of complementary RNA (cRNA) per mRNA molecule by linear amplification. This technique first synthesises double-stranded cDNA containing an RNA polymerase promoter, which subsequently serves as template to generate cRNA copies by *in vitro* transcription. If necessary, the resultant cRNA can be further amplified by a second round of linear amplification. Conceptually, three other methods used for global mRNA amplification are based on the generation of cDNA by reverse transcription and subsequent amplification by PCR using sequence independent-primers (SIP RT-PCR) (Brady *et al.*, 1990; Belyavsky *et al.*, 1989; Dixon *et al.*, 1998). Finally, a protocol has been described that combines linear RNA amplification with SIP RT-PCR, which is capable of generating a representative full-length cDNA library from individual cells (Ying *et al.*, 1999).

The present section focuses on the potential of SIP RT-PCR for biological research. The currently best characterised and most widely used SIP RT-PCR protocol was developed by Gerard Brady and colleagues (Brady *et al.*, 1990). The defining



- ◀ **Figure 3.91 Steps of the restriction fragment differential display (RFDD) technique (Gravesen *et al.*, 2000).** RFDD is a variation of restriction display polymerase chain reaction (RD-PCR). Apart from the possibility of examining prokaryotic RNA, the main feature of this technique is the ligation of two adapter primers. The first adapter is a standard double-stranded primer with a sticky end. The second adapter is a special primer with an extension protection group (EP primer). The protection group blocks the copy process of the polymerase. The labelled primer used in the following PCR step is complementary to the single-stranded region blocked by the protection group. Restriction fragments with EP adapter primers hybridised to either end (left arrow) cannot be amplified. By contrast, restriction fragments with standard adapter primers hybridised to either end are (weakly) amplified but are not displayed due to the lack of labelling. Only fragments with one standard and one EP adapter primer are both amplified and displayed (centre arrow).

characteristics of this PCR protocol, also called poly(A)-PCR, are the conditions of the reverse transcriptase reaction, which limit the size of the first-strand cDNA to 300 to 700 bases, thus minimising the selection against long transcripts during amplification. This distinct quality is crucial for gene expression profiling as it preserves the relative abundances of specific transcripts present in the initial mRNA population. The latter has been documented in spiking experiments where two-fold changes in input cRNA copies per cell were detected with a high linear coefficient within a range of 80 to 10,000 cRNA copies per cell (Theilgaard-Mönch *et al.*, 2001). Due to its quantitative performance and its simplicity poly(A)-PCR has been applied in many studies to generate cDNA from multiple samples prior to gene expression analysis by standard techniques. Standard techniques that have been combined successfully with poly(A)-PCR for gene expression profiling include virtual Northern blot analysis, real-time RT-PCR, and cDNA arrays (Al Taher *et al.*, 2000; Billia *et al.*, 2001; Brady *et al.*, 1995; Bryant *et al.*, 1999; Cheng *et al.*, 1996; Theilgaard-Mönch *et al.*, 2001).

3.3.14.2 Principle and basics

A schematic diagram of the poly(A)-PCR protocol described by Brady *et al.* (1990) is shown in figure 3.92. For the reverse transcription of mRNA into cDNA, small amounts of cells (one to one hundred cells) or, alternatively, of total RNA (one to fifty nanograms) are resuspended into PCR tubes containing first-strand buffer with a detergent. In the following, the samples are heated for cell lysis and denaturation of RNA. After cooling and addition of reverse transcriptase, samples are incubated for a short period (15 minutes) followed by heat inactivation of the enzyme. The resultant first-strand cDNA is then poly(A)-tailed using a terminal transferase. Aliquots of the poly(A)-reaction are then directly resuspended into PCR buffer containing a poly(T) primer, and poly(A)-tailed cDNA is amplified by 35 to 50 cycles of PCR. For subsequent gene expression profiling by virtual Northern blot analysis, real-time RT-PCR, or cDNA arrays the poly(A)-PCR products are purified and quantified by spectrophotometry.

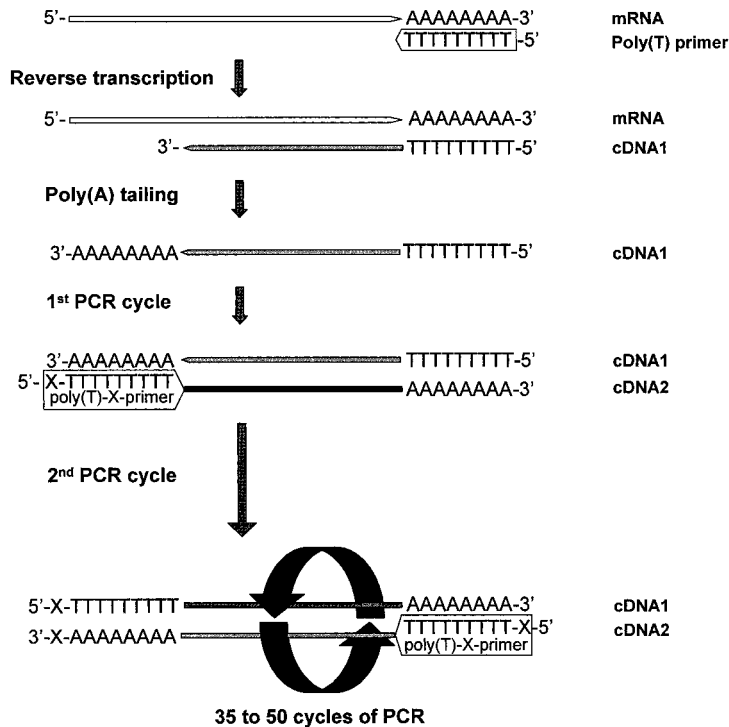


Figure 3.92 Schematic diagram of global mRNA amplification by reverse transcription and subsequent amplification by PCR using sequence-independent primers (SIP RT-PCR; also called poly(A)-PCR). In the first step, cells are lysed in first-strand buffer followed by reverse transcription of mRNA using an oligo(dT)

primer. In the second step, the first-strand cDNA is polyadenylated by terminal transferase to generate a 5'-oligo(dT) and 3'-poly(A)-tailed cDNA that finally is amplified by PCR using a sequence independent 5'-(T)₂₄X-3' (X is adenine, cytosine or guanine, but not thymine) primer mix.

3.3.14.3 Presentation of typical results

Figure 3.93 depicts a typical virtual Northern blot analysis of poly(A)-PCR products generated from individual cells. Figure 3.94 illustrates real-time RT-PCR analysis of a poly(A)-PCR product and figure 3.95 demonstrates a cDNA array hybridised with radiolabelled poly(A)-PCR products.

3.3.14.4 Examples

Techniques such as laser capture microdissection (LCM), fluorescence-activated cell sorting (FACS), and sibling cell analyses have greatly enhanced the ability to isolate small homogeneous populations and single cells based on morphology, tissue location, immunophenotype, and functional potential. Some of these techniques for cell isolation have been combined successfully with poly(A)-PCR in

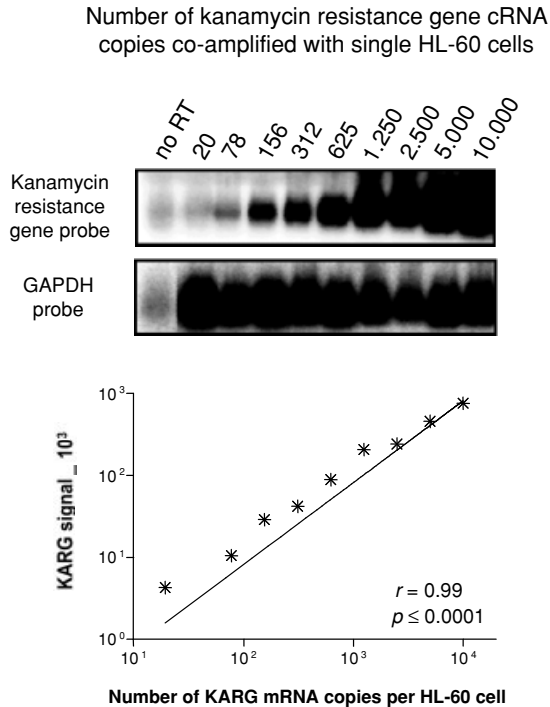


Figure 3.93 Virtual Northern blot analysis of poly(A)-PCR products generated from individual cells. Polyadenylated complementary RNA (cRNA) encoding a fragment of the kanamycin resistance gene (KARG) was generated by *in vitro* transcription, quantified by spectrophotometry, and added in two-fold dilutions to lysates of single HL-60 cells. After global mRNA amplification by poly(A)-PCR, the resultant cDNA products were run on a one percent agarose gel, transferred to a nylon membrane, and hy-

bridised with radiolabelled probes for the kanamycin resistance gene and the GAPDH housekeeping gene (upper part of the figure). Signals of the kanamycin resistance gene were quantified by PhosphorImager™ analysis and plotted against the number of KARG copies per single cell, and the indicated linear coefficient was determined (lower part of the figure). 'No RT' indicates a control sample that was amplified without addition of reverse transcriptase.

studies of normal and malignant cells. Guy Sauvageau and colleagues, for example, have isolated functionally primitive and committed haematopoietic progenitor cells by fluorescence-activated cell sorting (FACS), and subjected the subpopulations to poly(A)-PCR and subsequent virtual Northern blot analysis. Using this strategy it was possible to demonstrate the expression of several Hox genes in haematopoietic progenitor cells and to detect a regulated program of Hox gene expression during haematopoietic differentiation (Sauvageau *et al.*, 1994). A similar approach has been used to compare Hox gene expression profiles in normal and leukaemic haematopoietic progenitor cells (Kawagoe *et al.*, 1999). Another study, which employed FACS for the isolation of haematopoietic stem cells based on their *fms*-like tyrosine kinase (*flt3*) immunophenotype, was able to demonstrate a strong

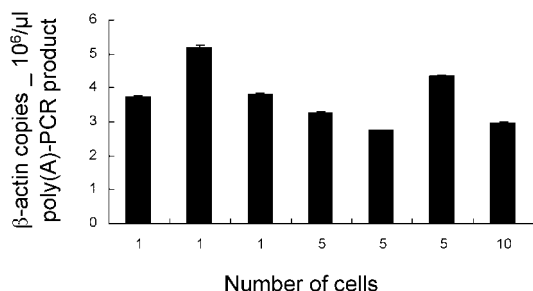


Figure 3.94 Real-time RT-PCR analysis of a poly(A)-PCR product. B-cells were sorted by fluorescence-activated cell sorting (FACS) directly into PCR tubes and poly(A)-PCR was performed. The number of β -actin copies in poly(A)-PCR products generated from one, five and ten B-cells were subsequently measured by real-time RT-PCR analysis. Mean values and standard deviations calculated from three real-time RT-PCR measurements of each poly(A)-PCR product are shown.

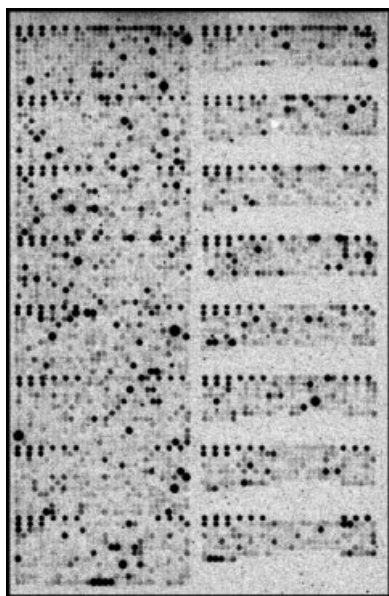


Figure 3.95 cDNA array hybridised with radio-labelled poly(A)-PCR products. Single myeloma plasma cells were sorted by fluorescence-activated cell sorting and subjected to poly(A)-PCR. The resultant product was radiolabelled and hybridised to a cDNA microarray. Gene expression profiles were visualised by PhosphorImager™ analysis.

correlation between *flt3* surface protein expression and *flt3* transcript levels (Adolfsson *et al.*, 2001).

A most elegant technique that has been combined with poly(A)-PCR and subsequent gene expression profiling is that of sibling cell analysis. This technique first generates eight to 16 identical siblings by synchronic cell division of individual haematopoietic progenitor cells *in vitro*. Subsequently, sibling cells are subjected to either gene expression analysis or cultured individually in media supportive of multi-lineage differentiation. Based on the fate of the cultured cells, the analysed sibling generation is classified as multi-lineage, bi-lineage, or uni-lineage-committed haematopoietic progenitor cells. This approach has been used to investigate the temporal order of gene expression during haematopoietic differentiation by the analysis of progenitor cells representing defined stages in the haematopoietic hier-

archy. Genes that have been analysed in these sibling cell studies include transcription factors, membrane proteins, and granule proteins (Billia *et al.*, 2001; Brady *et al.*, 1995). A recent study compared levels of gene expression in single cells and in a population of cells collected from an identical culture of the promyelocytic HL-60 cell line. Data from these experiments showed that individual cells had varying expression levels of specific transcripts. Notably, the average expression levels detected in single cells corresponded to the levels in the population (Theilgaard-Mönch *et al.*, 2001). These findings indicate that poly(A)-PCR and subsequent gene expression analysis is a powerful tool to assess the cellular diversity in apparently highly homogeneous populations. Finally, the combination of poly(A)-PCR and quantitative real-time RT-PCR methods has allowed a highly sensitive and accurate determination of specific target cDNA molecules (Al Taher *et al.*, 2000). Figure 3.94 shows an example of β -actin copy quantification in poly(A)-PCR products by real-time RT-PCR.

3.3.14.5 Discussion

Global mRNA amplification by poly(A)-PCR and subsequent gene expression analysis is a powerful tool for mRNA phenotyping of small homogeneous cell populations and single cells. Despite its enormous potential for biological research, few studies have evaluated the sensitivity and variability of protocols combining poly(A)-PCR with standard techniques for gene expression profiling. One study, for example, which employed poly(A)-PCR and virtual Northern blotting for parallel analysis of total RNA (0.6 nanograms per analysis) and lysed cells (25 per analysis), demonstrated 1.3-fold to 1.8-fold differences of expression levels for high as well as low abundant genes (Theilgaard-Mönch *et al.*, 2001). The same study also demonstrated that the gene expression levels detected in parallel by hybridisation of radiolabelled poly(A)-PCR products to cDNA arrays were comparable to those detected by Northern blot analysis of total RNA. These results indicated a low variability of global mRNA amplification as the relative frequencies of transcripts in a given mRNA population were not essentially distorted by poly(A)-PCR. Two independent studies substantiated these observations in spiking experiments by demonstrating that two-fold changes in input cRNA copies per single cell were detected with a high linear coefficient (Theilgaard-Mönch *et al.*, 2001; Cheng *et al.*, 1996). The two spiking experiments further showed that poly(A) PCR combined with virtual Northern blot analysis has a detection limit below 80 and 250 cRNA copies per single cell. An even higher sensitivity, *i.e.* one mRNA copy per 100 cells, was shown in a recent study combining poly(A)-mRNA with real-time RT-PCR (Al Taher *et al.*, 2000). Taken together, these findings demonstrate that methods based on poly(A)-PCR can reliably measure genes expressed at low levels in single cells. Notably, this reliability of poly(A)-PCR is highly dependent on the conditions of the reverse transcriptase reaction, which limit the size of the first-strand cDNA to the 300 to 700 bases, and, thus, avoids a bias against long transcripts during amplification. This, however, renders poly(A)-PCR unsuitable for analysis of 5'-sequences. Analysis of 5'-sequences can be achieved by more complex protocols

that generate full length cDNA from small amounts of cells or total RNA (Be-lyavsky *et al.*, 1989; Ying *et al.*, 1999).

A mayor advantage of the poly(A)-PCR protocol is that it can be applied directly to living cells without preceding RNA purification. This quality is crucial for the analysis of single cells as it makes the protocol robust by avoiding sample loss accompanied with purification procedures or mRNA degradation. A potential drawback of protocols employing direct cell lysis, however, is the presence of genomic DNA that might result in false-positives. This does not appear to be a major problem when poly(A)-PCR is combined with virtual Northern blot analysis or use of DNA arrays. However, poly(A)-PCR and subsequent real-time RT-PCR will generate a background signal from genomic DNA that must be considered when calculating gene expression levels.

Given the variety of protocols available for global mRNA amplification, the choice of a method highly depends on the requirements of the individual study. The poly(A)-PCR protocol provides a powerful means for gene expression studies employing limited amounts of cells, due to its simplicity, robustness, and most importantly the well-documented preservation of transcript abundances during amplification.

3.3.15

Targeted display

3.3.15.1 Introduction

The aim of most methods for analysing gene expression is to detect as many mRNAs as possible including abundant and rare transcripts. The simple and rapid targeted display method is based on the differential display technique. Not only does it have a high targeting rate, it also allows the generation of targeted fragments derived from protein-coding regions of mRNAs and not from the 3'-untranslated region as in the original differential display RT-PCR method (Brown *et al.*, 1999; Liang & Pardee, 1992).

3.3.15.2 Principle and basics

Targeted display is based on a method previously described by Lopez-Nieto & Nigam (1996), who first demonstrated a non-random distribution frequency for octanucleotide sequences in the protein-coding regions of sense DNA strands among 1,000 randomly selected human mRNAs. It was further found that out of all possible octanucleotides a group of 30 octanucleotides is highly over-represented in most cDNAs derived from human samples. These 30 octanucleotides can be used for the design of primer pairs in both sense and antisense orientation, which are linked with a specific five base pair tail to obtain a differential display RT-PCR-based format of 13 base pairs (targeted display primers). Computer analysis of PCRs using these primers showed that more than 70 percent of genes in the mammalian genome can be targeted using the targeted display approach (Brown *et al.*, 1999a).

The targeted display procedure is as follows. First, total RNA is isolated from samples of interest using standard methods and reverse transcribed into cDNA using oligo(dT) primers. Following amplification by PCR using the targeted display primers, fragments are separated by agarose gel electrophoresis, isolated from the gel and cloned for identification (Brown *et al.*, 1999). The principle of the targeted display method is illustrated in figure 3.96.

This method can be extended to target specific gene families by selecting octanucleotides that (i) are highly over-represented in the protein-coding DNA sequence of the original gene family and (ii) can theoretically target 90 percent of the sequences within this gene family.

3.3.15.3 Results obtained by targeted display

Two studies have been published using this technique (Brown *et al.*, 1999). In these studies, the nerve growth factor-induced differentiation of the rat pheochromocytoma cell line PC12 was investigated by targeting transcripts that are differentially regulated during treatment with nerve growth factor.

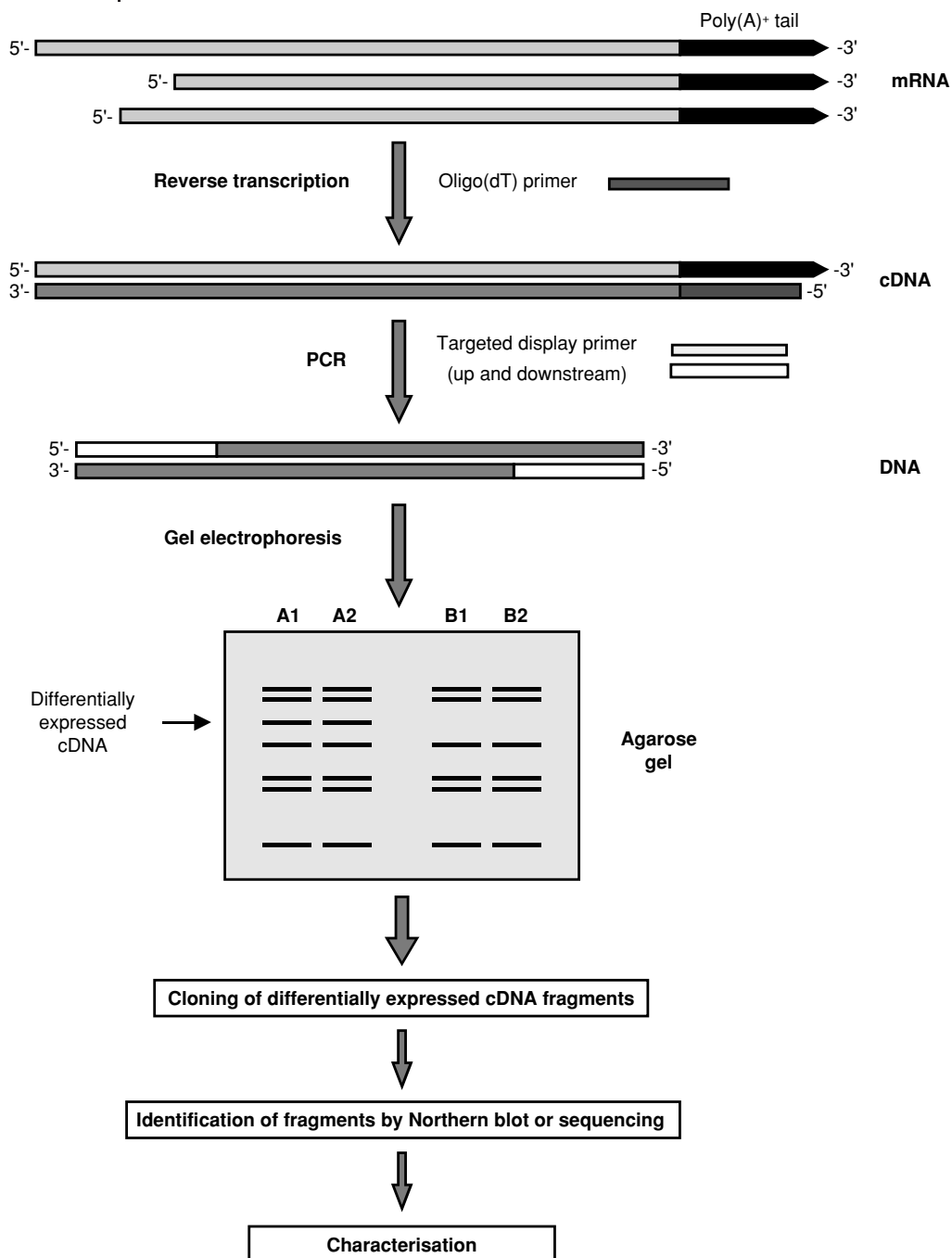
For this purpose, total RNA was isolated from PC12 cells which were incubated with nerve growth factor for varying time-periods. PCR amplification of the synthesised cDNA was then performed using the targeted display primers. A typical agarose gel of targeted display products is shown in figure 3.97. Genes that are differentially expressed due to nerve growth factor exposure could clearly be distinguished from non-regulated genes. Following isolation of the bands, the fragments were confirmed by Northern blot analysis or RT-PCR and sequenced after cloning.

Homology searches using public databases revealed that out of 32 isolated genes only seven were present in the databases GenBank and EMBL (Brown *et al.*, 1999b). To prove the role of the identified sequences in the differentiation of PC12 cells one clone, the serine/threonine kinase MARK1 (microtubule-associated protein/microtubule affinity-regulating kinase 1), was selected for further study, which indicated an involvement of MARK1 in the nerve growth factor-induced differentiation of PC12 cells.

3.3.15.4 Discussion

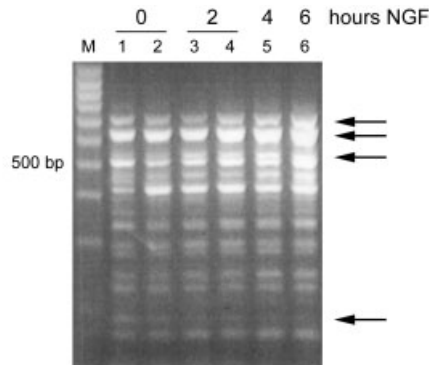
There are several advantages in targeted display that makes it a good alternative to the already established differential display RT-PCR method. Theoretical calculations showed that the targeted display primers should be able to recognise nearly 14,000 of the estimated 10,000 to 20,000 mRNAs in a typical mammalian cell. By choosing appropriate primer pairs, it is possible to target specific gene families. Because the primers normally target protein-coding regions, the location of exon-containing regions within large genomic subsets can be identified. Additionally, the sensitivity of this method is high enough to target rare mRNA species.

Because targeted display is a PCR-based technique, problems concerning background noise, reproducibility with altered conditions and the sensitivity to DNA contaminations are problems as with other PCR-based methods. Furthermore,



- ◀ **Figure 3.96 Flowchart of the targeted display principle.** After isolation of total RNA, mRNA is reverse transcribed into cDNA using an oligo(dT) primer. Complementary DNA is amplified by PCR using specifically-designed octanucleotide primers (up and downstream targeted display primers). After separation of amplified cDNA in an agarose gel, differentially expressed targets are excised from the gel, cloned and identified by sequencing. Differential expression of such genes is verified by independent methods such as Northern blot analysis or RT-PCR.

Figure 3.97 Analysis of gene products, identified by the targeted display method. The agarose gel reflects a typical band pattern revealed by amplification of cDNA using a single octanucleotide targeted display primer pair in the targeted display PCR. Origin total RNA was derived from undifferentiated PC12 cells (lanes 1 and 2), PC12 cells incubated for two hours (lanes 3 and 4), for four hours (lane 5) and for six hours with nerve growth factor (NGF) (lane 6). Lanes 1 and 2 and lanes 3 and 4 demonstrate the high reproducibility of duplicates used in the targeted display PCRs, when performed under the same conditions. Putative differentially expressed genes are indicated by arrows.



when several different PCR amplifications are performed, the number of targeted display primer pairs used is very large and the amount of starting material can be of limiting nature.

Overall, targeted display offers many advantages for the specific and selective targeting of genomic subsets, so that it may be of use in the clarification of gene functions.

3.4

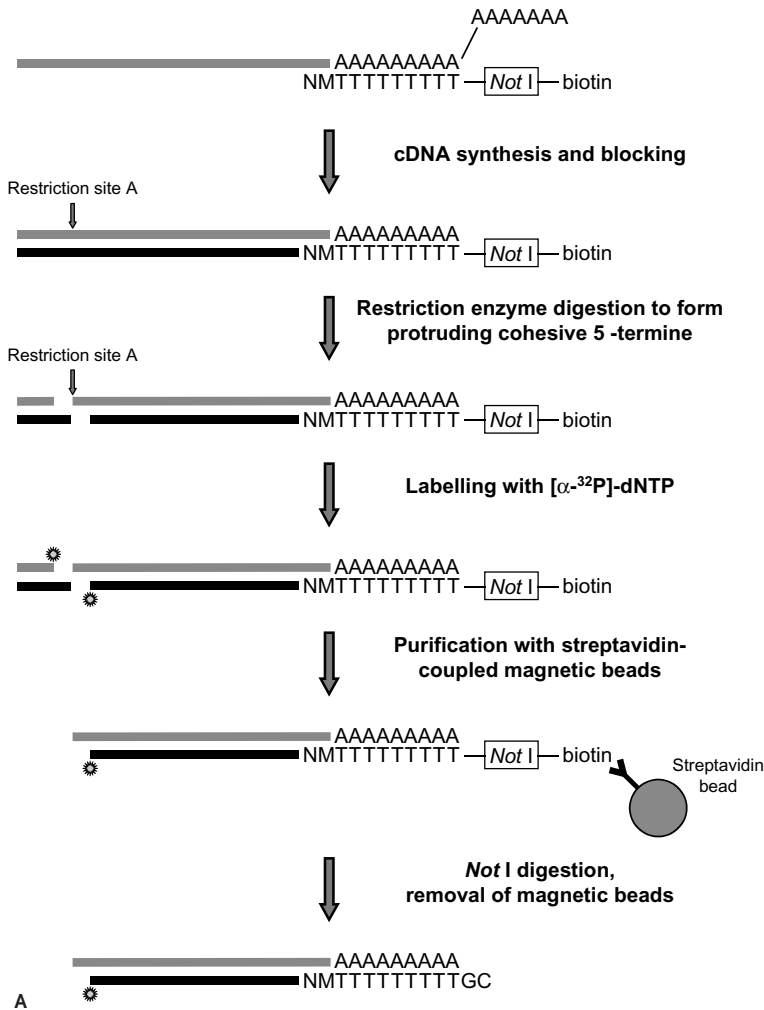
Further methods for analysing gene expression based on other principles

3.4.1

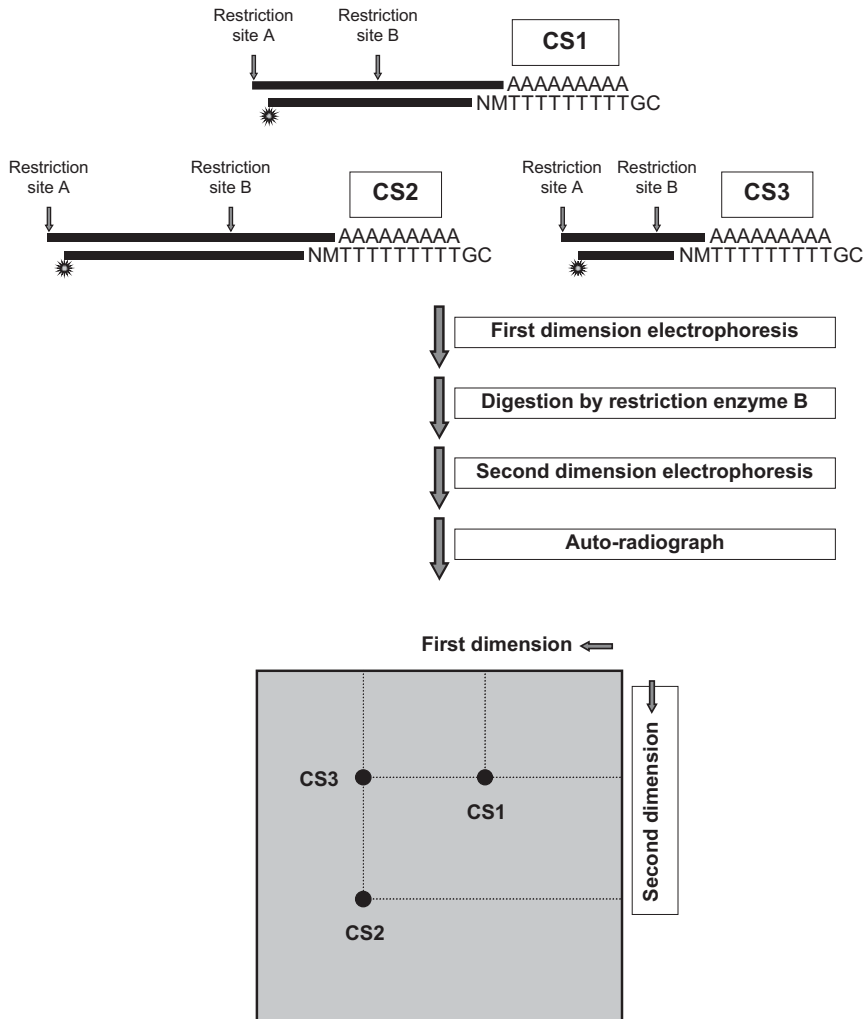
RLCS — Restriction landmark cDNA scanning

3.4.1.1 Introduction

Several years ago a simple genome scanning method designated as restriction landmark genomic scanning (RLGS) was developed by adopting the concept of using restriction enzyme sites as landmarks in conjunction with high-resolution two-dimensional gel electrophoresis. Its significant characteristic is that more than a thousand loci have been visualised quantitatively (Hatada *et al.*, 1991; Hayashizaki *et al.*, 1993), and it has been applied to many areas such as genetic mapping (Okuiizumi *et al.*, 1997), mutation analysis in cancer DNA (Hayashizaki *et al.*,



1993), detection of tissue-specific or developmental stage-dependent DNA methylation, and to DNA imprinting and genome analysis of bacterial drug-resistance (Fujimura & Murakami, 1997; Kawai *et al.*, 1993; Shibata & Plass, 1997). Restriction landmark cDNA scanning (RLCS) has been developed as a further application of RLGS to cDNA scanning and has proved to be very powerful method for systematic investigation of gene expression by allowing the simultaneous display of many cDNA species spots on a two-dimensional gel (Suzuki *et al.*, 1996; Yaoi *et al.*, 1997).



B

Figure 3.98 Schematic representation of the principle of restriction landmark cDNA scanning (RLCS). A) Preparation of the RLCS sample. The thin horizontal line and thick double horizontal lines indicate poly(A)⁺ mRNA and double-stranded cDNA, respectively. Asterisks show the incorporated radiolabelled nucleotides. An oligo(dT) anchor primer with MN (where M is adenine, cytosine, guanine; N is

adenine, cytosine, guanine, or thymine) at the 3'-end and [α -³²P]-dGTP were used for cDNA synthesis and labelling, respectively.

B) Separation of cDNA species by two-dimensional gel electrophoresis. cDNA fragments in a RLCS sample (CS1, CS2 and CS3) were separated by two-dimensional gel electrophoresis and detected as spots CS1, CS2 and CS3, respectively.

3.4.1.2 Principle and basics

The principle of RLCS is essentially the same as that of RLGS. The cDNA fragments that contain as landmarks cleaved and labelled restriction sites are separated by high-resolution two-dimensional gel electrophoresis followed by auto-radiography (Suzuki *et al.*, 1996; Yaoi *et al.*, 1997). It is important to prepare cDNA species of a uniform length for each mRNA species in order to display the various cDNA species as discrete gel spots. The outline of RLCS is schematically shown in figure 3.98.A and 3.98.B. As indicated in figure 3.98.A, an oligo(dT) anchor primer is designed for cDNA synthesis that primes at a particular location in the upstream end of the poly(A)⁺ mRNA (Yaoi *et al.*, 1997). For example, it comprises a 15-mer oligo(dT) stretch plus an additional two nucleotides MN at the 3'-end (M is adenine, cytosine, or guanine; N is adenine, cytosine, guanine, or thymine). In principle, different cDNA species will be displayed by selecting a different base as N and M. This primer also contains a tag molecule such as a biotin residue and multiple restriction enzyme sites at the 5'-end to allow simplified cloning and recovery of synthesised cDNA. After cDNA synthesis using these primers, blocking by ddNTPs is performed to prevent non-specific labelling. The cDNA synthesised is then digested with a restriction enzyme 'A' which preferably creates preferably protruding cohesive 5'-termini, because the restriction sites can easily be radiolabelled by Sequenase[®] and are also advantageous for spot cloning. The labelled fragments are recovered using streptavidin-conjugated magnetic beads and released from the beads by restriction enzymes such as *Not* I, the recognition site for which is included in the anchor primer. As shown in figure 3.98.B, the labelled cDNA samples (*e.g.*, the fragments CS1, CS2, and CS3) which are uniform in length for individual mRNA species are separated by electrophoresis on high-resolution two-dimensional agarose gels. After electrophoresis in the first dimension, the cDNA fragments in the gel are completely digested *in situ* with restriction enzyme 'B', and subjected to polyacrylamide gel electrophoresis in the second dimension. The gel is then dried and submitted to auto-radiography. Individual cDNA species are displayed as gel spots according to the positions of the enzyme A site and its nearest downstream enzyme 'B' site.

3.4.1.3 Typical results

Figure 3.99 represents typical RLCS patterns for cDNA samples prepared from adult mouse liver and brain (cerebral cortex, cerebellum and brain stem). Synthesis of cDNA was performed with a commercial cDNA synthesis kit (SuperScript[™], Invitrogen GmbH, Karlsruhe, Germany) using several micrograms of poly (A)⁺ mRNA and one microgram of the anchor primer (with N as adenine: 5'-AM(T)₁₅CCC GCC GGC GAG CGC TAG ATC TTG ATC AG-biotin-3'). After blocking with four kinds of ddNTPs and Sequenase[®], cDNA equivalent to one microgram starting poly(A)⁺ mRNA was completely digested with *Bam* HI and *Bgl* II as enzyme 'A'. Labelling was performed using Sequenase[®] and [α -³²P]-dGTP in the presence of ddCTP and ddTTP. To avoid non-specific labelling, the concentration and reaction time of the DNA polymerase should be determined separately.

After magnetically recovering the labelled cDNA samples with streptavidin-labelled magnetic beads, these were applied to the two-dimensional gel electrophoresis. The electrophoretic apparatus was obtained from Biocraft Laboratories, Inc. (Fair Lawn, New Jersey, USA). The first and second dimensional gel electrophoresis was performed in vertical disc gels (one percent agarose gel in Teflon™ tubing) and glass plate gels (six percent polyacrylamide gel), respectively. The agarose gel rod after the first electrophoresis was treated with *Hin* fl as enzyme 'B' and connected to the second electrophoretic plate gel. Many spots were observed: about 500 spots for the liver (figure 3.99.A), and more than 1,000 spots for each brain region (figure 3.99.A to 3.99.D). The spot pattern of the liver was quite different from that of each brain region. The patterns of the three brain regions are very similar but not identical. Spots whose intensities were similar (spots S1 and S2) or different (spots S3, S4, and S5) were easily distinguishable among the three regions (figure 3.100). Northern blot hybridisation was carried out using hybridisation probes which had been isolated from gel spots by the PCR-mediated spot cloning method (Suzuki *et al.*, 1994), and the Northern signals coincided well with the corresponding spot intensities. This means that the spot intensity reflects the corresponding gene expression level. Furthermore, the sensitivity of the RLCS was calculated to be enough to detect very rare mRNA species. Usually, 0.2 to 0.3 micrograms of double-stranded cDNA synthesised is applied to a gel, when one microgram of poly(A)⁺ mRNA is used. Since the mean size of the cDNA is 1,000 base pairs,

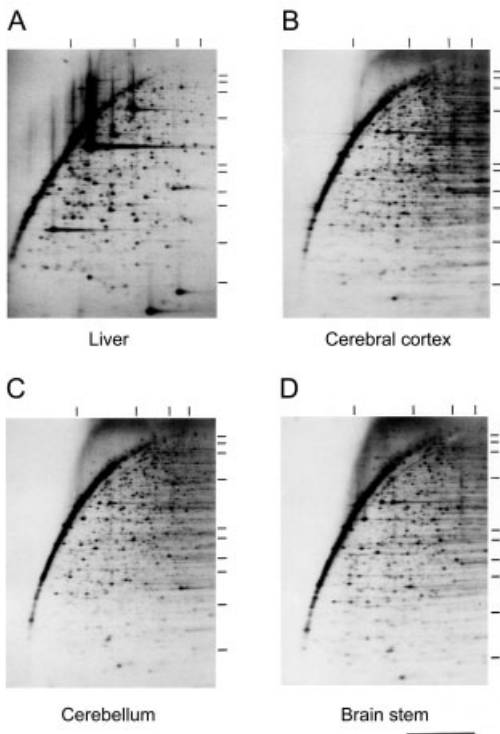
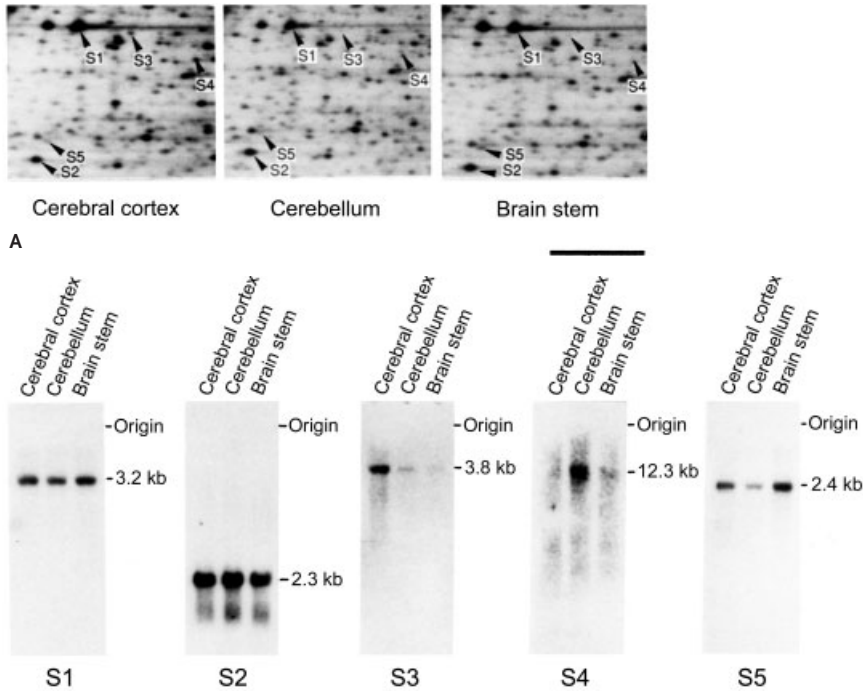


Figure 3.99 Typical restriction landmark cDNA scanning (RLCS) profiles for adult mouse liver A), cerebral cortex B), cerebellum C), and brain stem D). In each profile, *Bam* HI and *Bgl* II were used for restriction enzyme 'A' and *Hin* fl for restriction enzyme 'B' (see text for further information). A λ Eco T14I DNA digest and a ϕ X174 *Hae* III DNA digest were used as the first and second dimension molecular markers, respectively. The scales in each figure are 1.88, 1.49, 0.93 and 0.42 kilobases (from right to left) and 1,357, 1,078, 872, 603, 310, 281/271, 234, 194, 118 and 72 base pairs (from top to bottom) for the first and second dimensions, respectively. The thick horizontal bar corresponds to ten centimetres. The figure is reproduced from Suzuki *et al.* (1996) with permission from *Nucleic Acids Research*.



B

Figure 3.100 Correlation between the difference in intensity of the spots among restriction landmark cDNA scanning (RLCS) profiles and expression level of the corresponding mRNA species among RNA samples used for RLCS.

A) Profiles of the spots from each clone. A part of the RLCS profiles is from figure 3.99.B to 3.99.D. The spots cloned in the present study (S1 to S5) are shown by arrowheads. The thick horizontal bar is five centimetres. **B)** RNA blot hybridisation using cloned spot DNA fragments

as probes. Two micrograms of poly(A)⁺ mRNA from the three brain regions were used for each lane. Cloned spot DNA fragments used as probes are described under the photographs. The fragment size in each band with a signal was estimated using a 0.24 to 9.5 kilobases RNA ladder. The size of 12.3 kilobases in S4 was estimated by extrapolation. The figure is reproduced from Suzuki *et al.* (1996) with permission from *Nucleic Acids Research*.

the total copy number of the cDNA is estimated as $[(0.2 \text{ to } 0.3 \times 10^{-6}) / (660 \times 1,000 \text{ base pairs})] \times (6 \times 10^{23}) = (1.8 \text{ to } 2.7) \times 10^{11}$ copies. Since the population of very rare mRNA species is thought to be 0.0001 percent of total mRNA (Sambrook *et al.*, 1989), the number of cDNA is calculated as $(1.8 \text{ to } 2.7) \times 10^{11} \times 0.0001 \text{ percent} = (1.8 \text{ to } 2.7) \times 10^5$. In RLGS, one microgram of genomic DNA, equivalent to 3×10^5 copies for mammalian genome, is applied to a gel and almost all spots observed correspond to diploid intensity. Some spots are also detected with an intensity several fold weaker than diploid spots. Thus, RLCS should have the sensitivity to detect very rare mRNA species (Suzuki *et al.*, 1996; Yaoi *et al.*, 1997).

Using RLCS, various differentially expressed genes were identified: (i) chick retinal dehydrogenase 3 (RALDH-3; also termed member A3 of aldehyde dehydrogenase 1 family, ALDH1A3; aldehyde dehydrogenase 6; ALDH6; acetaldehyde dehydrogenase 6), a novel retinal dehydrogenase, expressed in the ventral region of the retina (Suzuki *et al.*, 2000); (ii) a novel gene designated as N-copine (also termed copine VI; CPNE6) whose expression in the hippocampus was up-regulated by kainate stimulation (Nakayama *et al.*, 1998); (iii) the uridine kinase gene was identified among up-regulated genes in ventral horn after crush injury to rat sciatic nerves (Yuh *et al.*, 1999); and (iv) several genes were identified, the expression of which changed during neuronal cell death evoked by nerve growth factor depletion (Mayumi-Matsuda *et al.*, 1999).

3.4.1.4 Discussion

Using the RLCS method, more than 1,000 discrete spots, each spot essentially corresponding to an individual gene species, can be displayed simultaneously on an RLCS gel. Differences in the intensity of the spots among RLCS profiles correlated well with the expression levels of the corresponding mRNA species. RLCS also proved sensitive enough to allow detection even of very rare mRNA species. Therefore, RLCS makes it possible to survey global gene expression as well as to screen for differentially expressed genes. However, the number of spots displayed is still limited. To increase the number of spots to be displayed, more than two RLCS gels with different restriction enzymes could be used. Since conventional RLCS gave redundant spots (the spots corresponding to the same genes appear in more than two RLCS profiles), RLCS was improved to reduce the spot redundancy (Mayumi *et al.*, 1998). Alternatively, to obtain RLCS spots as many as possible, various kinds of two-base anchored oligo(dT) primers were used, for example, 5'-dT₁₆VN-3', where N is any base, and V is adenine, cytosine, or guanine. Although each signal on the profile should ideally show one-to-one correspondence to a transcript without redundancy, selection of RNA species by two-base anchored oligo(dT) primer is not complete due to ambiguous recognition of the primers by reverse transcriptase. Thus, the reverse transcription reaction was improved to increase priming specificity in the presence of trehalose and different oligonucleotide-binding blockers (Mayumi-Matsuda *et al.*, 1999). These modifications should allow the number of cDNA species that can be displayed to be increased.

3.4.2

RNA patterns method

3.4.2.1 Introduction

An RNA pattern or poly(A)⁺ mRNA fingerprint is defined as a pattern of small DNA products representing a particular preparation of poly(A)⁺ mRNA or its corresponding cDNA. The RNA patterns method was developed by Chenchik *et al.* in 1993 in order to analyse the poly(A)⁺ mRNA composition of tissues or cells without any kind of amplification such as exponential amplification by PCR or linear am-

plification by *in vitro* transcription. In the RNA patterns method, poly(A)⁺ mRNA is reverse transcribed and radiolabelled using a set of normal dNTPs and one dideoxynucleotide for termination of the chain elongation during cDNA synthesis. The resulting short cDNA termination products are analysed by polyacrylamide gel electrophoresis (PAGE).

The aim of this method is to determine differences in the poly(A)⁺ mRNA composition and to monitor simultaneously changes in gene activity by semiquantitative analysis of several hundred individual mRNAs that are present in different preparations. With a set of 30 different oligonucleotide primers for use in reverse transcription, a comprehensive analysis of all mRNAs with abundances higher than 0.1 percent of the total RNA amount is claimed to be possible (Chenchik *et al.*, 1993b).

3.4.2.2 Principle and basics

Figure 3.101 depicts a schematic illustration of the principle of the RNA patterns method. Total RNA is isolated from a given tissue or cell using standard methods. Reverse transcription of poly(A)⁺ mRNA is then performed using Moloney murine leukemia virus (MMLV) reverse transcriptase and short oligonucleotide primers such as the ones listed in table 3.11. The method of RNA patterns comprises at least one oligonucleotide primer and a supply of the four nucleotides adenine, cytosine, guanine, and thymine of which one is a terminator nucleotide (Chenchik *et al.*, 1997) to achieve several short cDNA products, which are characteristic for one specific mRNA population.

In principle, all dideoxynucleotides may be used as terminators, but according to Chenchik *et al.* (1997) ddUTP is the dideoxynucleotide of choice. During cDNA synthesis a high background is often present due to non-terminated products and remaining oligonucleotide primers.

Reduction of the background is also possible by choosing a terminator with high stability against snake venom phosphodiesterase and hydrolysis of RNA by incuba-

Table 3.11 Oligonucleotide primer sequences used for displaying poly(A)⁺mRNA with the RNA patterns method for comparison of normal, cirrhotic and fetal human livers shown in figure 3.102 (Chenchik *et al.*, 1993b).

Primer	Sequence
F1	5'-TGC AGG CTG-3'
F6	5'-TGT GGC AGG-3'
P55	5'-TGC TGG GTG-3'
P41	5'-TGC TGG CAG-3'
P10	5'-TGC TGG AGG-3'
P35	5'-TGC TGG CGG-3'
P36	5'-TGC TGG TGG-3'
P57	5'-TGC TGG CTG-3'
P32	5'-TGC TGG AGT-3'

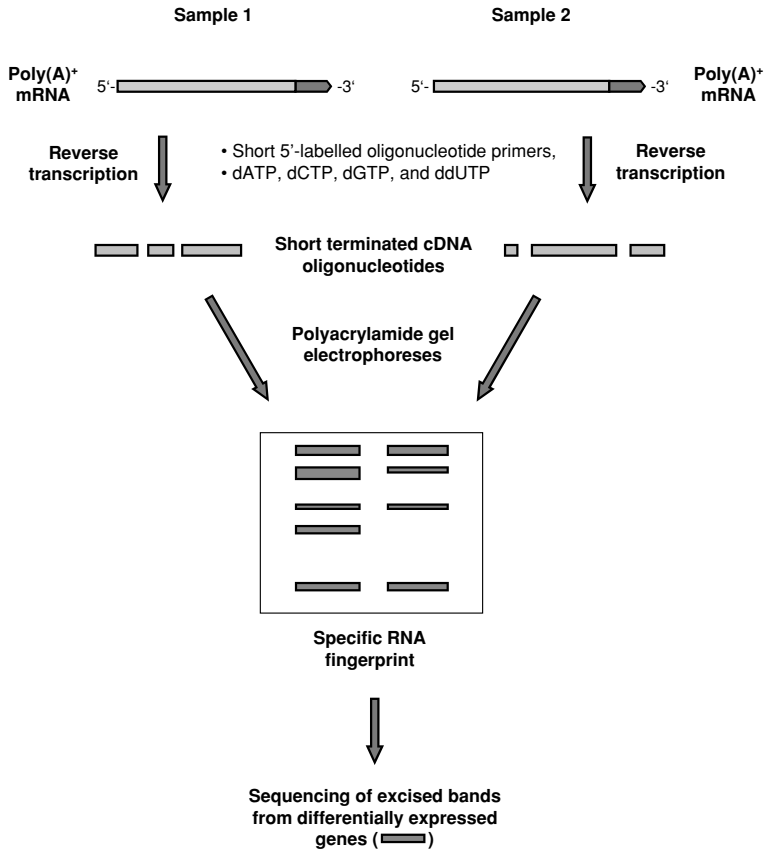


Figure 3.101 Schematic illustration of the RNA patterns method. Poly(A)⁺ mRNA is isolated from a given tissue and reverse transcription is carried out. For this, labelled primer(s), three natural dNTPs and one terminator dNTP are supplied. The resulting short terminated, labelled cDNA oligonucleotides are separated by polyacrylamide gel electrophoresis. Differentially expressed genes are identified by comparing the different RNA patterns. By slicing bands of interest and sequencing the oligonucleotides, the necessary sequence information is given for further analysis like PCR amplification or cloning.

tion with this enzyme subsequent to reverse transcription (see below). The resulting purified short cDNA termination products are separated by polyacrylamide gel electrophoresis, which results in characteristic poly(A)⁺ mRNA fingerprints of the cells or tissues. Radioactive labelling improves the sensitivity of the method, although other non-radioactive labels with appropriate detection systems like fluorescent labelling with laser scanning may also be possible.

For the RNA patterns method, the choice of primers is important for the quality of results. The following criteria should be used to design the primers (Chenchik *et al.*, 1993 and 1997): (i) In principal, primers with eight to ten nucleotides are suitable, whereas the best results are received with primers which are nine nucleotides

in length. (ii) The most commonly sequences in the coding regions of target genes should be used for the choice of primers in such a way that the probability of two complementary sequences appearing in any individual mRNA does not exceed three percent (Chenchik *et al.*, 1997). (iii) Minimising any repetitive, palindromic and homopolymeric sequences and binding to secondary non-complementary sites during subsequent cDNA synthesis is recommended, as well as (iv) maximising the specificity of binding to the complementary template sequence.

In order to provide a sufficiently detailed mRNA fingerprint, the number of different primers should be chosen in such a way as to allow differentiation between RNA fingerprints of various RNA preparations. As a guideline, a set of 30 PCR primers is sufficient for a comprehensive analysis of all mRNAs with abundances higher than 0.1 percent of the total RNA amount (Chenchik *et al.*, 1993b).

During reverse transcription, each primer produces ten to sixty bands per track following polyacrylamide gel electrophoresis under denaturing conditions. Increasing complexity of the poly(A)⁺ mRNA population results in the appearance of a greater number of bands per lane (Chenchik *et al.*, 1993a).

The RNA patterns procedure generally produces a background by non-terminated products and remaining primers. To reduce this background, non-terminated products are hydrolysed by incubation with snake venom phosphodiesterase after synthesis of cDNA. For this approach, it is necessary to use FAM or biotin-labelled ddUMP as terminators because extension products containing these 5'-modified analogues of ddUTP are at least 100-fold more resistant to snake venom phosphodiesterase than the corresponding non-terminated oligonucleotides (Chenchik *et al.*, 1997).

Another possibility for minimising the background is to use a labelled terminator like [α -³²P]-ddATP instead of a ³²phosphorus-labelled primer. In this case, only the terminated products appear in the polyacrylamide gel electrophoresis and a phosphodiesterase step is not required.

Following detection by auto-radiography, RNA patterns are compared, and bands of interest may be excised from the polyacrylamide gel. Direct chemical sequencing of the excised cDNA fragments usually provides enough information to identify the differentially expressed genes by database searches. Alternatively, excised cDNAs can be amplified by PCR followed by cloning in a vector.

3.4.2.3 Examples and presentation of typical results

Using the RNA patterns method, it has been possible to visualise changes in the RNA patterns between normal, cirrhotic and fetal livers in humans (Chenchik *et al.*, 1993b). The data shown in figure 3.102 show differences in the poly(A)⁺ mRNA profile of normal and fetal livers after reverse transcription with primers F1, F6, P55, P41, P10, P35, P36 or P57 (for primer sequences, see table 3.11). Differences between normal and cirrhotic livers after reverse transcription were identified using primers F6 (bands 1 and 2), P55 (band 3), P35 (band 4), P36 (bands 5 and 6) and P57 (band 7). For primer P32 no differentially expressed genes were observed.

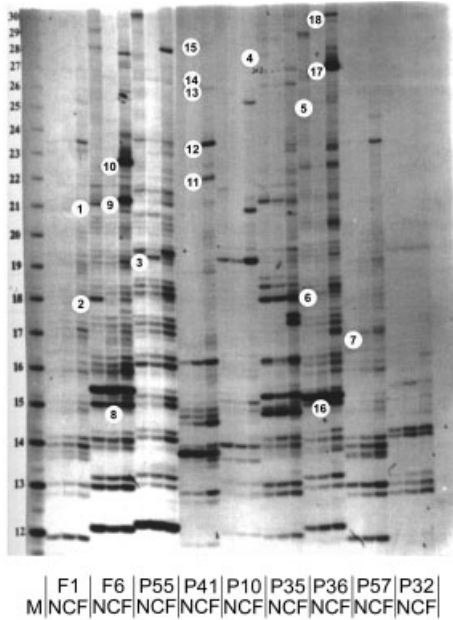


Figure 3.102 Typical results obtain by the RNA patterns method. Comparison of poly(A)⁺ mRNA patterns for normal (N), cirrhotic (C) and fetal (F) human livers using primers: F1, F6, P55, P41, P10, P35, P36, P57, P32 (for details, see table 3.11, and Chenchik *et al.*, 1993b). The resulting RNA pattern after cDNA synthesis with radioactive detection is shown. Differentially expressed poly(A)⁺ mRNAs are marked (number one to eighteen). A size standard is shown in lane M. The figure is reproduced from Chenchik *et al.* (1993b) with permission from *FEBS Letters*.

The RNA patterns method has also been used to monitor gene expression during differentiation of F9 fetal carcinoma cells (Chenchik *et al.*, 1993b). These cells can be differentiated to parietal endodermic cells by treatment with retinoic acid and dibutyryl cAMP (Strickland *et al.*, 1978). Using a set of 30 primers, at least ten differentially expressed mRNAs were identified including among the $\alpha 1$ chain of type IV collagen and laminin A mRNA that were exclusively expressed in differentiated F9 cells. These mRNAs are well-known markers of differentiated F9 cells (Goodfellow *et al.*, 1984; Sarvetnick *et al.*, 1989).

3.4.2.4 Discussion

The RNA patterns approach allows poly(A)⁺mRNA fingerprinting and the identification of differentially expressed mRNAs. It is claimed to be a sensitive method that detects rare mRNAs without an amplification step. Since PCR is not used, mRNA fingerprints displayed with this method should represent the true relative abundance of the mRNA population in the sample.

The choice of primers is very important. A ten-fold change in the number of bands can be achieved by a single nucleotide substitution around the 3'-end of the primer or by changing its length by a single nucleotide (Chenchik *et al.*, 1997).

As with other methods using polyacrylamide gel electrophoresis for displaying the mRNA expression profile, fragments with a similar length may co-migrate

so that excised bands may contain multiple cDNA fragments, making it difficult to identify differentially expressed mRNA and increasing the incidence of false-positives.

Although sensitivity of the RNA patterns method is sufficient for most purposes, use of samples with limited amounts of starting material such as dissected tissues or primary cells is limited due to the lack of an amplification step.

Overall, the RNA patterns method seems to be an approach that, at least in principle, allows displaying of mRNA fingerprints and thereby the identification of differential gene expression at the RNA level. However, the limited sensitivity of the RNA patterns method is likely to confine its use to experimental situations where the amount of starting material is not restricted.

3.5

Summary and conclusions

Each of the methods described in this chapter has its own niche and its particular advantages and disadvantages. Nevertheless, several general conclusions can be drawn. All hybridisation-based methods suffer from the problem that rare cDNA species in the tester population will not hybridise to a significant extent with their congener cDNA species in the driver population within any reasonable time-frame. This means that these methods inevitably suffer from a background false-positive rate. All PCR-based methods, on the other hand, suffer from the problems inherent to this amplification technique namely false priming, distortion of the apparent relative concentrations of starting material, and saturation effects at high cycle numbers. Further problems arise with all methods that include a step involving enzymatic degradation and annealing of adapters, for example, during the creation of linker molecules. Such hydrolysis and annealing steps are never completely efficient, nor are they ever entirely specific so that a background of false-positive results is inevitable with these methods, too. In addition, reverse transcription of mRNA to cDNA – a fundamental step of nearly all expression analysis experiments – is itself imperfect and may lead, for example, to under-representation of rare species.

As a guiding principle, it is therefore useful to confirm results obtained with one set of methods by means of a different set of methods. For example, differentially expressed genes revealed using a PCR-based method should generally be confirmed using a hybridisation-based approach.

3.6

References

Recommended readings

cDNA preparation and characterization.

Weissman, S., editor. *Methods in Enzymology* **1999**; 303.

Internet Resources

Brain EST Database (BED; Taisho Laboratory of Functional Genomics, Nara Institute of Science and Technology, Nara, Japan); available at <http://love2.aist-nara.ac.jp/>.

cDNA-AFLP (cDNA-based amplified-fragment length polymorphism fingerprinting) protocol (Laboratory of Plant Breeding, Wageningen University, Wageningen, The Netherlands); available at <http://www.dpw.wau.nl/pv/aflp/cDNA-AFLP%20Protocol.htm>.

DNA mfold (Rensselaer Polytechnic Institute, 331 Amos Eaton Hall, Troy, New York, USA); available at <http://bioinfo.math.rpi.edu/~mfold/dna/form1.cgi>.

Ordered differential display (ODD) protocols (Laboratory of Genes for Regeneration, Shemyakin & Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, Moscow, Russia); available at <http://www.ibch.ru/~lgr/protocols.htm>.

Q-RT-PCR software (Center for Human Genetics, Institute of Molecular Medicine, University of Texas at Houston, Texas, USA), available at <http://www.grad.ttuhsu.edu/archive/index.html>.

UniGene database (National Center for Biotechnology Information, NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/UniGene/>.

Cited publications

Al Taher, A., Bashein, A., Nolan, T., Hollingsworth, M., and Brady, G. Global cDNA amplification combined with real-time RT-PCR: accurate quantification of multiple human potassium channel genes at the single cell level. *Yeast* **2000**; *17*(3):201-210.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **1990**; *215*(3):403-410.

Adolfsson, J., Borge, O. J., Bryder, D., Theilgaard-Mönch, K., Åstrand-Grundström, I., Sitnicka, E., Sasaki, Y., and Jacobsen, S. E.

Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity* **2001**; *15*(4):659-669.

Alfke, H., Stumm, G., Schnieder, I., Klose, K. J., and Schlegel, J. Differential gene expression of vascular smooth muscle cells. Detection by RNA arbitrarily primed polymerase chain reaction. *Vasa* **1999**; *28*(3):164-168.

Allegra, P., Sterner, R., Clayton, D. F., and Allfrey, V. G. Affinity chromatographic purification of nucleosomes containing transcriptionally active DNA sequences. *Journal of Molecular Biology* **1987**; *196*(2):379-388.

Altman, J., and Bayer, S. A. Development of the cerebellar system: in relation to its evolution, structure, and functions. CRC Press, Boca Raton, Florida, USA **1996**.

Alwine, J. C., Kemp, D. J., and Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the USA* **1977**; *74*(12):5350-5354.

Andras, S. C., Power, J. B., Cocking, E. C., and Davey, M. R. Strategies for signal amplification in nucleic acid detection. *Molecular Biotechnology* **2001**; *19*(1):29-44.

Antalis, T. M., and Godbolt, D. Isolation of intact nuclei from hematopoietic cell types. *Nucleic Acids Research* **1991**; *19*(15):4301.

Antonarakis, S. E., Lyle, R., Chrast, R., and Scott, H. S. Differential gene expression studies to explore the molecular pathophysiology of Down syndrome. *Brain Research: Brain Research Reviews* **2001**; *36*(2-3):265-274.

Antony, T., Atreyi, M., and Rao, M. V. Spectroscopic studies on the binding of methylene blue to poly(riboadenylic acid). *Journal of Biomolecular Structure and Dynamics* **1993**; *11*(1):67-81.

Aryee, D. N. T., Sommergruber, W., Muehlbacher, K., Dockhorn-Dworniczak, B., Zoubek, A., and Kovar, H. Variability in gene expression patterns of Ewing tumor cell lines differing in EWS-FLI1 fusion type. *Laboratory Investigation* **2000**; *80*(12):1833-1844.

- Averboukh, L., Douglas, S. A., Zhao, S., Lowe, K., Maher, J., and Pardee, A. B. Better gel resolution and longer cDNAs increase the precision of differential display. *Biotechniques* **1996**; 20(5):918-921.
- Babiss, L. E., Zimmer, S. G., and Fisher, P. B. Reversibility of progression of the transformed phenotype in Ad5-transformed rat embryo cells. *Science* **1985**; 228(4703):1099-1101.
- Bachem, C. W. B., van der Hoeven, R. S., de Bruijn, S. M., Vreugdenhil, D., Zabeu, M., and Visser, R. G. F. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. *Plant Journal* **1996**; 9(5):745-753.
- Bachem, C. W. B., Oomen, R. J. F. J., and Visser, R. G. F. Transcript imaging with cDNA-AFLP: A step-by-step protocol. *Plant Molecular Biology Reporter* **1998**; 16(2):157-173.
- Bachem, C. W. B., van der Hoeven, R. S., Luckert, J., Oomen, R. J. F. J., Casarini, E., Jacobsen, E., and Visser, R. G. F. Functional genomic analysis of potato tuber life-cycle. *Potato Research* **2000a**; 43(4):297-312.
- Bachem, C. W. B., Oomen, R. J. F. J., Kuyt, S., Horvath, B. M., Claassens, M. M. J., Vreugdenhil, D., and Visser, R. G. F. Antisense suppression of a potato alpha-SNAP homologue leads to alterations in cellular development and assimilate distribution. *Plant Molecular Biology* **2000b**; 43(4):473-482.
- Bachem, C. W. B., Horvath, B. M., Trindade, L., Claassens, M., Davelaar, E., Jordi, W., and Visser, R. G. F. A potato tuber expressed mRNA with homology to steroid dehydrogenases affects gibberellin levels and plant development. *Plant Journal* **2001**; 25(6):595-604.
- Bailey, J. M., and Davidson, N. Methylmercury as a reversible denaturing agent for agarose gel electrophoresis. *Analytical Biochemistry* **1976**; 70(1):75-85.
- Barabas, K., and Molnar, J. Lack of correlation between intercalation and plasmid curing ability of some tricyclic compounds. *Acta Microbiologica Academiae Scientiarum Hungaricae* **1980**; 27(1):55-61.
- Barnes, W. M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the USA* **1994**; 91(6):2216-2220.
- Becker, P., Hufnagle, W., Peters, G., and Herrmann, M. Detection of differential gene expression in biofilm-forming versus planktonic populations of *Staphylococcus aureus* using micro-representational-difference analysis. *Applied and Environmental Microbiology* **2001**; 67(7):2958-2965.
- Becker-Andre, M., and Harlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTT). *Nucleic Acids Research* **1989**; 17(22):9437-9446.
- Bauer, D., Muller, H., Reich, J., Riedel, H., Ahrenkiel, V., Warthoe, P., and Strauss, M. Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research* **1993**; 21(18):4272-4280.
- Beck, M. T., Holle, L., and Chen, W. Y. Combination of PCR subtraction and cDNA microarray for differential gene expression profiling. *Biotechniques* **2001**; 31(4):782-784, 786.
- Belyavsky, A., Vinogradova, T. and Rajewsky, K. PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. *Nucleic Acids Research* **1989**; 17(8):2919-2932.
- Belyavsky, A. V., Shmelkov, S. V., and Visser, J. W. M. Identification of differentially expressed genes in sorted cell populations by two-dimensional gene-expression fingerprinting. In: Hematopoietic stem cell protocols. Klug, C. A., and Jordan C. T., editors. Humana Press, Totowa, New Jersey, USA **2002**:301-325.
- Berk, A. J., and Sharp, P. A. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **1977**; 12(3):721-732.
- Bertioli, D. J., Schlichter, U. H., Adams, M. J., Burrows, P. R., Steinbiss, H. H., and Antoniw, J. F. An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research* **1995**; 23(21):4520-4523.
- Biezen, v. d. E. A., Juwana, H., Parker, J. E., and Jones, J. D. cDNA-AFLP display for the isolation of *Peronospora parasitica* genes expressed during infection in *Arabidopsis thaliana*. *Molecular Plant Microbe Interactions* **2000**; 13(8):895-898.

- Billia, F., Barbara, M., McEwen, J., Trevisan, M. and Iscove, N. N. Resolution of pluripotential intermediates in murine hematopoietic differentiation by global complementary DNA amplification from single cells: confirmation of assignments by expression profiling of cytokine receptor transcripts. *Blood* **2001**; 97(8):2257-2268.
- Boedtker, H. The reaction of ribonucleic acid with formaldehyde. I. Optical absorbance studies. *Biochemistry* **1967**; 6(9): 2718-2727.
- Boffa, L. C., Walker, J., Chen, T. A., Sterner, R., Mariani, M. R., and Allfrey, V. G. Factors affecting nucleosome structure in transcriptionally active chromatin. Histone acetylation, nascent RNA and inhibitors of RNA synthesis. *European Journal of Biochemistry* **1990**; 194(3):811-823.
- Bonaldo, M. F., Lennon, G., and Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research* **1996**; 6(9):791-806.
- Bongrazio, M., Grafe, M., Pries, A. R., Gaetgens, P., and Zakrzewicz, A. Improvement of RNA fingerprinting efficiency for the analysis of differential gene expression in human cardiac macro- and microvascular endothelial cells. *Pharmacological Research* **2001**; 43(6):553-560.
- Bonnet, G., Tyagi, S., Libchaber, A., and Kramer, F. R. Thermodynamic basis of the enhanced specificity of structured DNA probes. *Proceedings of National Academy of Sciences of the USA* **1999**; 96(11):6171-6176.
- Bowler, L. D., Hubank, M., and Spratt, B. G. Representational difference analysis of cDNA for the detection of differential gene expression in bacteria: development using a model of iron-regulated gene expression in *Neisseria meningitidis*. *Microbiology* **1999**; 145(Part 12):3529-3537.
- Bowler, L. D., Bart, A., and van der Ende, A. Meningococcal disease. Humana Press, Totowa, New Jersey, USA **2001**.
- Brady, G., Barbara, M., and Iscove, N. N. Representative *in vitro* cDNA amplification from individual hematopoietic cells and colonies. *Methods in Molecular and Cellular Biology* **1990**; 2:17-25.
- Brady, G., and Iscove, N. N. Construction of cDNA libraries from single cells. *Methods in Enzymology* **1993**; 225:611-623.
- Brady, G., Billia, F., Knox, J., Hoang, T., Kirsch, I. R., Voura, E. B., Hawley, R. G., Cumming, R., Buchwald, M., and Siminovitch, K. Analysis of gene expression in a complex differentiation hierarchy by global amplification of cDNA from single cells. *Current Biology* **1995**; 5(8):909-922.
- Brady, G. Expression profiling of single mammalian cells – small is beautiful. *Yeast* **2000**; 17(3):211-217.
- Brazma, A., and Vilo, J. Gene expression data analysis. *FEBS Letters* **2000**; 480(1):17-24.
- Breslauer, K. J., Frank, R., Blöcker, H., and Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences of the USA* **1986**; 83(11):3346-3750.
- Breyne, P., and Zabeau, M. Genome-wide expression analysis of plant cell cycle modulated genes. *Current Opinions in Biotechnology* **2001**; 4(2):136-142.
- Britten, R. J., and Kohne, D. E. Repeated sequences in DNA: hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **1968**; 161(841):529-540.
- Brown, T., and Mackey, K. Analysis of RNA by northern and slot blot hybridization. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1997**:4.9.1-4.9.16.
- Brown, J., Hillarby, C., and Brandwood, C. Use of poly A RT-PCR coupled with subtractive hybridisation to isolate novel genes involved in osteogenesis. *Journal of Bone and Mineral Research* **1997**; 12(Supplement 1):S281.
- Brown, A. J., Hutchings, C., Burke, J. F., and Mayne, L. V. Targeted display: a new technique for the analysis of differential gene expression. *Methods in Enzymology* **1999a**; 303:392-408.
- Brown, A. J., Hutchings, C., Burke, J. F., and Mayne, L. V. Application of a rapid method (targeted display) for the identification of differentially expressed mRNAs following NGF-induced neuronal differentiation in PC12 cells. *Molecular and Cellular Neuroscience* **1999b**; 13(2):119-130.
- Brugmans, B., Fernandez del Carmen, A., Bachem, C. W. B., Van Os, H., Van Eck, H. J., and Visser, R. G. F. A novel method for the construction of genome wide transcriptome maps. *The Plant Journal* **2002**; 31(2): 211-222.

- Bryant, Z., Subrahmanyam, L., Tworoger, M., LaTray, L., Liu, C. R., Li, M. J., van den Engh, G., and Ruohola-Baker, H. Characterization of differentially expressed genes in purified *Drosophila* follicle cells: toward a general strategy for cell type-specific developmental analysis. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(10):5559-5564.
- Buckbinder, L., and Brown, D. D. Thyroid hormone-induced gene expression changes in the developing frog limb. *Journal of Biological Chemistry* **1992**; 267(36):25786-25791.
- Burn, T. C., Petrovick, M. S., Hohaus, S., Rollins, B. J., and Tenen, D. G. Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood* **1994**; 84(8):2776-27783.
- Bustin, S. A., and Dorudi, S. Molecular assessment of tumour stage and disease recurrence using PCR-based assays. *Molecular Medicine Today* **1998**; 4(9):389-396.
- Bustin, S. A., Gyselman, V. G., Williams, N. S., and Dorudi, S. Detection of cytokeratins 19/20 and guanylyl cyclase C in peripheral blood of colorectal cancer patients. *British Journal of Cancer* **1999**; 79(11-12):1813-1820.
- Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology* **2000**; 25(2):169-193.
- Bustin, S. A., Li, S. R., Phillips, S., and Dorudi, S. Expression of HLA class II in colorectal cancer: evidence for enhanced immunogenicity of microsatellite-instability-positive tumours. *Tumour Biology* **2001**; 22(5):294-298.
- Butler, E. T., and Chamberlin, M. J. Bacteriophage SP6-specific RNA polymerase. I. Isolation and characterization of the enzyme. *Journal of Biological Chemistry* **1982**; 257(10):5772-5778.
- Byers, R. J., Hoyland, J. A., Dixon, J., and Freemont, A. J. Subtractive hybridization – genetic takeaways and the search for meaning. *International Journal of Experimental Pathology* **2000**; 81(6):391-404.
- Caetano-Anolles, G. Scanning of nucleic acids by *in vitro* amplification: new developments and applications. *Nature Biotechnology* **1996**; 14(13):1666-1674.
- Calkhoven, C. F., and Ab, G. Multiple steps in the regulation of transcription factor level and activity. *Biochemical Journal* **1996**; 317(Part 2):329-342.
- Callard, D., Lescure, B., and Mazzolini, L. A method for the elimination of false positives generated by the mRNA differential display technique. *Biotechniques* **1994**; 16(6):1096-1103.
- Campalans, A., Pages, M., and Messegue, R. Identification of differentially expressed genes by the cDNA-AFLP technique during dehydration of almond (*Prunus amygdalus*). *Tree Physiology* **2001**; 21(10):633-643.
- Candalieri, G. A., Rao, Y., Floh, A., Sandler, S. D., and Aubin, J. E. cDNA fingerprinting of osteoprogenitor cells to isolate differentiation stage-specific genes. *Nucleic Acid Research* **1999**; 27(4):1079-1083.
- Carulli, J. P., Artinger, M., Swain, P. M., Root, C. D., Chee, L., Tulig, C., Guerin, J., Osborne, M., Stein, G., Lian, J., and Lomedico, P. T. High throughput analysis of differential gene expression. *Journal of Cellular Biochemistry* **1998**; 72(Supplement 30-31):286-296.
- Casal, J., and Leptin, M. Identification of novel genes in *Drosophila* reveals the complex regulation of early gene activity in the mesoderm. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(19):10327-10332.
- Chalbos, D., Vignon, F., Keydar, I., and Rochefort, H. Estrogens stimulate cell proliferation and induce secretory proteins in a human breast cancer cell line (T47D). *Journal of Clinical Endocrinology and Metabolism* **1982**; 55(2):276-283.
- Chamberlin, M., and Ryan, T. Bacteriophage DNA-dependent RNA polymerase. In: The enzymes. Boyer, P. D., Lardy, H., and Myrback, K., editors. Academic Press, New York, New York, USA **1982**; XV:87-108.
- Chen, T. A., and Allfrey, V. G. Rapid and reversible changes in nucleosome structure accompany the activation, repression, and superinduction of murine fibroblast proto-oncogenes *c-fos* and *c-myc*. *Proceedings of the National Academic Sciences of the USA* **1987**; 84(15):5252-5256.
- Chen, T. A., Sterner, R., Cozzolino, A., and Allfrey, V. G. Reversible and irreversible changes in nucleosome structure along the *c-fos* and *c-myc* oncogenes following inhibition of transcription. *Journal of Molecular Biology* **1990**; 212(3):481-493.
- Chen-Cleland, T. A., Boffa, L. C., Carpaneto, E. M., Mariani, M. R., Valentin, E., Mendez, E., and Allfrey, V. G. Recovery of transcrip-

- tionally active chromatin restriction fragments by binding to organomercurial-agarose magnetic beads. A rapid and sensitive method for monitoring changes in higher order chromatin structure during gene activation and repression. *Journal of Biological Chemistry* **1993**; 268(31):23409-23416.
- Chenchik, A. A., Diachenko, L. B. and Beaealashvili, R. S. Analysis of poly(A)⁺ RNA patterns in human tissues. *Federation of European Biochemical Societies* **1993a**; 321(1):98-101.
- Chenchik, A. A., Diachenko, L. B. and Beaealashvili, R. S. Application of poly(A)⁺ RNA patterns method for searching of differentially expressed genes. *Federation of European Biochemical Societies* **1993b**; 324(2):136-139.
- Chenchik, A., Diachenko, L., Moqadam, F., Tarabykin, V., Lukyanov, S. and Siebert, P. D. Full-length cDNA cloning and determination of mRNA 5' and 3' ends by amplification of adaptor-ligated cDNA. *Biotechniques* **1996**; 21(3):526-534.
- Chenchik, A. A., Diachenko, L. B., Beaealashvili, R. S., and Carter, C. J. RNA fingerprinting to determine RNA population differences. *US5665544* **1997**.
- Cheng, T., Shen, H., Giokas, D., Gere, J., Tenen, D. G., and Scadden, D. T. Temporal mapping of gene expression levels during the differentiation of individual primary hematopoietic cells. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(23):13158-13163.
- Chianale, J., Vollrath, V., Wielandt, A. M., Miranda, S., Gonzales, R., Fresno, M., Quintana, C., Gonzales, S., Andrade, L., and Guzman, S. Differences between nuclear run-off and mRNA levels for multidrug resistance gene expression in the cephalocaudal axis of the mouse intestine. *Biochimica Biophysica Acta* **1995**; 1264(3):369-376.
- Choi, S. C., Kim, J., and Han, J. K. Identification and developmental expression of par-6 gene in *Xenopus laevis*. *Mechanisms of Development* **2000**; 91(1-2):347-350.
- Choi, S. C., Chang, J. Y., and Han, J. K. A novel *Xenopus* acetyltransferase with a dynamic expression in early development. *Biochemical and Biophysical Research Communications* **2001**; 285(5):1338-1343.
- Chomczynski, P., and Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry* **1987**; 162(1):156-159.
- Chou, Q., Russell, M., Birch, D. E., Raymond, J., and Bloch, W. Prevention of pre-PCR mispriming and primer dimerisation improves low-copy-number amplifications. *Nucleic Acids Research* **1992**; 20(7):1717-1723.
- Chu, C. W., Hwang, S. J., Luo, J. C., Wang, Y. J., Lu, R. H., Lai, C. R., Tsay, S. H., Wu, J. C., Chang, F. Y., and Lee, S. D. Comparison of clinical, virologic and pathologic features in patients with acute hepatitis B and C. *Journal of Gastroenterology and Hepatology* **2001**; 16(2):209-214.
- Collins, M. L., Irvine, B., Tyner, D., Fine, E., Zayati, C., Chang, C., Horn, T., Ahle, D., Detmer, J., Shen, L. P., Kolberg, J., Bushnell, S., Urdea, M. S., and Ho, D. D. A branched DNA signal amplification assay for quantification of nucleic acid targets below 100 molecules/ml. *Nucleic Acids Research* **1997**; 25(15):2979-2984.
- Cooper, B. Collateral gene expression changes induced by distinct plant viruses during the hypersensitive resistance reaction in *Chenopodium amaranticolor*. *Plant Journal* **2001**; 26(3):339-349.
- Corcelette, S., Masse, T., and Madjar, J. J. Initiation of translation by non-AUG codons in human T-cell lymphotropic virus type I mRNA encoding both Rex and Tax regulatory proteins. *Nucleic Acids Research* **2000**; 28(7):1625-1634.
- Cui, K., Feldman, L., and Sytkowski, A. J. Isolation of differentially expressed genes by cloning transcriptionally active DNA fragments. *Methods* **1999**; 17(3):265-271.
- Daele, V. I., Loose, d. M., Bockstaele, v. E., and Roldan, R. I. Molecular analysis of self-incompatibility in ryegrass. *Mededelingen Faculteit Landbouwkundige en Toegepaste Biologische Wetenschappen Universiteit Gent* **2000**; 65(3b):443-446.
- De la Monte, S. M., Ng, S. C., and Hsu, D. W. Aberrant GAP-43 gene expression in Alzheimer's disease. *American Journal of Pathology* **1995**; 147(4):934-946.
- Debono, E., Halfon, P., Bourliere, M., Gerolami-Santandrea, V., Gastaldi, M., Castellani, P., Cartouzou, G., Botta-Fridlund, D., Cau, P., and Gauthier, A. Absence of hepatitis C genome in semen of infected men by polymerase chain reaction, branched DNA and *in situ* hybridization. *Liver* **2000**; 20(3):257-261.

- Debouck, C. Differential display or differential dismay. *Current Opinion in Biotechnology* **1995**; 6(5):597-599.
- Dellagi, A., Birch, P. R., Heilbronn, J., Lyon, G. D., and Toth, I. K. cDNA-AFLP analysis of differential gene expression in the prokaryotic plant pathogen. *Erwinia carotovora*. *Microbiology* **2000**; 146(Part 1):165-171.
- Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D., and Siebert, P. D. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(12):6025-6030.
- Dimaculangan, D. D., Chawla, A., Boak, A., Kagan, H. M., and Lazar, M. A. Retinoic acid prevents downregulation of ras resection gene/lysyl oxidase early in adipocyte differentiation. *Differentiation* **1994**; 58(1):47-52.
- Dixon, A. K., Richardson, P. J., Lee, K., Carter, N. P., and Freeman, T. C. Expression profiling of single cells using 3 prime end amplification (TPEA) PCR. *Nucleic Acids Research* **1998**; 26(19):4426-4431.
- Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K., and Mattick, J. S. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research* **1991**; 19(14):4008.
- Doris, P. A., Hayward-Lester, A., and Hays, J. K., Sr. Q-RT-PCR: data analysis software for measurement of gene expression by competitive RT-PCR. *Computer Application in Biosciences* **1997**; 13(6):587-591.
- Doris, P. A., Oefner, P. J., Chilton, B. S., and Hayward-Lester, A. Quantitative analysis of gene expression by ion-pair high-performance liquid chromatography. *Journal of Chromatography. A* **1998**; 806(1):47-60.
- Doyu, M., Sawada, K., Mitsuma, N., Niwa, J., Yoshimoto, M., Fujii, Y., Sobue, G., and Kato, K. Gene expression profile in Alzheimer's brain screened by molecular indexing. *Brain Research. Molecular Brain Research* **2001**; 87(1):1-11.
- Duggen, J. D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. Expression profiling using cDNA microarrays. *Nature Genetics* **1999**; 21(1 Supplement):10-14.
- Duguid, J. R., Rohwer, R. G., and Seed, B. Isolation of cDNAs of scrapie-modulated RNAs by subtractive hybridization of a cDNA library. *Proceedings of the National Academy of Sciences of the USA* **1988**; 85(15):5738-5742.
- Duguid, J. R., and Dinanuer, M. C. Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research* **1990**; 18(9):2789-2792.
- Dunbar, S. A., and Jacobson, J. W. Application of the Luminex LabMAP in rapid screening for mutations in the cystic fibrosis transmembrane conductance regulator gene: a pilot study. *Clinical Chemistry* **2000**; 46(9):1498-1500.
- Durrant, W. E., Rowland, O., Piedras, P., Hammond-Kosack, K. E., and Jones, J. D. cDNA-AFLP reveals a striking overlap in race-specific resistance and wound response gene expression profiles. *The Plant Cell* **2000**; 12(6):963-977.
- Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettl, M., and Coleman, P. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the USA* **1992**; 89(7):3010-3014.
- Ekmekcioglu, S., Ellerhorst, J., Mhashilkar, A. M., Sahin, A. A., Read, C. M., Prieto, V. G., Chada S., and Grimm, E. A. Down-regulated melanoma differentiation associated gene (*mda-7*) expression in human melanomas. *International Journal of Cancer* **2001**; 94(1):54-59.
- El-Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W., and Vogelstein, B. WAF1, a potential mediator of p53 tumor suppression. *Cell* **1993**; 75(4):817-825.
- Enomoto, M., Nishiguchi, S., Shiomi, S., Tanaka, M., Fukuda, K., Ueda, T., Tamori, A., Habu, D., Takeda, T., Yano, Y., and Otani, S. Comparison of real-time quantitative polymerase chain reaction with three other assays for quantitation of hepatitis C virus. *Journal of Gastroenterology and Hepatology* **2001**; 16(8):904-909.
- Fink, L., Seeger, W., Ermert, L., Hanze, J., Stahl, U., Grimminger, F., Kummer, W., and Bohle, R. M. Real-time quantitative RT-PCR after laser-assisted cell picking. *Nature Medicine* **1998**; 4(11):1329-1333.
- Fink, L., Kinfe, T., Stein, M. M., Ermert, L., Hanze, J., Kummer, W., Seeger, W., and Bohle, R. M. Immunostaining and laser-as-

- sisted cell picking for mRNA analysis. *Laboratory Investigation* **2000**; 80(3):327-333.
- Fischer, A., Saedler, H., and Theissen, G. Restriction fragment length polymorphism-coupled domain-directed differential display: a highly efficient technique for expression analysis of multigene families. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(12):5331-5335.
- Fisher, P. B., Prignoli, D. R., Hermo, H., Jr., Weinstein, I. B., and Pestka, S. Effects of combined treatment with interferon and mezerein on melanogenesis and growth in human melanoma cells. *Journal of Interferon Research* **1985**; 5(1):11-22.
- Fisher, P. B., Hermo, H., Jr., Solowey, W. E., Dietrich, M. C., Edwards, G. M., Weinstein, I. B., Langer, J. A., Pestka, S., Giacomini, P., Kusama, M., and Ferrone, S. Effect of recombinant human fibroblast interferon and mezerein on growth, differentiation, immune interferon binding and tumor associated antigen expression in human melanoma cells. *Anticancer Research* **1986**; 6(4):765-774.
- Flood, J., Drew, W. L., Miner, R., Jekic-McMullen, D., Shen, L. P., Kolberg, J., Garvey, J., Follansbee, S., and Poscher, M. Diagnosis of cytomegalovirus (CMV) polyradiculopathy and documentation of *in vivo* anti-CMV activity in cerebrospinal fluid by using branched DNA signal amplification and antigen assays. *Journal of Infectious Diseases* **1997**; 176(2):348-352.
- Fraenkel-Conrat, H. Reaction of nucleic acids with formaldehyde. *Biochimica et Biophysica Acta* **1954**; 15(1):307-309.
- Franz, O., Roeder, T., and Gewecke, M. Analysis of differential gene expression in the central nervous system of *Schistocerca gregaria* by differential display PCR. *Journal of Comparative Physiology* **1998**; 182(5):627-633.
- Franz, O., Bruchhaus, I., and Roeder, T. Verification of differential gene transcription using virtual Northern blotting. *Nucleic Acids Research* **1999**; 27(1):e3.
- Frohman, M., Dush, M. K., and Martin, G. R. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the USA* **1988**; 85(23):8998-9002.
- Fuchs, B., Zhang, K., Bolander, M. E., and Sarkar, G. Identification of differentially expressed genes by mutually subtracted RNA fingerprinting. *Analytical Biochemistry* **2000a**; 286(1):91-98.
- Fuchs, B., Zhang, K., Bolander, M. E., and Sarkar, G. Differential mRNA fingerprinting by preferential amplification of coding sequences. *Gene* **2000b**; 258(1-2):155-163.
- Fuchs, B., Zhang, K., Schabel, A., Bolander, M. E., and Sarkar, G. Identification of twenty-two candidate markers for human osteogenic sarcoma. *Gene* **2001**; 278(1-2):245-252.
- Fujimura, T., and Murakami, K. Increase of methicillin resistance in *Staphylococcus aureus* caused by deletion of a gene whose product is homologous to lytic enzymes. *Journal of Bacteriology* **1997**; 179(20):6294-6301.
- Fukuda, T., Kido, A., Kajino, K., Tsutsumi, M., Miyauchi, Y., Tsujiuchi, T., Konishi, Y., and Hino, O. Cloning of differentially expressed genes in highly and low metastatic rat osteosarcomas by a modified cDNA-AFLP method. *Biochemical and Biophysical Research Communications* **1999**; 261(1):35-40.
- Fulton, R. J., McDade, R. L., Smith, P. L., Kienker, L. J., and Kettman, J. R., Jr. Advanced multiplexed analysis with FlowMetrix™ system. *Clinical Chemistry* **1997**; 43(9):1749-1756.
- Gao, L., Cueto, M. A., Asselbergs, F., and Atadja, P. Cloning and functional characterization of HDAC11, a novel member of the human histone deacetylase family. *Journal of Biological Chemistry* **2002**; 277(28):25748-25755.
- Geck, P., Szelei, J., Jimenez, J., Sonnenschein, C., and Soto, A. M. Early gene expression during androgen-induced inhibition of proliferation of prostate cancer cells: A new suppressor candidate on chromosome 13, in the BRCA2-Rb1 locus. *Journal of Steroid Biochemistry and Molecular Biology* **1999**; 68(1-2):41-50.
- Geijsen, N., Uings, I. J., Pals, C., Armstrong, J., McKinnon, M., Raaijmakers, J. A. M., Lammers, J.-W. J., Koenderman, L., and Coffey, P. J. Cytokine-specific transcriptional regulation through an IL-5R α interacting protein. *Science* **2001**; 293(5532):1136-1138.
- Gravesen, A., Warthoe, P., Knochel, S., and Thirstrup, K. Restriction fragment differential display of pediocin-resistant *Listeria monocytogenes* 412 mutants shows consistent overexpression of a putative beta-glucoside-specific PTS system. *Microbiology* **2000**; 146(6):1381-1389.

- Gravesen, A., Sorensen, K., Aarestrup, F. M., and Knochel, S. Spontaneous nisin-resistant *Listeria monocytogenes* mutants with increased expression of a putative penicillin-binding protein and their sensitivity to various antibiotics. *Microbial Drug Resistance* **2001**; 7(2):127-135.
- Gray, J. W., and Collins, C. Genome changes and gene expression in human solid tumors. *Carcinogenesis* **2000**; 21(3):443-452.
- Greenberg, M. E., and Bender, T. P. Identification of newly transcribed RNA. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1997**:4.10.1-4.10.11.
- Greenberg, S. A. DNA microarray gene expression analysis technology and its application to neurological disorders. *Neurology* **2001**; 57(5):755-761.
- Grootjans, J. J., Zimmermann, P., Reekmans, G., Smets, A., Degeest, G., Durr, J., and David, G. Syntenin, a PDZ protein that binds syndecan cytoplasmic domains. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(25):13683-13688.
- Guan, Z., Lu, L., Zheng, Z., Liu, J., Yu, F., Lu, S., Xin, Y., Liu, X., Hong, J., and Zhang, W. A spontaneous recurrent seizure-related *Rattus* NSF gene identified by linker capture subtraction. *Brain Research: Molecular Brain Research* **2001**; 87(1):117-123.
- Gubbay, J., Doyle, J. P., Skinner, M., and Heintz, N. Changing patterns of gene expression identify multiple steps during regression of rat prostate *in vivo*. *Endocrinology* **1998**; 139(6):2935-2943.
- Gubler U., and Hoffman, B. J. A simple and very efficient method for generating cDNA libraries. *Gene* **1983**; 25(2-3):263-269.
- Goodfellow, P. N. The use of cloned gene probes to study differentiation in teratocarcinomas. *Cell Differentiation* **1984**; 15(2-4):257-267.
- Hakvoort, T. B., Leegwater, A. C., Michiels, F. A., Chamuleau, R. A., and Lamers, W. H. Identification of enriched sequences from a cDNA subtraction-hybridization procedure. *Nucleic Acids Research* **1994**; 22(5):878-879.
- Hampson, I. N., Pope, L., Cowling, G. J., and Dexter, T. M. Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridization probes. *Nucleic Acids Research* **1992**; 20(11):2899.
- Hampson, I. N., and Hampson, L. Chemical crosslinking subtraction (CCLS). *Methods in Molecular and Cellular Biology* **1997**; 68:221-236.
- Han, E., and Hilsenbeck, S. G. Array-based gene expression profiling to study aging. *Mechanism of Ageing and Development* **2001**; 122(10):999-1018.
- Hara, E., Kato, T., Nakada, S., Sekiya, L. and Oda, K. Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research* **1991**; 19(25):7097-7104.
- Harris, S. A., Enger, R. J., Riggs, B. L., and Spelsberg, T. C. Development and characterization of a conditionally immortalized human fetal osteoblastic cell line. *Journal of Bone and Mineral Research* **1995**; 10(2):178-186.
- Hartley, D. P., and Klaassen, C. D. Detection of chemical-induced differential expression of rat hepatic cytochrome P450 mRNA transcripts using branched DNA signal amplification technology. *Drug Metabolism and Disposition* **2000**; 28(5):608-616.
- Haselkorn, R., and Doty, P. The reaction of formaldehyde with polynucleotides. *Journal of Biological Chemistry* **1961**; 236(10):2738-2745.
- Hatada, I., Hayashizaki, Y., Hirotsune, S., Komatsubara, H., and Mukai, T. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(21):9523-9527.
- Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Hatada, I., Shibata, H., Kawai, J., Hirose, K., Watanabe, S., Fushiki, S., Wada, S., Sugimoto, T., Kobayakawa, K., Kawara, T., Kat-suki, M., Sibuya, T., and Mukai, T. Restriction landmark genomic scanning method and its various applications. *Electrophoresis* **1993**; 14(4):251-258.
- Hayward-Lester, A., Oefner, P. J., Sabatini, S., and Doris, P. A. Accurate and absolute quantitative measurement of gene expression by single-tube RT-PCR and HPLC. *Genome Research* **1995**; 5(5):494-499.
- Hayward-Lester, A., Oefner, P. J., and Doris, P. A. Rapid quantification of gene expression by competitive RT-PCR and ion-pair reversed-phase HPLC. *Biotechniques* **1996**; 20(2):250-257.

- Heid, C. A., Stevens, J., Livak, K. J., and Williams, P. M. Real time quantitative PCR. *Genome Research* **1996**; 6(10):986-994.
- Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Wooley, D. E., and Davis, R. W. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(6):2150-2155.
- Hendriks, T., Vreugdenhil, D., and Stiekema, W. J. Patatin and four serine proteinase inhibitor genes are differentially expressed during potato tuber development. *Plant Molecular Biology* **1991**; 17(3):385-394.
- Hennig, H., Luhm, J., Hartwig, D., Kluter, H., and Kirchner, H. A novel RT-PCR for reliable and rapid HCV RNA screening of blood donations. *Transfusion* **2001**; 41(9):1100-1106.
- Heremans, H., Billiau, A., Cassiman, J. J., Mulier, J. C., and de Somer, P. *In vitro* cultivation of human tumor tissues. II. Morphological and virological characterization of three cell lines. *Oncology* **1978**; 35(6):246-252.
- Herrin, D. L., and Schmidt, G. W. Rapid, reversible staining of Northern blots prior to hybridization. *Biotechniques* **1988**; 6(3):196-200.
- Higuchi, R., Dollinger, G., Walsh, P. S., and Griffith, R. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology (New York)* **1992**; 10(4):413-417.
- Higuchi, R., Fockler, C., Dollinger, G., and Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology (New York)* **1993**; 11(9):1026-1030.
- Hillarby, M. C., King, K. E., Brady, G., Grant, M. E., Wallis, G. A., and Boot-Handfort, R. P. Localization of gene expression during endochondral ossification. *Annals of the New York Academy of Sciences* **1996**; 785:263-266.
- Hirai, M. Y., Suzuki, H., Yamazaki, M., and Saito, K. Biochemical and partial molecular characterization of bitter and sweet forms of *Lupinus angustifolius*, an experimental model for study of molecular regulation of quinolizidine alkaloid biosynthesis. *Chemical Pharmaceutical Bulletins* **2000**; 48(10):1458-1461.
- Hirate, Y., Mieda, M., Harada, T., Yamasu, K., and Okamoto, H. Identification of ephrin-A3 and novel genes specific to the midbrain-MHB in embryonic zebrafish by ordered differential display. *Mechanisms of Development* **2001**; 107(1-2):83-96.
- Hirose, T., Ideue, T., Wakasugi, T., and Sugiura, T. The chloroplast *infA* gene with a functional UUG initiation codon. *FEBS Letters* **1999**; 445(1):169-172.
- Hoard, D. E. The applicability of formal titration to the problem of end-groups determinations in polynucleotides. A preliminary investigation. *Biochimica. et Biophysica Acta* **1960**; 40(1):62-70.
- Holland, P. M., Abramson, R. D., Watson, R., and Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(16):7276-7280.
- Horn, T., and Urdea, M. S. Forks and combs and DNA: the synthesis of branched oligodeoxyribonucleotides. *Nucleic Acids Research* **1989**; 17(17):6959-6967.
- Horn, T., Chang, C. A., and Urdea, M. S. Chemical synthesis and characterization of branched oligodeoxyribonucleotides (bdNA) for use as signal amplifiers in nucleic acid quantification assays. *Nucleic Acids Research* **1997**; 25(23):4842-4849.
- Huang, F., Adelman, J., Jiang, H., Goldstein, N. I., and Fisher, P. B. Differentiation induction subtraction hybridization (DISH): a strategy for cloning genes displaying differential expression during growth arrest and terminal differentiation. *Gene* **1999a**; 236(1):125-131.
- Huang, F., Adelman, J., Jiang, H., Goldstein, N. I., and Fisher, P. B. Identification and temporal expression pattern of genes modulated during irreversible growth arrest and terminal differentiation in human melanoma cells. *Oncogene* **1999b**; 18(23):3546-3552.
- Huang, E. Y., Madireddi, M. T., Gopalkrishnan, R. V., Leszczyniecka, M., Su, Z., Lebedeva, I. V., Kang, D., Jiang, H., Lin, J. J., Alexandre, D., Chen, Y., Vozhilla, N., Mei, M. X., Christiansen, K. A., Sivo, F., Goldstein, N. I., Mhashilkar, A. B., Chada, S., Huberman, E., Pestka, S., and Fisher, P. B. Genomic structure, chromosomal localization and expression profile of a novel melanoma differentiation associated (*mda-7*) gene with cancer specific growth suppressing and apoptosis inducing properties. *Oncogene* **2001**; 20(48):7051-7063.
- Huang, J., Bachem, C. W. B., Jacobsen E., and Visser, R. G. F. Molecular analysis of differ-

- entially expressed genes during postharvest deterioration in cassava (*Manihot esculenta* Crantz) tuberous roots. *Euphytica* **2001**; *120*(1):85-93.
- Hubank, M., and Schatz, D. G. Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Research* **1994**; *22*(25):5640-5648.
- Hubank, M., and Schatz, D. G. cDNA representational difference analysis: a sensitive and flexible method for identification of differentially expressed genes. *Methods in Enzymology* **1999**; *303*:325-349.
- Hwang, S. J., Lu, R. H., Wood, M. L., Wang, Y. J., Chang, F. Y., and Lee, S. D. Comparison of the nucleic acid-based crosslinking hybridization assay and the branched DNA signal amplification assay in the quantitative measurement of serum hepatitis B virus DNA. *Journal of Clinical Laboratory Analysis* **1999**; *13*(6):296-300.
- Iannone, M. A., Taylor, J. D., Chen, J., Li, M.-S., Rivers, P., Slentz-Kesler, K. A., and Weiner, M. P. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. *Cytometry* **2000**; *39*(2):131-140.
- Imamura, M., and Ozawa, E. Differential expression of dystrophin isoforms and utrophin during dibutyl- α -AMP-induced morphological differentiation of rat brain astrocytes. *Proceedings of the National Academy of Sciences of the USA* **1998**; *95*(11):6139-6144.
- Innis, M. A., Gelfand, D. H., Sninsky, J. J., and White, T. J. PCR Protocols. Innis, M. A., Gelfand, D. H., Sninsky, J. J., and White, T. J., editors. Academic Press, San Diego, California, USA **1990**.
- Iqbal, S. S., Chambers, J. P., Brubaker, R. R., Goode, M. T., and Valdes, J. J. Detection of *Yersinia pestis* using branched DNA. *Molecular and Cellular Probes* **1999**; *13*(4):315-320.
- Ivanova, N. B., and Belyavsky, A. V. Identification of differentially expressed by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research* **1995**; *23*(15):2954-2958.
- Ivanova, N. B., and Belyavsky, A. V. Restriction endonuclease-based gene expression fingerprinting. In: Gene cloning and analysis: Current innovations. Schaefer, B. C., editor. Horizon Scientific Press, Wymondham, United Kingdom **1997**:43-60.
- Ivanova, N. B., Luchinskaia, N. N., Popsueva, A. E., Ponomarev, E. D., and Belyavsky, A. V. Identification of mRNA, localized at various segments of the *Xenopus laevis* embryo at early stages of the gastrula. *Proceedings of the Russian Academy of Sciences* **1998**; *359*(1):116-119.
- Ivanova, N. B., Fesenko, I. V., and Beliaevski, A. V. A new method of comparative analysis of gene expression and identification of differentially expressed mRNA. *Molecular Biology (Moscow)* **1999**; *28*(6):1367-1375.
- Iwao, K., Matoba, R., Ueno, N., Ando, A., Miyoshi, Y., Matsubara, K., Noguchi, S., and Kato, K. Molecular classification of primary breast tumors possessing distinct prognostic properties. *Human Molecular Genetic* **2002**; *11*(2):199-206.
- James, B. D., and Higgins, S. J. Nucleic acid hybridization. IRL Press Ltd., Oxford, United Kingdom **1985**.
- Jiang, H., and Fisher, P. B. Use of a sensitive and efficient subtraction hybridization protocol for the identification of genes differentially regulated during the induction of differentiation in human melanoma cells. *Molecular and Cellular Differentiation* **1993**; *1*:285-299.
- Jiang, H., Su, Z. Z., Boyd, J., and Fisher, P. B. Gene expression changes associated with reversible growth suppression and the induction of terminal differentiation in human melanoma cells. *Molecular and Cellular Differentiation* **1993**; *1*(1):41-66.
- Jiang, H., Lin, J., and Fisher, P. B. A molecular definition of terminal cell differentiation in human melanoma cells. *Molecular and Cellular Differentiation* **1994**; *2*(2):221-239.
- Jiang, H., Lin, J., Su, Z. Z., Goldstein, N. I., and Fisher, P. B. Subtraction hybridization identifies a novel melanoma differentiation associated gene, *mda-7*, modulated during human melanoma differentiation, growth and progression. *Oncogene* **1995a**; *11*(12):2477-2486.
- Jiang, H., Lin, J., Su, Z. Z., Kerbel, R. S., Herlyn, M., Weissman, R. B., Welch, D., and Fischer, P. B. The melanoma differentiation associated gene *mda-6*, which encodes the cyclin-dependent kinase inhibitor p21, is differentially expressed during growth, differentiation and progression in human melanoma cells. *Oncogene* **1995b**; *10*(9):1855-1864.

- Jiang, H., Su, Z. Z., Lin, J. J., Goldstein, N. I., and Fisher, P. B. The melanoma differentiation associated gene *mda-7* suppresses cancer cell growth. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(17):9160-9165.
- Jiang, H., Kang, D., Alexandre, D., and Fisher, P. B. RaSH, a rapid subtraction hybridization approach for identifying and cloning differentially expressed genes. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(23):12684-12689.
- Johnson, E. M., Sterner, R., and Allfrey, V. G. Altered nucleosomes of active nucleolar chromatin contain accessible histone H3 in its hyperacetylated forms. *Journal of Biological Chemistry* **1987**; 262(15):6943-6946.
- Jung, M., Kondratyev, A. D., and Dritschilo, A. Elongation factor 1 delta is enhanced following exposure to ionizing radiation. *Cancer Research* **1994**; 54(10):2541-2543.
- Kacharmina, J. E., Crino, P. B., and Eberwine, J. Preparation of cDNA from single cells and subcellular regions. *Methods in Enzymology* **1999**; 303:3-18.
- Kafatos, F. C., Jones, C. W., and Efstratiadis, A. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acid Research* **1979**; 7(6):1541-1552.
- Kang, D. C., LaFrance, R., Su, Z. Z., and Fisher, P. B. Reciprocal subtraction differential RNA display: an efficient and rapid procedure for isolating differentially expressed gene sequences. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(23):13788-13793.
- Kang, D. C., Jiang, H., Wu, Q., Pestka, S., and Fisher, P. B. Cloning and characterization of human ubiquitin-processing protease-43 from terminally differentiated human melanoma cells using a rapid subtraction hybridization protocol RaSH. *Gene* **2001**; 267(2): 233-242.
- Kang, D. C., Gopalkrishnan, R. V., Wu, Q., Jankowsky, E., Pyle, A. M., and Fisher, P. B. *mda-5*: An interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. *Proceedings of the National Academy of Sciences of the USA* **2002**; 99(2): 637-642.
- Kassevetis, G. A., Butler, E. T., Roulland, D., and Chamberlin, M. J. Bacteriophage SP6-specific RNA polymerase. II. Mapping of SP6 DNA and selective *in vitro* transcription. *Journal of Biological Chemistry* **1982**; 257(10):5779-5787.
- Kato, K. Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Research* **1995**; 23(18):3685-3690.
- Kato, K. RNA fingerprinting by molecular indexing. *Nucleic Acids Research* **1996**; 24(2):394-395.
- Kato, K. Adapter-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Research* **1997**; 5(22):4694-4696.
- Kawagoe, H., Humphries, R. K., Blair, A., Sutherland, H. J., and Hogge, D. E. Expression of HOX genes, HOX cofactors, and MLL in phenotypically and functionally defined subpopulations of leukemic and normal human hematopoietic cells. *Leukemia* **1999**; 13(5):687-698.
- Kawai, J., Hirotsune, S., Hirose, K., Fushiki, S., Watanabe, S., and Hayashizaki, Y. Methylation profiles of genomic DNA of mouse developmental brain detected by restriction landmark genomic scanning (RLGS). *Nucleic Acids Research* **1993**; 21(24):5604-5608.
- Kawamoto, S., Ohnishi, T., Kita, H., Chisaka, O., and Okubo, K. Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. *Genome Research* **1999**; 9(12):1305-1312.
- Ke, Y., Rudland, P. S., Jib, C., Smith, P., and Foster, C. S. Systematic differential display: A strategy for a complete assessment of differential gene expression. *Analytical Biochemistry* **1999**; 269(1):201-204.
- Kennedy, H., and Dehay, C. The importance of developmental timing in cortical specification. *Perspectives on Developmental Neurobiology* **1993**; 1(2):93-99.
- Kersten, S., Mandard, S., Tan, N. S., Escher, P., Metzger, D., Chambon, P., Gonzalez, F. J., Desvergne, B., and Wahli, W. Characterization of the fasting-induced adipose factor FIAF, a novel peroxisome proliferator-activated receptor target gene. *Journal of Biological Chemistry* **2000**; 275(37):28488-28493.
- Khandjian, E. W. UV crosslinking of RNA to nylon membrane enhances hybridization signals. *Molecular Biology Reports* **1986**; 11(2):107-115.
- Kim, S. H., Park, H. C., Yeo, S. Y., Hong, S. K., Choi, J. W., Kim, C. H., Weinstein, B. M., and

- Huh, T. L. Characterization of two frizzled8 homologues expressed in the embryonic shield and prechordal plate of zebrafish embryos. *Mechanisms of Development* **1998**; 78(1-2):193-198.
- Kim, S., Zeller, K., Dang, C. V., Sandgren, E. P., and Lee, L. A. A strategy to identify differentially expressed genes using representational difference analysis and cDNA arrays. *Analytical Biochemistry* **2001a**; 288(2):141-148.
- Kim, D. S., Lee, S. J., Park, S. Y., Yoo, H. J., Kim, S. H., Kim, K. J., and Cho, H. J. Differentially expressed genes in rat dorsal root ganglia following peripheral nerve injury. *Neuroreport* **2001b**; 12(15):3401-3405.
- Kimura, B., Kawasaki, S., Fujii, T., Kusunoki, J., Itoh, T., and Flood, S. J. Evaluation of TaqMan PCR assay for detecting *Salmonella* in raw meat and shrimp. *Journal of Food Protection* **1999**; 62(4):329-335.
- Kohne, D. E., Levison, S. A., and Byers, M. J. Room temperature method for increasing the rate of DNA reassociation by many thousandfold: the phenol emulsion reassociation technique. *Biochemistry* **1977**; 16(24):5329-5341.
- Kornmann, B., Preitner, N., Rifat, D., Fleury-Olela, F., and Schibler, U. Analysis of circadian liver gene expression by ADDER, a highly sensitive method for the display of differentially expressed mRNAs. *Nucleic Acids Research* **2001**; 29(11):E51-51.
- Kreuzer, K. A., Bohn, A., Lass, U., Peters, U. R., and Schmidt, C. A. Influence of DNA polymerases on quantitative PCR results using TaqMan probe format in the LightCycler instrument. *Molecular and Cellular Probes* **2000**; 14(2):57-60.
- Kumar, R., and Haugen, J. D. Human and rat osteoblast-like cells express stathmin, a growth-regulatory protein. *Biochemical and Biophysical Research Communications* **1994**; 201(2):861-865.
- Kunkel, L. M., Monaco, A. P., Middlesworth, W., Oches, H. D., and Latt, S. A. Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. *Proceedings of the National Academy of Sciences of the USA* **1985**; 82(14):4778-4782.
- Lackner, K. J., Kilwinski, J., Langmann, T., Aslanidis, C., and Schmitz, G. Multiplex DNA- und RNA-Analyse an fluoreszenten Microbeads als Alternative zum DNA-Array. *Medizinische Genetik* **1999**; 11(1):16-17.
- Lander, E. S., Linton, L. M., Birren, B., Nussbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., Le Vine, R., McEwan, P., McKernan, K., Meldrum, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Peltier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J.,

- Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Horkamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrino, A., Morgan, M. J., and Szustakowski, J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**; 409(6822):860-921.
- Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(24):13057-13062.
- Lavery, D. J., Lopez-Molina, L., Fleury-Olela, F., and Schibler, U. Selective amplification via biotin- and restriction-mediated enrichment (SABRE), a novel selective amplification procedure for detection of differentially expressed mRNAs. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(13):6831-6836.
- Lavery, D. J., Lopez-Molina, L., Margueron, R., Fleury-Olela, F., Conquet, F., Schibler, U., and Bonfils, C. Circadian expression of the steroid 15 α -hydroxylase (Cyp2a4) and coumarin 7-hydroxylase (Cyp2a5) genes in mouse liver is regulated by the PAR leucine zipper transcription factor DBP. *Molecular and Cellular Biology* **1999**; 19(10):6488-6499.
- Lavery, D. J., Fonjallaz, P., Fleury-Olela, F., and Schibler, U. Analysis of differential gene expression using the SABRE enrichment protocol. *Methods in Molecular Biology* **2000**; 99:321-345.
- Lebedeva, I. V., Su, Z. Z., Chang, Y., Kitada, S., Reed, J. C., and Fisher, P. B. The cancer growth suppressing gene *mda-7* induces apoptosis selectively in human melanoma cells. *Oncogene* **2002**; 21(5):708-718.
- Lee, S. W., Tomasetto, C., and Sager, R. Positive selection of candidate tumor-suppressor genes by subtractive hybridization. *Proceedings of the National Academic Sciences of the USA* **1991**; 88(7):2825-2829.
- Lee, L. G., Connell, C. R., and Bloch, W. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Research* **1993**; 21(16):3761-3766.
- Lee, C. K., Klopp, R. G., Weindrich, R., and Prolla, T. A. Gene expression profile of aging and its retardation by caloric restriction. *Science* **1999**; 285(5432):1390-1393.
- Leszczyniecka, M., Roberts, T., Dent, P., Grant, S., and Fisher, P. B. Differentiation therapy of human cancer: basic science and clinical applications. *Pharmacology and Therapeutics* **2001**; 90(2-3):105-156.
- Lewin, B. *Genes*. Oxford University Press, New York, New York, USA **1997**:170.
- Liang, P., and Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **1992**; 257(5072):967-971.
- Liang, P., Averboukh, L., Keyomarsi, K., Sager, R., and Pardee, A. B. Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research* **1992**; 52(24):6966-6968.
- Liang, P., and Pardee, A. B. Recent advances in differential display. *Current Opinion in Immunology* **1995**; 7(2):274-280.
- Liang, P., Bauer, D., Averboukh, L., Warthoe, P., Rohrwild, M., Muller, H., Strauss, M., and Pardee, A. B. Analysis of altered gene expression by differential display. *Methods in Enzymology* **1995**; 254:304-321.
- Lichtlen, P., Wang, Y., Belser, T., Georgiev, O., Certa, U., Sack, R., and Schaffner, W. Target gene search for the metal-responsive transcription factor MTF-1. *Nucleic Acids Research* **2001** 29(7):1514-1523.
- Lim, D. J., Liu, X. L., Sutkowski, D. M., Braun, E. J., Lee, C., and Kozlowski, J. M. Growth of an androgen-sensitive human prostate cancer

- cell line, LNCaP, in nude mice. *Prostate* **1993**; 22(2):109-118.
- Lin, J. J., Jiang, H., and Fisher, P. B. Characterization of a novel melanoma differentiation associated gene-9, *mda-9*, that is down-regulated during terminal cell differentiation. *Molecular and Cellular Differentiation* **1996**; 4(3):317-333.
- Lin, J. J., Jiang, H., and Fisher, P. B. Melanoma differentiation associated gene-9, *mda-9*, is a human gamma interferon responsive gene. *Gene* **1998**; 207(2):105-110.
- Lisi, P. J., Huang, C. W., Hoffman, R. A., and Teipel, J. W. A fluorescence immunoassay for soluble antigens employing flow cytometric detection. *Clinica Chimica Acta* **1982**; 120(2):171-179.
- Lisitsyn, N., and Wigler, M. Cloning the differences between two complex genomes. *Science* **1993**; 259(5097):946-951.
- Lisitsyn, N. A. Representational difference analysis: finding the differences between genomes. *Trends in Genetics* **1995**; 11(8):303-307.
- Lisitsyn, N., and Wigler, M. Representational difference analysis in detection of genetic lesions in cancer. *Methods in Enzymology* **1995**; 254:291-304.
- Lopez-Nieto, C. E., and Nigam, S. K. Selective amplification of protein-coding regions of large sets of genes using statistically designed primer sets. *Nature Biotechnology* **1996**; 14(7):857-861.
- Lukyanov, K., Diatchenko, L., Chenchik, A., Nanisetti, A., Siebert, P., Usman, N., Matz, M., and Lukyanov, S. Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect. *Biochemical and Biophysical Research Communications* **1997**; 230(2):285-288.
- Luo, J. H., Puc, J. A., Slosberg, E. D., Yao, Y., Bruce, J. N., Wright, T. C., Jr., Becich, M. J., and Parsons, R. Differential subtraction chain, a method for identifying differences in genomic DNA and mRNA. *Nucleic Acids Research* **1999**; 27(19):e24.
- Macas, J., Lambert, G. M., Dolezel, D., and Galbraith, D. W. Nuclear expressed sequence tag (NEST) analysis: a novel means to study transcription through amplification of nuclear RNA. *Cytometry* **1998**; 33(4):460-468.
- Madireddi, M. T., Su Z. Z., Young, C. S. H., Goldstein, N. I., and Fisher, P. B. *mda-7*, a novel melanoma differentiation associated gene with promise for cancer gene therapy. *Advances in Experimental Medicine and Biology* **2000**; 465:239-261.
- Mahadeva, H., Starkey, M. P., Sheikh, F. N., Mundy, C. R., and Samani, N. J. A simple and efficient method for the isolation of differentially expressed genes. *Journal of Molecular Biology* **1998**; 284(5):1391-1398.
- Marzluff, W. F., and Huang, R. C. C. Transcription of RNA in isolated nuclei. In: Transcription and translation: A practical approach, Hames, B. D., and Higgins, S. J., editors. IRL Press, Oxford, United Kingdom **1984**:89-130.
- Matoba, R., Kato, K., Kurooka, C., Maruyama, C., Sakakibara, Y., and Matsubara, K. Correlation between gene functions and developmental expression patterns in the mouse cerebellum. *European Journal of Neuroscience* **2000a**; 12(4):1357-1371.
- Matoba, R., Kato, K., Saito, S., Kurooka, C., Maruyama, C., Sakakibara, Y., and Matsubara, K. Gene expression in mouse cerebellum during its development. *Gene* **2000b**; 4(2):125-131.
- Matoba, R., Saito, S., Ueno, N., Maruyama, C., Matsubara, K., and Kato, K. Gene expression profiling of mouse postnatal cerebellar development. *Physiological Genomics* **2000c**; 4(2):155-164.
- Matz, M., Usman, N., Shagin, D., Bogdanova, E., and Lukyanov, S. Ordered differential display: a simple method for systematic comparison of gene expression profiles. *Nucleic Acids Research* **1997**; 25(12):2541-2542.
- Matz, M. V., and Lukyanov, S. A. Different strategies of differential display: areas of application. *Nucleic Acids Research* **1998**; 26(24):5537-5543.
- Matz, M. V., Shagin, D. A., Usman, N. Y., Bogdanova, E. A., Fradkov, A. F., Soboleva, T. A., and Lukyanov, S. A. Cloning of region-specific genetic markers of planarian *Dugesia ti-grina* by means of a new method: differential display. *Bioorganicheskaya Khimiya* **1998**; 24(12):910-915.
- Matz, M. V. Amplification of representative cDNA samples from microscopic amounts of invertebrate tissue to search for new genes. In: Green fluorescent protein: Applications and protocols. Hicks, B. W., editor. Humana Press, Totowa, New Jersey, USA **2002**; in press.
- Mayumi, K., Yaoi, T., Kawai, J., Kojima, S., Watanabe, S., and Suzuki, H. Improved re-

- striction landmark cDNA scanning and its application to global analysis of genes regulated by nerve growth factor in PC12 cells. *Biochimica et Biophysica Acta* **1998**; 1399(1):10-18.
- Mayumi-Matsuda, K., Kojima, S., Nakayama, T., Suzuki, H., and Sakata, T. Scanning gene expression during neuronal cell death evoked by nerve growth factor depletion. *Biochimica et Biophysica Acta* **1999**; 1489(2-3):293-302.
- McClelland, M., and Welsh, J. DNA fingerprinting by arbitrarily primed PCR. *PCR Methods & Applications* **1994**; 4(1):559-565.
- McClelland, M., Ralph, D., Cheng, R., and Welsh, J. Interactions among regulators of RNA abundance characterized using RNA fingerprinting by arbitrarily primed PCR. *Nucleic Acids Research* **1994**; 22(21):4419-4431.
- McClelland, M., Mathieudaude, F., and Welsh, J. RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics* **1995**; 11(6):242-246.
- Melton, D. A., Krieg, P. A., Rcbagliati, M. R., Maniatis, T., Zinn, K., and Green, M. R. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Research* **1984**; 12(18):7035-7056.
- Merschner, S., Hanselmann, R., Welter, C., and Dooley, S. T. Nuclear Runoff Transcription Analysis Using Chemiluminescent Detection. *BioTechniques* **1994**; 16(6):1024-1026.
- Meszaros, M., and Morton, D. B. Subtractive hybridization strategy using paramagnetic Oligo (dT) beads and PCR. *BioTechniques* **1996**; 20(3):413-418.
- Metcalfe, B. L., Sellers, K. W., Jeng, M. J., Huentelman, M. J., Katovich, M. J., and Raizada, M. K. Gene therapy for cardiovascular disorders: is there a future? *Annals of the New York Academy of Sciences* **2001**; 953:31-42.
- Mhashilkar, A. M., Schrock, R. D., Hindi, M., Liao, J., Sieger, K., Kourouma, F., Zou-Yang, X. H., Onishi, E., Takh, O., Vedvick, T. S., Fanger, G., Stewart, L., Watson, G. J., Snary, D., Fisher, P. B., Saeki, T., Roth, J. A., Ramesh, R., and Chada, S. Melanoma differentiation associated gene-7 (*mda-7*): a novel anti-tumor gene for cancer gene therapy. *Molecular Medicine* **2001**; 7(4):271-282.
- Miller, R. D., and Riblet, R. Improved phenol emulsion DNA reassociation technique (PERT) using thermal cycling. *Nucleic Acids Research* **1995**; 23(12):2339-2340.
- Mizuno, Y., Carninci, P., Okazaki, Y., Tateno, M., Kawai, J., Amanuma, H., Muramatsu, M., and Hayashizaki, Y. Increased specificity of reverse transcription priming by trehalose and oligo-blockers allows high-efficiency window separation of mRNA display. *Nucleic Acids Research* **1999**; 27(5):1345-1349.
- Mok, S. C., Wong, K. K., Chan, R. K., Lau, C. C., Tsao, S. W., Knapp, R. C., and Berkowitz, R. S. Molecular cloning of differentially expressed genes in human epithelial ovarian cancer. *Gynecologic Oncology* **1994**; 52(2):247-252.
- Money, T., Reader, S., Qu, L. J., Dunford, R. P., and Moore, G. AFLP-based mRNA fingerprinting. *Nucleic Acids Research* **1996**; 24(13):2616-2617.
- Mou, L., Miller, H., Li, J., Wang, E., and Chalifour, L. Improvements to the differential display method for gene analysis. *Biochemical and Biophysical Research Communications* **1994**; 199(2):564-569.
- Nakayama, T., Yaoi, T., Yasui, M., and Kuwajima, G. N-copine: a novel two C2-domain-containing protein with neuronal activity-regulated expression. *FEBS Letters* **1998**; 428(1-2):80-84.
- Nolte, F. S. Branched DNA signal amplification for direct quantitation of nucleic acid sequences in clinical specimens. *Advances in Clinical Chemistry* **1998**; 33:201-235.
- Norton, D. M., and Batt, C. A. Detection of viable *Listeria monocytogenes* with a 5' nuclease PCR assay. *Applied and Environmental Microbiology* **1999**; 65(5):2122-2127.
- Nussbaum, R. L., Lesko, J. G., Lewis, R. A., Ledbetter, S. A., and Ledbetter, D. H. Isolation of anonymous DNA sequences from within a submicroscopic X chromosome deletion in a patient with choroideremia, deafness, and mental retardation. *Proceedings of the National Academy of Sciences of the USA* **1987**; 84(18):6521-6525.
- Okano, K., Uematsu, C., Matsunaga, H., and Kambara, H. Characteristics of selective polymerase chain reaction (PCR) using two-base anchored primers and improvement of its specificity. *Electrophoresis* **1998**; 19(18):3071-3078.
- Okuzumi, H., Okazaki, Y., and Hayashizaki, Y. RLGS spot mapping method. In: Restriction landmark genomic scanning (RLGS). Hayashizaki, Y., and Watanabe, S., editors.

- Springer-Verlag, Heidelberg, Germany **1997**:57-93.
- Pardee, A. B. G1 events and regulation of cell proliferation. *Science* **1989**; 246(4930):603-607.
- Pardinas, J. R., Combates, N. J., Prouty, S. M., Stenn, K. S., and Parimoo, S. Differential subtraction display: a unified approach for isolation of cDNAs from differentially expressed genes. *Analytical Biochemistry* **1998**; 257(2):161-168.
- Patel, M., and Sive, H. PCR-base subtractive cDNA cloning. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1996**:5.9.1-5.9.20.
- Patrone, G., Puppo F., Cusano R., Scaranari M., Ceccherini I., Puliti A., and Ravazzolo, R. Nuclear run-on assay using biotin labeling, magnetic bead capture and analysis by fluorescent-based RT-PCR. *BioTechniques* **2000**; 29(5):1012-1017.
- Pawlotsky, J. M., Martinot-Peignoux, M., Poveda, J. D., Bastie, A., Le Breton, V., Darthuy, F., Remire, J., Erlinger, S., Dhumeaux, D., and Marcellin, P. Quantification of hepatitis C virus RNA in serum by branched DNA-based signal amplification assays. *Journal of Virological Methods* **1999**; 79(2):227-235.
- Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA* **1988**; 85(8):2444-2448.
- Peferoen, M., Huybrechts, R., and De Loof, A. Vacuum-blotting: A new simple and efficient transfer of proteins from sodium-dodecyl sulfate-polyacrylamide gels to nitrocellulose. *FEBS Letters* **1982**; 145(3):369-372.
- Peliska, J. A., and Benkovic, S. J. Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. *Science* **1992**; 258(5085):1112-1118.
- Pestka, S., Kotenko, S. V., and Fisher P. B. IL-22 and *mda-7* (IL-24), members of the family of IL-10-related cytokines. In: Encyclopedia of hormones. Henry, H. L., and Norman, A. W., editors. Academic Press, Inc., Orlando, Florida, USA **2002**; in press.
- Player, A. N., Shen, L. P., Kenny, D., Antao, V. P., and Kolberg, J. A. Single-copy gene detection using branched DNA (bDNA) *in situ* hybridization. *Journal of Histochemistry and Cytochemistry* **2001**; 49(5):603-612.
- Prashar, Y., and Weissman, S. M. Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(2):659-663.
- Pretlow, T. G., Delmoro, C. M., Dilley, G. G., Spadafora, C. G., and Pretlow, T. P. Transplantation of human prostatic carcinoma into nude mice in Matrigel. *Cancer Research* **1991**; 51(14):3814-3817.
- Prior, C. P., Cantor, C. R., Johnson, E. M., Littau, V. C., and Allfrey, V. G. Reversible changes in nucleosome structure and histone H3 accessibility in transcriptionally active and inactive states of rDNA chromatin. *Cell* **1983**; 34(3):1033-1042.
- Profous-Juchelka, H. R., Reuben, R. C., Marks, P. A., and Rifkind, R. A. Transcriptional and post-transcriptional regulation of globin gene accumulation in murine erythroleukemia cells. *Molecular Cell Biology* **1983**; 3(2):229-232.
- Qin, L., Overmars, H., Helder, J., Popeijus, H., van der Voort, J. R., Groenink, W., van Koert, P., Schots, A., Bakker, J., and Smant, G. An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode *Globodera rostochiensis*. *Molecular Plant Microbe Interactions* **2000**; 13(8):830-836.
- Ralph, D., McClelland, M., and Welsh, J. RNA fingerprinting using arbitrarily primed PCR identifies differentially regulated RNAs in mink lung (Mv1Lu) cells growth arrested by transforming growth factor beta 1. *Proceedings of the National Academy of Sciences of the USA* **1993**; 90(22):10710-10714.
- Rangnekar, V. V., Waheed, S., and Rangnekar, V. M. Interleukin-1-inducible tumor growth arrest is characterized by activation of cell type-specific "early" gene expression programs. *Journal of Biological Chemistry* **1992**; 267(9):6240-6248.
- Reddy, P. G., Su, Z. Z., and Fisher, P. B. Identification and cloning of genes involved in progression of transformed phenotype. In: Chromosome and genetic analysis: Methods in molecular genetics. Adolph, K. W., editor. Academic Press, Inc., Orlando, Florida, USA **1993**:68-102.
- Ririe, K. M., Rasmussen, R. P., and Wittwer, C. T. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Analytical Biochemistry* **1997**; 245(2):154-160.

- Roeder, T. Solid-phase cDNA synthesis, a versatile approach. *Nucleic Acids Research* **1998**; 26(14):3451-3452.
- Roeder, T. Simple and efficient cloning of small polymerase chain reaction-generated DNA products. *Analytical Biochemistry* **2000**; 285(2):278-280.
- Rolfe, F. G., and Sewell, W. A. Analysis of human interleukin-5 gene transcription by a novel nuclear run-on method based on the polymerase chain reaction. *Journal of Immunological Methods* **1997**; 202(2):143-151.
- Roth, W. K., Lee, J. H., Ruster, B., and Zeuzem, S. Quantification of hepatitis C virus RNA by RT-PCR in comparison to the branched DNA method. *Zeitschrift für Gastroenterologie* **1998**; 36(1):5-11.
- Rouet, F., Montcho, C., Rouzioux, C., Leroy, V., Msellati, P., Kottan, J. B., You, B., Viho, I., Dabis, F. The Abidjan DITRAME Study Group (ANRS 049 clinical trial). Early diagnosis of paediatric HIV-1 infection among African breast-fed children using a quantitative plasma HIV RNA assay. *Aids* **2001**; 15(14):1849-1856.
- Rubenstein, J. L., Brice, A. E., Ciaranello, R. D., Denney, D., Porteus, M. H., and Usdin, T. B. Subtractive hybridization system using single-stranded phagemids with directional inserts. *Nucleic Acids Research* **1990**; 18(16):4833-4842.
- Ryan, M., Starkey, M., Faull, R., Emson, P., and Bahn, S. Indexing-based differential display-studies on post-mortem Alzheimer's brains. *Brain Research: Molecular Brain Research* **2001**; 88(1-2):199-202.
- Ryffel, G. U., and McCarthy, B. J. Complexity of cytoplasmic RNA in different mouse tissues measured by hybridization of polyadenylated RNA to complementary DNA. *Biochemistry* **1975**; 14(7):1379-1385.
- Rzucidlo, S. J., Gibbons, J., and Stice, S. L. Comparison by restriction fragment differential display RT-PCR of gene expression pattern in bovine oocytes matured in the presence or absence of fetal calf serum. *Molecular Reproduction and Development* **2001**; 59(1):90-96.
- Saeki, T., Mhashilkar, A., Chada, S., Branch, C., Roth, J. A., and Ramesh, R. Tumor-suppressive effects by adenovirus-mediated *mda-7* gene transfer in non-small cell lung cancer cell *in vitro*. *Gene Therapy* **2000**; 7(23):2051-2057.
- Sagerström, C. G., Sun, B. I., and Sive, H. L. Subtractive cloning: past, present, and future. *Annual Review in Biochemistry* **1997**; 66:751-783.
- Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B., and Erlich, H. A. Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature* **1986**; 324(6093):163-166.
- Saito, S., Matoba, R., Ueno, N., Matsubara, K., and Kato, K. Comparison of gene expression profiling during postnatal development of mouse dentate gyrus in comparison with that of mouse cerebellum. *Physiological Genomics* **2002**; 8(2):131-137.
- Sakai, K., Higuchi, H., Matsubara, K., and Kato, K. Microarray hybridization with fractionated cDNA: enhanced identification of differentially expressed genes. *Analytical Biochemistry* **2000**; 287(1):32-37.
- Sakamoto, W., Tan, S. H., Murata, M., and Motoyoshi, F. An unusual mitochondrial *atp9-rp116* cotranscript found in the maternal distorted leaf mutant of *Arabidopsis thaliana*: implication of GUG as an initiation codon in plant mitochondria. *Plant and Cell Physiology* **1997**; 38(8):975-979.
- Sambrook, J., Fritsch, E., and Maniatis, T. Molecular cloning. A laboratory manual. 2nd edition. Cold Spring Harbor Laboratory Press, New York, New York, USA **1989**.
- Santiard-Baron, D., Gosset, P., Nicole, A., Sinet, P. M., Christen, Y., and Ceballos-Picot, I. Identification of beta-amyloid-responsive genes by RNA differential display: early induction of a DNA damage-inducible gene, *gadd45*. *Experimental Neurology* **1999**; 158(1):206-213.
- Sanyal, A., O'Driscoll, S. W., Bolander, M. E., and Sarkar, G. An effective method of completely removing contaminating genomic DNA from an RNA sample to be used for PCR. *Molecular Biotechnology* **1997**; 8(2):135-137.
- Sargent, T. D., and Dawid, I. B. Differential gene expression in the gastrula of *Xenopus laevis*. *Science* **1983**; 222(4620):135-139.
- Sarkar, G., Koeberl, D. D., and Sommer, S. S. Direct sequencing of the activation peptide and the catalytic domain of the factor IX gene in six species. *Genomics* **1990**; 6(1):133-143.
- Sarkar, G., Turner, R. T., and Bolander, M. E. A direct method of unknown sequence retrieval adjacent to a known locus by using universal

- primers. *PCR Methods and Applications* **1993**; 2(4):318-322.
- Sarvetnick, N., Tsai, J.-Y., Rox, H., Pilder, S. H., and Silver, L. M. A mouse chromosome 17 gene encodes a testes-specific transcript with unusual properties. *Immunogenetics* **1989**; 30(1):34-41.
- Sauvageau, G., Lansdorp, P. M., Eaves, C. J., Hogge, D. E., Dragowska, W. H., Reid, D. S., Largman, C., Lawrence, H. J., and Humphries, R. K. Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells. *Proceedings of the National Academy of Sciences of the USA* **1994**; 91(25):12223-12227.
- Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Analytical Biochemistry* **1995**; 227(2):255-273.
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**; 270(5235):467-470.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davies, R. W. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(20):10614-10619.
- Schmidt, E. E., and Schibler, U. Cell size regulation, a mechanism that controls cellular RNA accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and the liver-enriched transcription factor DBP. *Journal of Cell Biology* **1995**; 128(4):467-483.
- Schmittgen, T. D., and Zakrajsek, B. A. Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. *Journal of Biochemical and Biophysical Methods* **2000**; 46(1-2):69-81.
- Schramm, G., Bruchhaus, I., and Roeder, T. A simple and reliable 5'-RACE approach. *Nucleic Acids Research* **2000**; 28(22):e96.
- Shaw-Smith, C. J., Coffey, A. J., Huckle, E., Durham, J., Campbell, E. A., Freeman, T. C., Walters, J. R., and Bentley, D. R. Improved method for detecting differentially expressed genes using cDNA indexing. *Biotechniques* **2000**; 28(5):958-964.
- Sheils, O. M., and Sweeney, E. C. TSH receptor status of thyroid neoplasms-TaqMan RT-PCR analysis of archival material. *Journal of Pathology* **1999**; 188(1):87-92.
- Shi, Y. B., and Brown D. D. The earliest changes in gene expression in tadpole intestine induced by thyroid hormone. *Journal of Biological Chemistry* **1993**; 268(27):20312-20317.
- Shi, M. M., Bleavins, M. R., and de la Iglesia, F. A. Technologies for detecting genetic polymorphisms in pharmacogenomics. *Molecular Diagnostics* **1999**; 4(4):343-351.
- Shibata, H., and Plass, C. Application of RLGS to screening endogenously imprinted genes. In: Restriction landmark genomic scanning (RLGS). Hayashizaki, Y., and Watanabe, S., editors. Springer-Verlag, Heidelberg, Germany **1997**:95-114.
- Shmelkov, S. V., Visser, J. W. M., and Belyavsky, A. V. Two-dimensional gene expression fingerprinting. *Analytical Biochemistry* **2001**; 290(1):26-35.
- Sibson, D. R. Process for categorising nucleotide sequence populations. *WO9401582* **1992**.
- Sibson, D. R., and Starkey, M. P. Increasing the average abundance of low abundance cDNAs by ordered subdivision of cDNA populations. In: Methods in molecular biology. Cowell, I. G., and Austin, C. A., editors. Humana Press, Inc., Totowa, New Jersey, USA **1997**; 69(cDNA library protocols).
- Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A., and Lukyanov, S. A. An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Research* **1995**; 23(6):1087-1088.
- Simm, M., Su, Z., Huang, E. Y., Chen, Y., Jiang, H., Volsky, D. J., and Fisher, P. B. Cloning of differentially expressed genes in an HIV-1 resistant T cell clone by rapid subtraction hybridization, RaSH. *Gene* **2001**; 269(1-2):93-101.
- Sive, H. L., and St John, T. A simple subtractive hybridization technique employing photoactivatable biotin and phenol extraction. *Nucleic Acids Research* **1988**; 16(22):10937.
- Smith, M. R., Devine, C. S., Cohn, S. M., and Lieberman, M. W. Quantitative electrophoretic transfer of DNA from polyacrylamide or agarose gels to nitrocellulose. *Analytical Biochemistry* **1984**; 137(1):120-124.
- Smith, D. R. Ligation-mediated PCR of restriction fragments from large DNA molecules. *PCR Methods and Applications* **1992**; 2(1):21-27.

- Smith, P. L., Walker-Peach, C. R., Fulton, J. R., and DuBois, D. B. A rapid, sensitive, multiplexed assay for detection of viral nucleic acids using the FlowMetrix system. *Clinical Chemistry* **1998**; 44(9):2054-2056.
- Smith, C. W., and Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemical Science* **2000**; 25(8):381-388.
- Srivastava, R. A. K., and Schonfeld, G. Measurements of rate of transcription in isolated nuclei by nuclear "run-off" assay. In: *Methods in molecular biology*. Harwood, A. J., editor. Humana Press, Totowa, New Jersey, USA **1994**; (*Protocols for gene analysis*):281-288.
- Sterner, R., Boffa, L. C., Chen, T. A., and Allfrey, V. G. Cell cycle-dependent changes in conformation and composition of nucleosomes containing human histone gene sequences. *Nucleic Acids Research* **1987**; 15(11):4375-4391.
- Stone, B., and Wharton, W. Targeted RNA fingerprinting: the cloning of differentially-expressed cDNA fragments enriched for members of the zinc finger gene family. *Nucleic Acids Research* **1994**; 22(13):2612-2618.
- Straus, D., and Ausubel, F. M. Genomic subtraction for cloning DNA corresponding to deletion mutations. *Proceedings of the National Academy of Sciences of the USA* **1990**; 87(5):1889-1893.
- Strauss, K. I., and Jacobowitz, D. M. Quantitative measurement of calretinin and β -actin mRNA in rat brain micropunches without prior isolation of RNA. *Molecular Brain Research* **1993**; 20(3):229-239.
- Strickland, S., and Mahdavi, V. The induction of differentiation in teratocarcinoma stem cells by retinoic acid. *Cell* **1978**; 15(2):393-403.
- Su, Z.-Z., Shi, Y., and Fisher, P. B. Subtraction hybridization identifies a transformation progression-associated gene PEG-3 with sequence homology to a growth arrest and DNA damage-inducible gene. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(17):9125-9130.
- Su, Z.-Z., Madireddi, M. T., Lin, J. J., Young, C. S. H., Kitada, S., Reed, J. C., Goldstein, N. I., and Fisher, P. B. The cancer growth suppressor gene *mda-7* selectively induces apoptosis in human breast cancer cells and inhibits tumor growth in nude mice. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(24):12200-14405.
- Su, Z. Z., Lebedeva, I. V., Gopalkrishnan, R. V., Goldstein, N. I., Stein, C. A., Reed, J. C., Dent, P., and Fisher, P. B. A combinatorial approach for selectively inducing programmed cell death in human pancreatic cancer cells. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(18):10332-10337.
- Su, Z.-Z., Kang, D.-C., Chen, Y., Pekarskaya, O., Chao, W., Volsky, D. J., and Fisher, P. B. Identification and cloning of human astrocyte genes displaying elevated expression after infection with HIV-1 or exposure to HIV-1 envelope glycoprotein by rapid subtraction hybridization, RaSH. *Oncogene* **2002**; 21(22):3592-3602.
- Suzuki, H., Kawai, J., Taga, C., Ozawa, N., and Watanabe, S. A PCR-mediated method for cloning spot DNA on restriction landmark genomic scanning (RLGS) gel. *DNA Research* **1994**; 1(5):175-180.
- Suzuki, H., Yaoi, T., Kawai, J., Hara, A., Kuwajima, G., and Watanabe, S. Restriction landmark cDNA scanning (RLCS): a novel cDNA display system two-dimensional gel electrophoresis. *Nucleic Acid Research* **1996**; 24(2):289-294.
- Suzuki, R., Shintani, T., Sakuta, H., Kato, A., Ohkawara, T., Osumi, N., and Noda, M. Identification of RALDH-3, a novel retinaldehyde dehydrogenase, expressed in the ventral region of the retina. *Mechanisms of Development* **2000**; 98(1-2):37-50.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. *Nature Genetics* **1999**; 22(3):281-285.
- Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A., and Tunncliffe, A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **1992**; 13(3):718-725.
- Theilgaard-Mönch, K., Cowland, J., and Borregaard, N. Profiling of gene expression in individual hematopoietic cells by global mRNA amplification and slot blot analysis. *Journal of Immunological Methods* **2001**; 252(1-2):175-189.
- Theissen, G., and Fischer, A. RC4D – restriction fragment length polymorphism-coupled domain-directed differential display. *Methods in Molecular Biology* **1997**; 85:123-133.

- Thelwell, N., Millington, S., Solinas, A., Booth, J., and Brown, T. Mode of action and application of Scorpion primers to mutation detection. *Nucleic Acids Research* **2000**; *28*(19):3752-3761.
- Thomas, P. S. Hybridization of denatured RNA and small DNA fragments transferred to nitrocellulose. *Proceedings of the National Academy of Sciences of the USA* **1980**; *77*(9):5201-5205.
- Tortola, S., Capella, G., Marcuello, E., Gunther, K., Aiza, G., Masramon, L., Reymond, M. A., and Peinado, M. A. Analysis of differential gene expression in human colorectal tumor tissues by RNA arbitrarily primed-PCR: a technical assessment. *Laboratory Investigation* **1998**; *78*(3):309-317.
- Tyagi, S., and Kramer, F. R. Molecular beacons: probes that fluoresce upon hybridization. *Nature Biotechnology* **1996**; *14*(3):303-308.
- Uematsu, C., Nishida, J., Okano, K., Miura, F., Ito, T., Sakaki, Y., and Kambara, H. Multiplex polymerase chain reaction (PCR) with color-tagged module-shuffling primers for comparing gene expression levels in various cells. *Nucleic Acids Research* **2001**; *29*(16):e84.
- Unrau, P., and Deugau, K. V. Non-cloning amplification of specific DNA fragments from whole genomic DNA digests using DNA 'indexers'. *Gene* **1994**; *145*(2):163-169.
- Urdea, M., Running, J. A., Horn, T., Clyne, J., Ku, L., and Warner, B. A novel method for the rapid detection of specific nucleotide sequences in crude biological samples without blotting or radioactivity; application to the analysis of hepatitis B virus in human serum. *Gene* **1987**; *61*(3):253-264.
- Urdea, M. S., Kolberg, J., Clyne, J., Running, J. A., Besemer, D., Warner, B., and Sanchez-Pescador, R. Application of a rapid non-radioisotopic nucleic acid analysis system to the detection of sexually transmitted disease-causing organisms and their associated antimicrobial resistances. *Clinical Chemistry* **1989**; *35*(8):1571-1575.
- Utans, U., Liang, P., Wyner, L. R., Karnovsky, M. J., and Russell, M. E. Chronic cardiac rejection: identification of five upregulated genes in transplanted hearts by differential mRNA display. *Proceedings of the National Academy of Sciences of the USA* **1994**; *91*(14):6463-6467.
- Van Cleve, M., Ostrerova, N., Tietgen, K., Cao, W., Chang, C., Collins, M. L., Kolberg, J., Urdea, M., and Lohman, K. Direct quantitation of HIV by flow cytometry using branched DNA signal amplification. *Molecular and Cellular Probes* **1998**; *12*(4):243-247.
- Van Hoof, A., and Green, P. J. Control of mRNA decay in plants. In: mRNA metabolism and post-transcriptional gene regulation. Harford, J. B., and Morris, D. R., editors. Wiley-Liss, New York, New York, USA **1997**:201-216.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **1995**; *270*(5235):484-487.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland,

- T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Murganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yoosheph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu X. The sequence of the human genome. *Science* **2001**; 291(5507):1304-1351.
- Vignali, D. A. A. Multiplexed particle-based flow cytometric assays. *Journal of Immunological Methods* **2000**; 243(1-2):243-255.
- Vila, M. R., Nicolas, A., Morote, J., de Torres, I., and Meseguer, A. Increased glyceraldehyde-3-phosphate dehydrogenase expression in renal cell carcinoma identified by RNA-based, arbitrarily primed polymerase chain reaction. *Cancer* **2000**; 89(1):152-164.
- Volkenandt, M., Dicker, A. P., Banerjee, D., Fanin, R., Schweitzer, B., Horikoshi, T., Danenberg, K., Danenberg, P., and Bertino, J. R. Quantitation of gene copy number and mRNA using the polymerase chain reaction. *Proceedings of the Society for Experimental Biology and Medicine* **1992**; 200(1):1-6.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Pelman, J., Kuiper, M., and Zabeau, M. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **1995**; 23(21):4407-4414.
- Wrang, M. L., Moller, F., Alsbo, C. W., and Diemer, N. H. Changes in gene expression following induction of ischemic tolerance in rat brain: detection and verification. *Journal of Neuroscience Research* **2001**; 65(1):54-58.
- Walker, J., Chen, T. A., Sterner, R., Berger, M., Winston, F., and Allfrey, V. G. Affinity chromatography of mammalian and yeast nucleosomes. Two modes of binding of transcriptionally active mammalian nucleosomes to organomercurial-agarose columns, and contrasting behavior of the active nucleosomes of yeast. *Journal of Biological Chemistry* **1990**; 265(10):5736-5746.
- Wan, J. S., Sharp, S. J., Poirier, G. M.-C., Wagaman, P. C., Chambers, J., Pyati, J., Hom, Y.-L., Galindo, J. E., Huvar, A., Peterson, P. A., Jackson, M. R., and Erlander, M. G. Cloning differentially expressed mRNAs. *Nature Biotechnology* **1996**; 14(13):1685-1691.
- Wan, J. S., and Erlander, M. G. Cloning differentially expressed genes by using differential display and subtractive hybridization. *Methods in Molecular Biology* **1997**; 85:45-68.
- Wang, Z., and Brown, D. D. A gene expression screen. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(24):11505-11509.
- Wang, Z., Tufts, R., Haleem, R., and Cai, X. Genes regulated by androgen in the rat ventral prostate. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(24):12999-13004.
- Wang, S. M., Fears, S. C., Zhang, L., Chen, J.-J., and Rowley, J. D. Screening poly(dA/dT) cDNAs for gene identification. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(8):4162-4167.
- Watson, J. B., and Margulies, J. E. Differential cDNA screening strategies to identify novel stage-specific proteins in the developing mammalian brain. *Developmental Neuroscience* **1993**; 15(2):77-86.
- Watson, M. A., and Fleming, T. P. Isolation of differentially expressed sequence tags from human breast cancer. *Cancer Research* **1994**; 54(17):4598-4602.

- Waxman, S. Differentiation therapy. Challenges of modern medicine. Waxman, editor. Ares-Serono Symposia Publishers, Rome, Italy **1996**.
- Weaver, R. F., and Weissman, C. Mapping of RNA by modification of the Berk-Sharp procedure: The 5' termini of 15S β -globin mRNA and mature 10S β -globin mRNA have identical map coordinates. *Nucleic Acids Research* **1979**; 7(5):1175-1193.
- Weber, K., Bolander, M. E., and Sarkar, G. Rapid acquisition of unknown DNA sequence adjacent to known segment by multiplex restriction site PCR. *BioTechniques* **1998a**; 25(3):415-419.
- Weber, K., Bolander, M. E., and Sarkar, G. Selective differential fingerprinting: A method for identifying differentially expressed genes in a family between two samples. *Molecular Biotechnology* **1998b**; 10(1):77-81.
- Weinstein, J. N., and Boulamwini, J. Restriction display (RD-PCR) of differentially expressed mRNAs. *WO9129211* **1997**.
- Weis, L. Common malignant bone tumors: osteosarcoma. In: Surgery for bone and soft-tissue tumors. Simon, M. A., and Springfield, D., editors. Lippincott-Raven, New York, New York, USA **1998**:265-274.
- Welsh, J., and McClelland, M. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* **1990**; 19(19):7213-7218.
- Welsh, J., Chada, K., Dalal, S. S., Ralph, D., Cheng, R., and McClelland, M. Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research* **1992**; 20(19):4965-4970.
- Westbrock-Wadman, S., Sherman, D. R., Hickey, M. J., Coulter, S. N., Zhu, Y. O., Warren, P., Nguyen, L. Y., Shawar, R. M., Folger, K. R., and Stover, C. K. Characterization of a *Pseudomonas aeruginosa* efflux pump contributing to aminoglycoside impermeability. *Antimicrobial Agents and Chemotherapy* **1999**; 43(12):2975-2983.
- Whitcombe, D., Theaker, J., Guy, S. P., Brown, T., and Little, S. Detection of PCR products using self-probing amplicons and fluorescence. *Nature Biotechnology* **1999**; 17(8):804-807.
- Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **1990**; 18(22):6531-6535.
- Winkles, J. A. Serum- and polypeptide growth factor-inducible gene expression in mouse fibroblasts. *Progress in Nucleic Acid Research and Molecular Biology* **1998**; 58:41-78.
- Wittwer, C. T., Herrmann, M. G., Moss, A. A., and Rasmussen, R. P. Continuous fluorescence monitoring of rapid cycle DNA amplification. *Biotechniques* **1997**; 22(1):130-138.
- Won, D. I., Park, J. Y., Kim, J. M., and Kim, H. S. Comparison of the reverse transcription-PCR with the branched DNA assay for measurement of human immunodeficiency virus type 1 RNA levels in plasma of Korean patients. *Yonsei Medical Journal* **2001**; 42(2):204-208.
- Wong, K. K., and McClelland, M. Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences of the USA* **1994**; 91(2):639-643.
- Wu, C. G., Hakvoort, T. B., Lamers, W. H., and Chamuleau, R. A. Isolation of up- and down-regulated cDNAs associated with hepatocellular carcinoma by a subtraction-enhanced display technique. *Biochimica et Biophysica Acta* **1996**; 1315(3):169-175.
- Wytenbach, A., Swartz, J., Kita, H., Thykjaer, T., Carmichael, J., Bradley, J., Brown, R., Maxwell, M., Schapira, A., Orntoft, T. F., Kato, K., and Rubinsztein, D. C. Polyglutamine expansions cause decreased CRE mediated transcription and early gene expression changes prior to cell death in an inducible cell model of Huntington's disease. *Human Molecular Genetics* **2001**; 10(17):1829-1845.
- Yamanishi, Y., Boyle, D. L., Clark, M., Maki, R. A., Tortorella, M. D., Arner, E. C., and Firestein, G. S. Expression and regulation of aggrecanase in arthritis: the role of TGF- β . *Journal of Immunology* **2002**; 168(3):1405-1412.
- Yamashita, T., Hashimoto, S., Kaneko, S., Nagai, S., Toyoda, N., Suzuki, T., Kobayashi, K., and Matsushima, K. Comprehensive gene expression profile of a normal human liver. *Biochemical and Biophysical Research Communications* **2000**; 269(1):110-116.
- Yamashita, R., Matsubara, K., and Kato, K. A comprehensive collection of mouse zinc finger motifs compiled by molecular indexing. *Gene* **2001**; 274(1-2):101-110.
- Yang, M., and Sytkowski, A. J. Cloning differentially expressed genes by linker capture

- subtraction. *Analytical Biochemistry* **1996**; 237(1):109-114.
- Yang, M., and Sytkowski, A. J. Linker-capture subtraction. In: *Gene cloning and analysis by RT-PCR*. Siebert, P. D., and Larrick, J., editors. BioTechniques Books, Eaton Publishing, Natick, Massachusetts, USA **1998**.
- Yang, M., Loda, M., and Sytkowski, A. J. Identification of genes expressed differentially by LNCaP or PC-3 prostate cancer cell lines. *Cancer Research* **1998**; 58(16):3732-3735.
- Yang, D., Kusser, I., Kopke, A. K., Koof, B. F., and Matheson, A. T. The structure and evolution of the ribosomal proteins encoded in the *spc* operon of the archaeon (*Crenarchaeota*) *Sulfolobus acidocaldarius*. *Molecular Phylogenetics and Evolution* **1999**; 12(2):177-185.
- Yang, L., Tran, D. K., and Wang, X. BADGE, BeadsArray for the detection of gene expression, a high-throughput diagnostic bioassay. *Genome Research* **2001**; 11(11):1888-1898.
- Yaoi, T., Suzuki, H., Kawai, J., and Watanabe, S. RLCS, restriction landmark cDNA scanning. In: *Restriction landmark genomic scanning (RLGS)*. Hayashizaki, H., and Watanabe, S., editors. Springer-Verlag, Heidelberg, Germany **1997**:129-156.
- Ye, F., Li, M.-S., Taylor, J. D., Nguyen, Q., Colton, H. M., Casey, W. M., Wagner, M., Weiner, M. P., and Chen, J. Fluorescent microsphere-based readout technology for multiplexed human single nucleotide polymorphism analysis and bacterial identification. *Human Mutation* **2001**; 17(4):305-316.
- Ying, S. Y., Lui, H. M., Lin, S. L., and Chuong, C. M. Generation of full-length cDNA library from single human prostate cancer cells. *BioTechniques* **1999**; 27(3):410-414.
- Yu, Y. P., Lin, F., Bisceglia, M., Krill, D., Dhir, R., Becich, M., and Luo, J.-H. Identification of a novel gene with increasing rate of suppression in high grade prostate cancers. *American Journal of Pathology* **2001**; 158(1):19-24.
- Yuh, I., Yaoi, T., Watanabe, S., Okajima, S., Hirasawa, Y., and Fushiki, S. Up-regulated uridine kinase gene identified by RLCS in the ventral horn after crush injury to rat sciatic nerves. *Biochemical and Biophysical Research Communications* **1999**; 266(1):104-109.
- Zeng, J., Gorski, R. A., and Hamer, D. Differential cDNA cloning by enzymatic degrading subtraction (EDS). *Nucleic Acids Research* **1994**; 22(21):4381-4385.
- Zhang, L., and Medina, D. Gene expression screening for specific genes associated with mouse mammary tumor development. *Molecular Carcinogenesis* **1993**; 8(2):123-126.
- Zhang, H., Zhang, R., and Liang, P. Differential screening of differential display cDNA products by reverse northern. *Methods in Molecular Biology* **1997a**; 85:87-93.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. *Science* **1997b**; 276(5316):1268-1272.
- Zhang, J. S., Duncan, E. L., Chang, A. C., and Reddel, R. R. Differential display of mRNA. *Molecular Biotechnology* **1998**; 10(2):155-165.
- Zhao, S., Ooi, S. L., and Pardee, A. B. New primer strategy improves precision of differential display. *BioTechniques* **1995**; 18(5):842-850.
- Zhao, S., Ooi, S. L., Yang, F. C., and Pardee, A. B. Three methods for identification of true positive cloned cDNA fragment in differential display. *BioTechniques* **1996**; 20(3):400-404.
- Zijlmans, J. M., Visser, J. W., Kleiverda, K., Kluin, P. M., Willemze, R., and Fibbe, W. E. Modification of rhodamine staining allows identification of hematopoietic stem cells with preferential short-term or long-term bone marrow-repopulating ability. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(19):8901-8905.
- Zinn, K., DiMaio, D., and Maniatis, T. Identification of two distinct regulatory regions adjacent to the human β -interferon gene. *Cell* **1983**; 34(3):865-879.

4

High-throughput and industrial methods for mRNA expression analysis

4.1

Introduction

The volume of information available on genetic regulation means that it has now become feasible to examine the expression of a large number of genes in parallel at the RNA level. Such expression studies require the use of techniques that, in contrast to methods presented in the previous chapter, allow the simultaneous investigation of many genes within a short period of time and that are suitable, at least in principle, for scaling-up. The present chapter is devoted to such high-throughput methods for mRNA expression analysis.

We have divided the chapter into three sections. The first deals with techniques that exploit the ability of DNA to bind in specific manner to itself; the second deals with techniques based on the polymerase chain reaction (PCR) that exploit the possibility of using short or arbitrary oligonucleotides to simultaneously amplify a large number of cDNA fragments; and the third deals with techniques based on DNA sequencing which generally operate by the incorporation of short sequence tags into the gene whose expression is to be quantified.

Each of the methods described here has its particular niche and each has its specific drawbacks. While analysis of gene expression using DNA microarrays is undoubtedly a powerful and appealing technique, it suffers from the inherent drawback that only known sequences can be presented on the microarray. In addition, this technique is expensive and has been beset by problems related to quality control and lack of sensitivity. Differential display RT-PCR, for example, is at first sight an inexpensive method but is technically unforgiving and labour-intensive. However, differential display RT-PCR has the great advantage of revealing changes in the level of expression of unknown genes. The methods based on gene sequencing such as massively parallel signature sequencing (MPSS[®]) and serial analysis of gene expression (SAGE) also reveal the regulation of unknown genes but are expensive and require a level of infrastructure that effectively limits their use to the non-academic setting.

We hope that the contributions which follow will assist the reader in deciding which technique is best for his problem and best for his pocket.

4.2

Hybridisation-based techniques

4.2.1

DNA microarrays

4.2.1.1 Introduction

Many molecular biological and biochemical methods can already be implemented using microtechnology. Biochips less than the size of a postage stamp, for example, allow simultaneous measurement of the differential expression of several thousand genes. Biochips are suitable not only for the analysis of gene expression and for sequencing of nucleic acids but may also be used for sample preparation and analysis. At present, much effort is directed towards the development of fully integrated and automated 'lab-on-a-chip' systems. In the not too distant future, such systems will be available for routine genetic diagnostics and for many areas of genetic research where their use will lead to more accurate analyses and reduced costs. Following on the heels of microtechnology, one of the most important current areas of applied research today is the field of nanotechnology, the development of instruments and processes at the nanometer (10^{-9} metres) scale. Extreme miniaturisation of this degree offers many advantages including reduced productions costs, a greater degree of automation and increased flexibility of use. The field of biochip and, in particular, nucleic acid array design is eminently suited for a nanotechnology approach. This chapter aims therefore to outline the main processes currently used for the manufacture of nucleic acid microarrays, or 'DNA chips' as the high-density variety are often called (see below) and to briefly point to the direction that this may take in the future.

The complete coding information of the human genome is estimated to comprise somewhere between 30,000 and 40,000 genes, the total sequence of which probably makes up about ten percent of the total number of 3.2 billion nucleotides (Lander *et al.*, 2001; Venter *et al.*, 2001; Zhuo *et al.*, 2001). Nevertheless, at the time of writing, only about one tenth of all known genes have been matched functionally to a corresponding enzyme or protein. For most of the nucleic acid sequences that have been mapped to chromosomes, the expression information available in public databases is very limited. Also, for most expressed sequence tags (ESTs) that encode protein, the function of the encoded protein is unknown. A current focus of medical and biochemical research is therefore the generation of this information. Elucidating the cellular metabolism regulated by these genes will of course be a much greater and more challenging, but ultimately more interesting and rewarding task that is likely to take many decades. Both of these processes will be greatly accelerated by the application of DNA microarrays.

4.2.1.2 How DNA microarrays work

4.2.1.2.1 Function and structure of DNA microarrays

DNA microarrays function using the classical hybridisation techniques of molecular biology as were first developed by Edward Southern and are used in conventional Southern and Northern blots (Alwine *et al.*, 1977; Brown, 1993; Brown & Mackey, 1997; Dyson, 1991; Southern, 1975). The key to the method is the property of nucleic acid strands to bind specifically and strongly to each other using Watson-Crick base pairing (cytosine to guanine and thymine to adenine). The specificity of this binding is such that mismatches at a single position in a DNA strand of, say, 20 bases, under suitable conditions sufficiently reduce the binding affinity as to completely prevent hybridisation.

For the production of DNA micro arrays, nucleic acids with known sequences are immobilised on a suitable substrate in a known pattern. The nucleic acid under investigation is labelled and hybridised with the nucleic acid on the array. Under ideal conditions, Watson-Crick base pairing only occurs between nucleic acid strands that are exactly complementary. Moreover, under ideal circumstances, the intensity of the detected signal is directly proportional to the amount of the hybridised probe, so that the amount of a nucleic acid species present in a sample can be measured.

4.2.1.2.2 Low-density microarrays

Such low-density arrays are commercially available from a wide range of suppliers or may be produced by research groups in their own laboratories. In one application, nucleic acids are coupled with high affinity to a nylon membrane or to chemically modified glass slides. If necessary, the nucleic acids can be covalently cross-linked to the nylon membrane using ultraviolet light. Plasmids and products of polymerase chain reactions are usually applied ('spotted') to a membrane or slides with a density of several square centimetres at a density of 80 to 200 spots per square centimetre for low-density microarrays and 500 to 10,000 spots per square centimetre for medium-density microarrays (Schena & Davis, 1999). The method of detection and labelling depends on the material on which the nucleic acid probes are spotted. Radioactively labelled samples, for example, are usually used for nylon membrane-based arrays and evaluated by auto-radiography on X-ray films or by photo-stimulated storage phosphor imaging plates, while fluorescently-labelled samples, are usually used for glass slide-based arrays and analysed using fluorimetric readers. Commercially available arrays used a wide variety of detection techniques as described in detail in a following contribution of this book.

4.2.1.2.3 High-density microarrays

Arrays with more than 10,000 probes per square centimetre also exist. Today, it is possible to produce DNA microarrays with more than 250,000 probes per square centimetre (Schena & Davies, 1999). Such high-density arrays are what is generally meant by the term 'DNA chip'. The substrate used for immobilisation of high-density DNA arrays is usually glass, which is sometimes chemically modified, or sili-

con. The library of nucleic acids on the chip may comprise oligonucleotides of up to usually 25 to 60 bases in length, or PCR fragments or sometimes plasmids of between 500 and 5,000 bases.

The two main uses of high-density DNA arrays are quantitative measurement of gene expression and analysis of sequence of nucleic acids. The pattern and composition of the nucleic acids on the array must be chosen such as to reflect these uses.

4.2.1.2.4 Analysis of gene expression using DNA microarrays

To analyse gene expression, nucleic acids that match the sequences of the mRNAs to be measured are applied onto the array surface. In general, it is not useful to directly hybridise the mRNA from the tissue under investigation to the array, as no satisfactory detection system is available for unlabelled nucleic acids. For that reason, the mRNAs are usually reverse transcribed to cDNA. Because of the low efficiency of the reverse transcription process and the generally small amount of starting material (*e.g.*, from tissue biopsies or blood samples), the amount of cDNA obtained is too low for most applications. For this reason, the cDNA must be amplified. It is important, however, to choose a method that avoids changing the ratio of the nucleic acid species to each other. Two techniques are commonly used, either a PCR with a low number of cycles, or *in vitro* transcription. In case of *in vitro* transcription, a promoter for RNA polymerase is added to the cDNA strand during reverse transcription of the mRNA from the sample. The cDNA is then amplified by a factor of about 100 to 1,000 using RNA polymerase. During *in vitro* transcription, the RNA molecules are usually fluorescently-labelled. However, if RNA is not limiting the experiment, labelled cDNA can be directly used for hybridisation with the array. As a note, several modified procedures for the amplification of RNA samples and for increasing signal intensity for have been described (Ivashuta *et al.*, 2002; Karsten *et al.*, 2002; Mahadevappa & Warrington, 1999; Nallur *et al.*, 2001; Pabón *et al.*, 2001; Phillips & Eberwine, 1996; Stears *et al.*, 2000; Wang *et al.*, 2000a; Zhumabayeva *et al.*, 2001).

Two general procedures are available for the estimation of gene expression and, in particular, for estimation of differences in gene expression between different samples (figure 4.1.A and 4.1.B). In the first procedure, labelled nucleic acids from the two samples to be compared are hybridised to two identical arrays. In order to normalise the signal intensity of single genes, internal standards and/or complex statistical methods are used. Usually, housekeeping genes that have a steady rate of expression can be used as internal standards (for problems associated with use of housekeeping genes, see chapter 1). Alternatively, the mean signal intensity is calculated and used for normalisation. Several methods have been developed for normalisation, calibration, variance, quality and statistical analysis of DNA microarrays but their description goes far beyond the scope of this chapter. Finally, differences in normalised signal intensities indicate different rates of expression. See for further information on normalisation, calibration, variance, quality and statistical analysis: Alter *et al.*, 2000; Baldi & Long, 2001; Beissbarth *et al.*, 2000; Brazma *et al.*, 2001; Brown *et al.*, 2001; Chudin *et al.*, 2002; Claverie, 1999; Draghici *et al.*, 2001; Eickhoff *et al.*, 1999; Herwig *et al.*, 2001; Hill *et al.*, 2001; Ideker *et al.*,

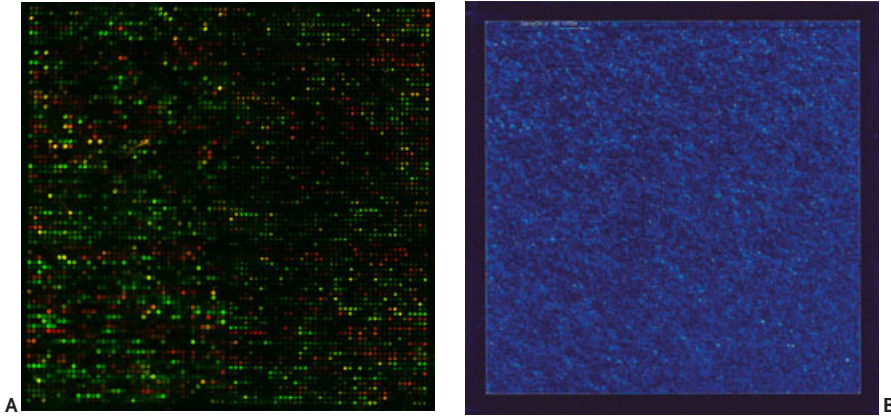


Figure 4.1 Two examples of DNA microarrays for gene expression analysis. A) DNA microarray using the two-colour detection system. Two RNA samples (labelled either with Cy3™ or Cy5™) were mixed and hybridised to an array containing 6,116 yeast RNAs and controls. The figure was kindly provided by Dr. Patrick O. Brown (Stanford University, California, USA). B) Human Genome U95A array (Affymetrix,

Inc., Santa Clara, San Diego, USA) hybridised with amplified and biotin-labelled cRNA (detected after two rounds of incubation with streptavidin-bound phycoerythrin) derived from differentiated macrophages derived from the human monocyte THP-1 cell line. The array contains more than 12,000 probe sets representing human transcripts.

2000; Kadota *et al.*, 2001; Kerr & Churchill, 2001; Kerr *et al.*, 2000; Lee *et al.*, 2000; Li & Wong, 2001; Long *et al.*, 2001; Manduchi *et al.*, 2000; Mutch *et al.*, 2001; Naef *et al.*, 2002; Newton *et al.*, 2001; Quackenbush, 2001; Rifkin *et al.*, 2000; Rocke & Durbin, 2001; Schadt *et al.*, 2000 and 2001; Schuchhardt *et al.*, 2000; Thomas *et al.*, 2001; Troyanskaya *et al.*, 2001; Tseng *et al.*, 2001; Tsodikov *et al.*, 2002; Tusher *et al.*, 2001; Wang *et al.*, 2001a; Yang *et al.*, 2001 and 2002; Yuen *et al.*, 2002; Zhang & Zhao, 2000; Zien *et al.*, 2001).

The second procedure for comparison of gene expression requires only one array for two samples. In this case, the two samples are labelled differently, usually by means of two different fluorophores, for example, Cy3™ (IUPAC name of the water-soluble cyanine dye: 2-[(1*E*,3*E*)-5-(1-{6-[2,5-dioxo-1-pyrrolidinyl]oxy}-6-oxohexyl)-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene)-1,3-propadienyl]-1-ethyl-3,3-dimethyl-3*H*-indolium-5-sulphonate) with an absorption wavelength of 552 nanometres and an emission wavelength of 568 nanometres, and Cy5™ (IUPAC name of the water-soluble cyanine dye: 2-[(1*E*,3*E*)-5-(1-{6-[2,5-dioxo-1-pyrrolidinyl]oxy}-6-oxohexyl)-3,3-dimethyl-5-sulfo-1,3-dihydro-2*H*-indol-2-ylidene)-1,3-pentadienyl]-1-ethyl-3,3-dimethyl-3*H*-indolium-5-sulphonate) with an absorption wavelength of 650 nanometres and an emission wavelength of 667 nanometres. Hybridisation on the array is performed using a one-to-one ratio of labelled nucleic acid probes. Genes with identical levels of expression are represented by a mixed colour, while with genes with differing rates of expression, one of the fluorophores dominates over the other.

The procedure described above is usually performed for arrays comprising PCR products or plasmids. The simple process is an advantage of this method. However, the inhomogeneity of hybridisation kinetics in a mixture that contains such a variety of different and relatively long nucleic acid molecules is a disadvantage, the reason being that the conditions for hybridisation such as temperature and salt concentration are never optimal for all molecules. As a result, unspecific or incomplete hybridisation may occur leading to incorrect results. In addition, since the nucleic acids affix to the substrate in random configuration, many will not be available for hybridisation with DNA in the sample for steric reasons.

A solution to this dilemma lies in the use of oligonucleotide microarrays comprising nucleic acid probes in the range of 15 to 25 bases (sometimes up to 60 bases) in length. These oligonucleotides can be selected for their chemical properties such as to ensure optimal hybridisation. The Affymetrix company (Santa Clara, San Diego, USA) has developed an elegant solution to this problem. At the time of writing, their array uses eleven to twenty different 25-mer oligonucleotides for each mRNA to be monitored. These oligonucleotides on the microarray are exactly complementary to different sequences within the 3'-untranslated region of mRNAs and are therefore called perfect match oligonucleotides (figure 4.2). Since the melting temperature of the hybrid of the microarray oligonucleotide and the sample

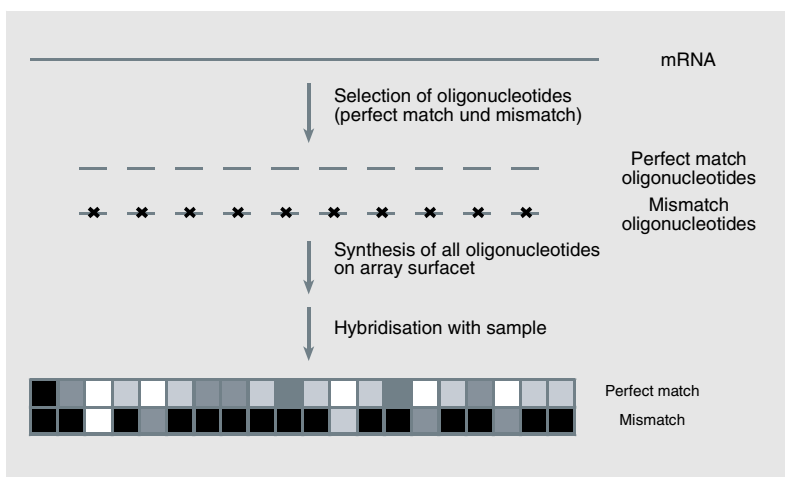


Figure 4.2 Expression analysis with a high-density oligonucleotide array (Affymetrix, Inc., Santa Clara, California, USA) using the perfect match/mismatch system. Several sections of the mRNA are represented on the surface of the array as 25-mers (perfect match probe). Mismatch oligonucleotides differ from perfect match oligonucleotides by virtue of a single base exchange in the centre of the oligonucleotide. Comparison of signal intensities between perfect match and mismatch probes is

used to decide on the specificity of binding. Area 1 shows an example where the sample hybridises neither to the perfect match nor to the mismatch oligonucleotide; such an area is excluded from evaluation. Similarly, area 3 is not evaluated because the sample hybridises to the perfect match and the mismatch oligonucleotide with equal intensity, indicating non-specific hybridisation. All other areas are assumed to show specific hybridisation and are subjected to analysis.

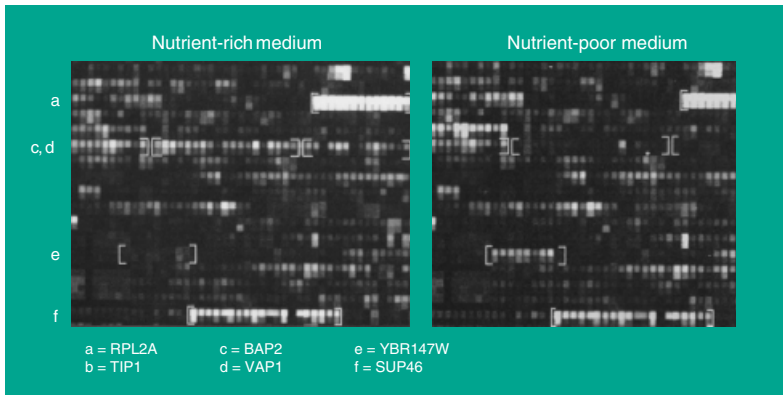


Figure 4.3 Details of fluorescence images of oligonucleotide microarrays that have been hybridised with cRNA samples recovered from mRNA of cultured yeast. Yeast cells were grown either in nutrient-rich or nutrient-poor medium at 30 °C. The tagged areas show perfect match and mismatch oligonucleotides that represent RNAs expressed with different intensities in the two cultures. With permission from Affymetrix, Inc., Santa Clara, California, USA.

DNA can be calculated on the basis of the sequence, oligonucleotides with the best hybridisation properties can be selected based on the mRNA sequence of the gene whose expression is to be studied. Thus, even when certain oligonucleotides hybridise poorly or not at all, sufficient remain to enable the results of the analyses with the array to be interpreted correctly. As a negative control in addition to the perfect match oligonucleotides, the Affymetrix microarray also contains the same number of mismatch oligonucleotides that differ from the perfect match sequences by the presence of a single mismatched base in the centre of the oligonucleotide. Ideally, there is no hybridisation under optimal conditions to the mismatched oligonucleotide so that signals must derive from unspecific hybridisation. The differences in signal intensity between the perfect match and the mismatch oligonucleotides allow evaluation of non-specific binding. An algorithm calculates the corrected intensities for the perfect match oligonucleotides that are compared to the intensity of a second array hybridised to another probe. The final reported difference in the amount of an mRNA between two samples results from integration of the results of comparison of all corrected perfect match oligonucleotides representing that mRNA (figure 4.3). The technique for manufacture of such arrays is described below. While undoubtedly a most useful technology, readers should be aware of the difficulties associated with such arrays as described in detail in the contribution by Peng Liang in the contribution on fluorescent differential display in this chapter.

4.2.1.2.5 Re-sequencing by hybridisation

It is noteworthy that the use of microarrays is not only limited to gene expression analysis. Use of high-density oligonucleotide arrays for re-sequencing by hybridisation allows the identification of mutations and single nucleotide polymorphisms (SNPs) in known gene sequences. Oligonucleotides complementary to certain

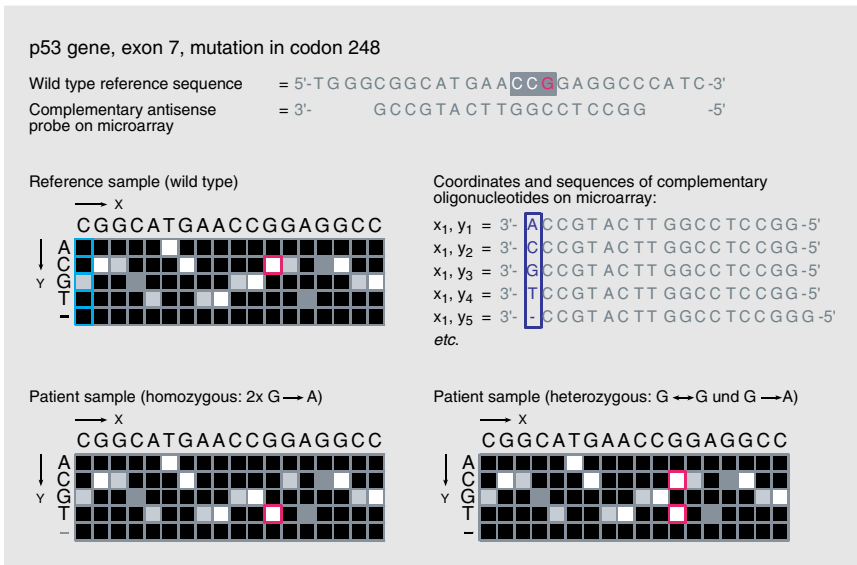


Figure 4.4 Re-sequencing of DNA using an oligonucleotide microarray. The sample hybridises only to oligonucleotides with a perfectly complementary sequence. Because the location and sequence of the oligonucleotides on the array are known, the sequence of the sample

can be deduced from the hybridisation pattern. The x-axis shows the sequence read from the hybridisation pattern on the array, the y-axis determines the identity of the base at that position on the array.

parts of the gene are fixed to the DNA microarray. After amplification and labeling, the sample to be investigated is hybridised to the oligonucleotides on the array. Use of an array that contains oligonucleotides complementary to each of the four bases at each position effectively allows an unknown DNA sequence in the sample to be determined based on the hybridisation pattern alone (figure 4.4). At present, sequencing by means of DNA arrays is limited by difficulties in optimising hybridisation conditions, so that confirmation of any mutations found is necessary before this technique can be used for medical diagnostic purposes. Moreover, this technique may be prone to error if multiple mutations or complex sequence deviations such as insertions or deletions are present. For these reasons, conventional sequencing is currently a much more efficient method. In the future, of course, this situation may change with advances in array and experimental design, as well as with progress in bioinformatics.

4.2.1.3 Production of DNA microarrays

DNA is applied to the surface of the array in one of two ways, immobilisation of pre-formed nucleic acids or synthesis of oligonucleotide *in situ*. Many immobilisation techniques exist, including *contact-tip deposition printing*, *micro-contact printing* (μ CP), *micro-fluidics networks* (μ FN) and *electro capture*. Piezoelectric printing and

micro wet printing (μ WP) may be used both for immobilisation and for *in situ* synthesis. The photolithographic technique is used for *in situ* synthesis alone. The following provides a brief description of each of these techniques.

4.2.1.3.1 Contact-tip deposition printing

This technique, commercialised in 1997 by the Synteni company (Fremont, California, USA; now Incyte Genomics, Inc., Palo Alto, California, USA), is used by a large number of research teams to generate their own DNA arrays. The nucleic acid to be immobilised is first dissolved. A needle is dipped into this solution so as to leave a defined amount of solution at its tip. This solution is then applied to the surface of the substrate (figure 4.5). In practise, several needles are used simultaneously. In most cases, the needles have a groove that serves as a reservoir for the solution as with a conventional ink pen (figure 4.6; Schena *et al.*, 1995).

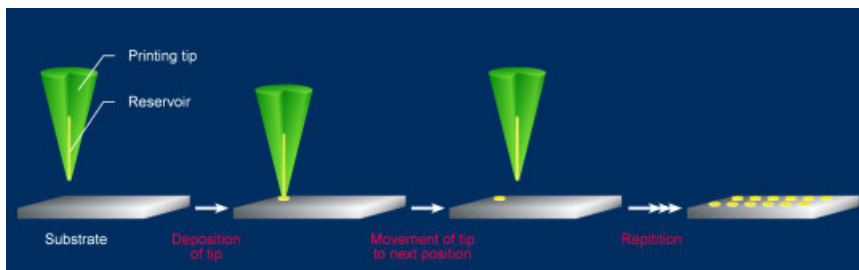


Figure 4.5 Production of a DNA microarray using contact tip deposition printing. See text for details.

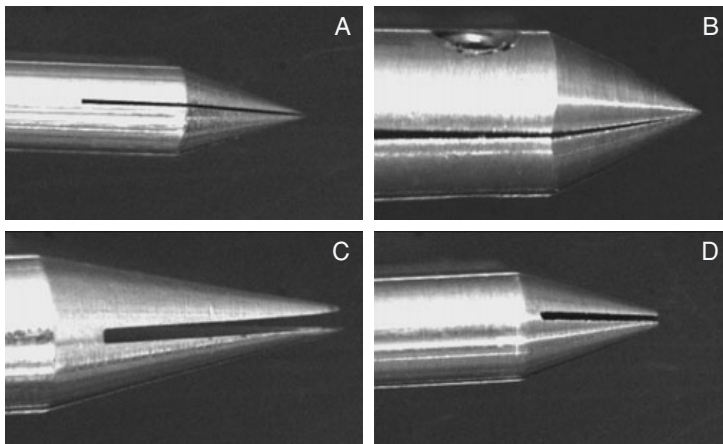


Figure 4.6 Examples of tips used for contact tip deposition printing. **A)** Tip with a 0.004 centimetre groove for generation of spots 50 to 120 millimetres in diameter. **B)** Tip with a 0.025 centimetre groove for generation of spots 100 millimetres in diameter. A screw (upper picture) is used to adjust the spot size. **C)** Tip with a 0.025 centimetres groove for generation of large spots of various sizes. **D)** Tip with a groove of 0.20 centimetres for preparation of large probe spots of variable diameter. All pictures courtesy of Dr. Patrick O. Brown (Stanford University, California, USA).

The company Genetic Microsystems (Woburn, Massachusetts, USA; now Affymetrix, Inc., Santa Clara, CA, USA) introduced a variation of this technique called the '*pin-and-ring array*' system in 1998 (Rose, 1998). A small ring of several millimetres in diameter is dipped into the nucleic acid solution. A membrane or lamella is formed in the ring by surface tension in exactly the same way as a soapy membrane forms in a child's toy for blowing soap bubbles. The ring is positioned over the array and a flat-topped pin is inserted through the lamella until it makes contact with the surface of the array. In this way, the small, defined portion of the lamella adhering to the tip of the pin is deposited on the array (figure 4.7).

Contact-tip deposition printing may be used to produce arrays of high density. For example, Incyte Genomics and Synteni produced arrays containing about 10,000 different cDNAs ranging in size from 500 to 5,000 bases on a glass surface of 3.6 square centimetres. The companies aimed to use their technique to produce arrays containing up to 100,000 probes on a surface measuring 6.5 square centimetres. The reproducible minimal spot diameter was in the range of 50 millimetres, with an inter-spot distance also in this range.

Nucleic acids bind poorly to plain glass, which is therefore generally treated before use. Poly-L-lysine, which binds molecules by ionic interaction, is often used for this purpose. Such arrays can be used only once, however, because the salt conditions required to remove the hybridised sample DNA also dissociate the immobilised probes from the surface of the array. In addition, the nucleic acid probes on such coated arrays are entwined, reducing the hybridisation kinetics by steric hindrance. This problem may be avoided or at least reduced by the use of linker

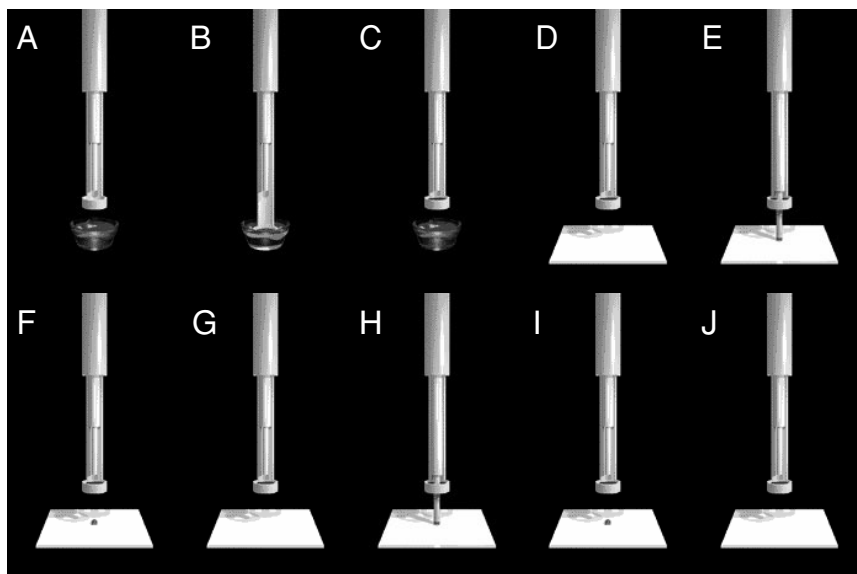


Figure 4.7 Schematic diagram of array production using *pin-and-ring array technology*. See text for details. The figure is reproduced with permission from Affymetrix, Inc. (Santa Clara, CA, USA).

molecules such as derivatives of oligoethyleneglycol to separate the immobilised DNA from the surface of the array.

The glass substrate may also be treated with silane, which binds covalently to the probe DNA by means of a Schiff-base reaction, thus preventing removal of the probe DNA during hybridisation and washing steps and allowing the arrays to be reused.

Glass slides which are used to detect radioactively-labelled nucleic acids may be coated with a nylon membrane to which the DNA probes are covalently cross-linked using ultraviolet light.

Due to the limited amount of technical effort required for contact-tip deposition printing, smaller laboratories are using this method to generate their own (custom-made) arrays. Commercial printing machines ('spotters') are broadly available. The group of Pat Brown at the University of Stanford in California has even provided instructions for a do-it-yourself printing apparatus on the internet (<http://cmgm.stanford.edu/pbrown/mguide/index.html>).

4.2.1.3.2 μ CP – Micro-contact printing

The *micro-contact printing method* (μ CP) works in a method analogous to contact-tip deposition printing. In μ CP, a polydimethylsilane (PDMS) stamp is used to transfer the nucleic acids to the carrier surface (figure 4.8). The advantage of this material is that it can be used to produce very small structures. At the time of writing, stamps are already being produced in reproducible fashion with defined structures of less than 50 nanometres. In theory, this opens up a new level of miniaturisation (Kumar & Whitesides, 1993). However, practical results obtained with the μ CP method have been disappointing. While μ CP has been used to immobilise antibodies on gold substrates (Morhard *et al.*, 1988), this technique has not yet succeeded in producing the density of probes generally required for the manufacture of DNA arrays.

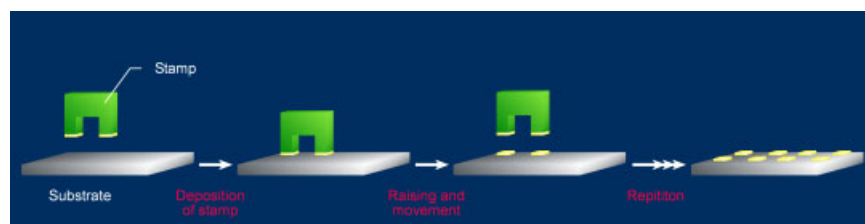


Figure 4.8 DNA array production using *micro-contact printing*. See text for details.

4.2.1.3.3 μ FN – Micro-fluidics network

Micro-fluidics network (μ FN) technology is a further development of μ CP. In μ FN, a PDMS stamp containing small channels is placed on a glass, gold, polystyrene or silicone/silicone dioxide surface (figure 4.8). These small channels are filled with a solution containing the substrate, which perfuses the surface of the array by capillary attraction. This system has been used to immobilise antibodies following activation of the substrate by N-hydroxysuccinimidyl esters, which undergo chemical

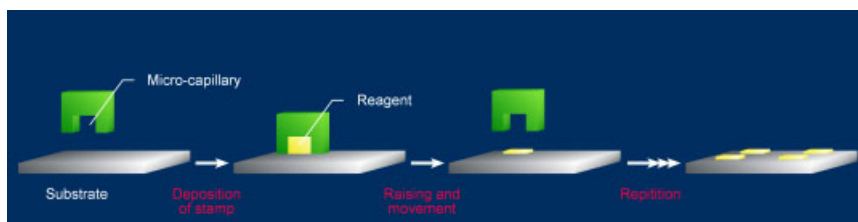


Figure 4.9 Array production using *micro-fluidics network technology*. See text for details.

coupling with the amino group of the proteins (Delamarche *et al.*, 1997). However, occlusion of the capillaries may occur and the number of suitable solvents is limited. As with μ CP, the feasibility of μ FN for production of DNA arrays has yet to be demonstrated. At the time of writing, PDMS stamps measuring three millimetres $\times$ one millimetre have been used to generate 100 capillaries with a depth of 1.5 millimetres, a width of three millimetres, and a length of three millimetres on a variety of substrates (Ligler *et al.*, 1988).

4.2.1.3.4 Piezoelectric printing

Piezoelectric printers use the technology that was developed for conventional ink-jet printers to dispense small amounts of DNA solution instead of conventional ink (figure 4.10; Theriault *et al.*, 1999). Using silicon and glass pipette tips of less than 100 micrometres in diameter, volumes of between 15 picolitres (15×10^{-12} litres) and 500 picolitres can be applied with a frequency of up to two kilohertz and a coefficient of variation of less than one percent. A scheme of a piezoelectric tip is given in figure 4.11 and figure 4.12 shows the generation of a drop at such a dispenser tip. Microarrays produced using piezoelectric printing with up to 10,000 spots per square centimetre are available. Commercially available piezoelectric systems allow production of arrays with spot diameters of about 200 micrometres, inter-spot distances of 300 micrometres and a spot density of more than 1,400 spots square centimetre (Cheung *et al.*, 1999; Schena & Davis, 1999). It has been shown that the spot size can be reduced even to 25 to 30 micrometres (Harris *et al.*, 2000; Okamoto *et al.*, 2000). Piezoelectric dispenser systems are able to target solutions with impressive accuracy (figure 4.13).

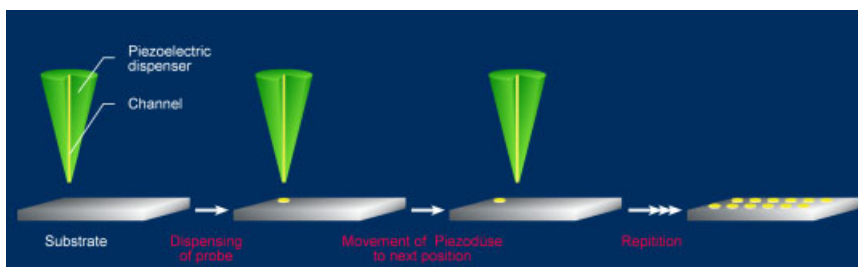


Figure 4.10 Principle of the *piezoelectric printing* method. See text for details.

Figure 4.11 Schematic diagram of a piezoelectric jet. The following abbreviations are used in the figure: m, metre; s, second; $|U|$, absolute value of voltage; μs , microsecond; V , volt. Picture courtesy of Microdrop Gesellschaft für Mikrodosiersysteme mbH (Norderstedt, Germany).

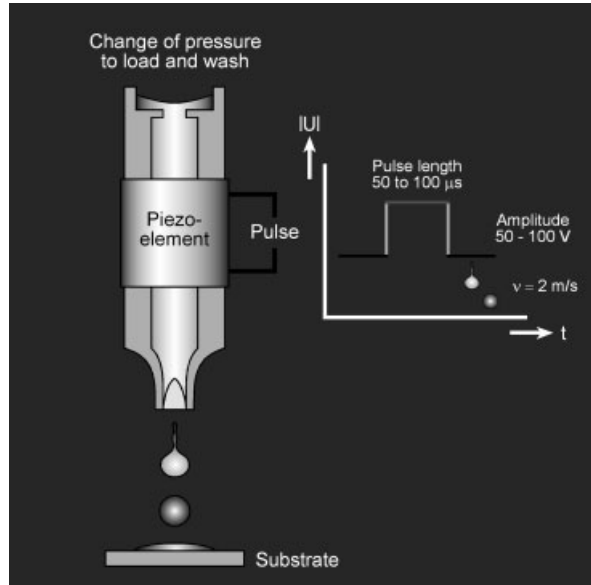


Figure 4.12 Formation of a liquid droplet at the dip of a piezoelectric dispenser. The figure was kindly provided by Microdrop Gesellschaft für Mikrodosiersysteme mbH (Norderstedt, Germany).

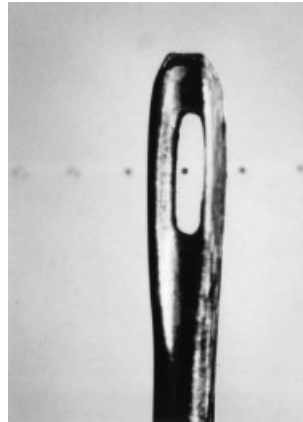


Figure 4.13 Piezoelectric jets can be controlled with such precision that a drop can easily be dispensed through the eye of a needle. Picture courtesy of Microdrop Gesellschaft für Mikrodosiersysteme mbH (Norderstedt, Germany).

Contamination leading to poor reproducibility is a problem associated with piezoelectric dispenser systems. This may result from vaporisation of the small probe volumes and by satellite spots, which result from splashing as the drop impacts the surface of the substrate. Attempts are being made to deal with this problem by developing surfaces with modified degrees of wettability. The Protogene company (Palo Alto, California, USA), for example, used substrates with hydrophilic areas separated by hydrophobic barriers (Blanchard *et al.*, 1996; Brennan, 2001). Alternatively, photolithographically-produced polyacrylamide cushions may be used. Polyacrylamide gel has a high capacity for nucleic acids, which reduces the problem of vaporisation (Dubiley *et al.*, 1997 and 1999; Fotin *et al.*, 1998; Guschin *et al.*, 1997; Proudnikov *et al.*, 1998; Yershov *et al.*, 1996).

Contact of the nozzle with the surface of the substrate should be avoided as this may produce vibrations, causing imprecise application of the probe and damage to either the substrate or the tip.

Piezoelectric dispensers may also be used to synthesise oligonucleotides *in situ* (figures 4.12.A and 4.12.B; Blanchard *et al.*, 1998; Theriault *et al.*, 1999). Highly

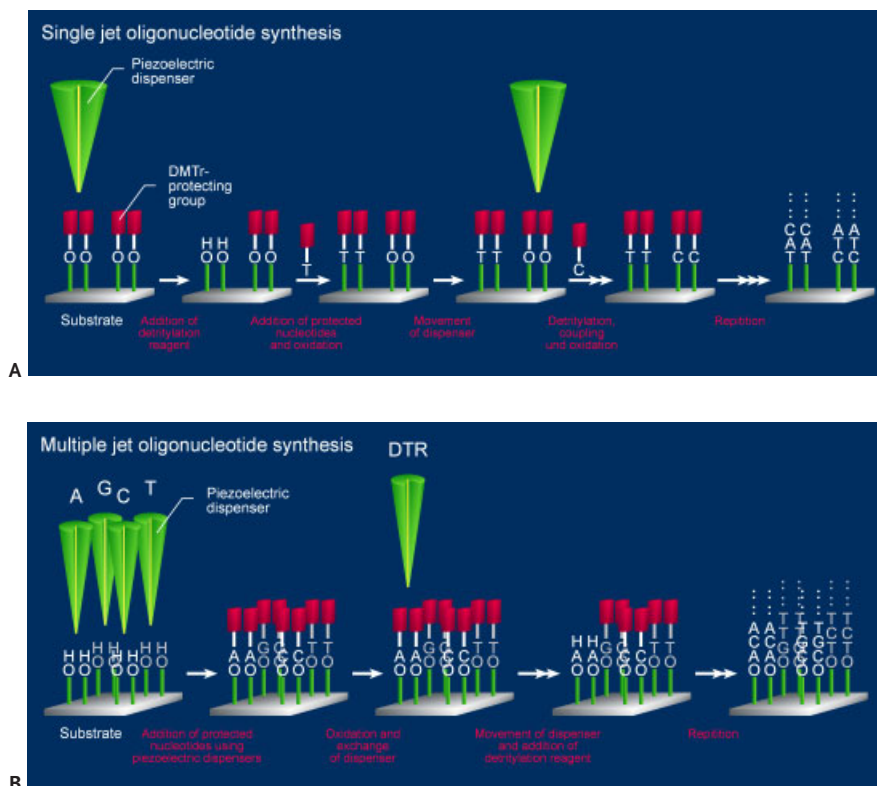


Figure 4.14 *In situ* oligonucleotide synthesis using a piezoelectric dispenser. **A)** Single jet procedure. **B)** Multi jet procedure (see text for details). The following abbreviations are used in this figure: DMTr, dimethoxytrityl protecting group; DTR, de-tritylation reagent.

specialised nozzles are used, which limits the ability to employ different solutions and reagents with various physical properties in the same dispenser. Two strategies may be used to solve this problem, both of which utilise a modified phosphoramidite procedure to synthesise the oligonucleotides. In the single jet system (figure 4.14.A), a single dispenser is used to apply a reagent that de-protects the 5'-hydroxyl group of the nucleotide derivatives at defined positions on the substrate. Coupling of a nucleotide to the unprotected hydroxyl group and phosphorus oxidation is then achieved by moistening the entire substrate surface with the appropriate reagent. In the multiple jet system, five different dispensers are used, each of which contains one of the four phosphoramidite nucleotides needed for oligonucleotide synthesis (figure 4.14.B). The fifth dispenser is needed for the de-tritylation reagent.

4.2.1.3.5 Electro capture

A completely different procedure to piezoelectric printing has been developed by the Nanogen corporation (San Diego, California, USA) (figures 4.15 and 4.16) (Cheng *et al.*, 1998; Edman *et al.*, 1997; Gilles *et al.*, 1999; Radtkey *et al.*, 2000; Sosnowski *et al.*, 1997; Westin *et al.*, 2000). This active programmable electronic device technology is based on a silicon chip containing an array of one square centimetre, which can be loaded with 25, 64, 100 or platinum electrodes. These chips are manufactured by generating a layer of silicone dioxide on a silicon carrier by means of thermic oxidation. This layer is coated with aluminium and then with a photoresist. Using a mask and ultraviolet light, the resist is developed and removed in certain areas, to which a 20 nanometres layer of chrome is applied, followed by platinum electrodes of 500 nanometres in thickness. The remaining resist and the aluminium coating are then removed. The whole chip is then covered by a two millimetres dielectric layer of trisilicone tetranitride (Si_3N_4), and afterwards with photoresist. This resist is developed and removed exactly above the platinum electrodes, exposing them. The Si_3N_4 layer is then removed so that the electrodes are freely accessible from the top and isolated at the borders by Si_3N_4 . Finally, the residual resist is removed and the array is coated with streptavidin-derivatised agarose (figure 4.15 and 4.16). The agarose layer supports the immobilisation of biotinylated molecules on the surface of the array.

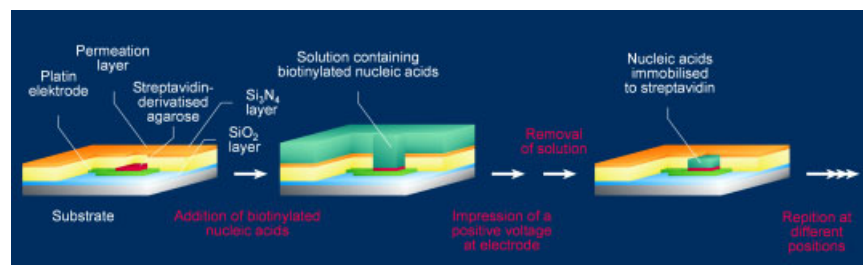


Figure 4.15 *Electro capture (electrochemical focussing)*. See text for details.

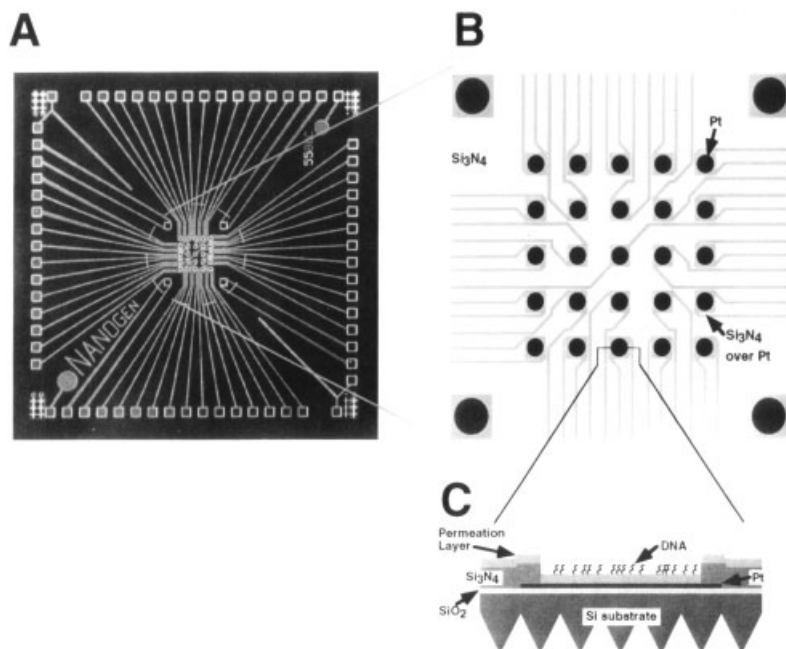


Figure 4.16 Diagram of an electric silicon chip.

A) Overview of a chip with 25 platinum electrodes with an edge length of one centimetre. The bright squares in the peripheral areas of the array represent the platinum electrodes by which the chip is connected to the power source. The platinum connections are visible as bright lines running between the peripheral electrodes and the smaller electrodes in the centre of the chip. **B)** Section of the area of the electrodes in the centre of the chip. This one square centimetre central area consists of an electrode with a diameter of 160 micrometres in each corner and 25 electrodes, each with a diameter of 80 micrometres, arranged in a square. Pt: areas of the platinum electrodes

with free access. Si₃N₄: areas coated with the dielectric trisilicone tetranitride. Si₃N₄ over Pt: Areas of the platinum electrodes that are isolated by a covering of Si₃N₄. **C)** View of the area of the electrode. The location of the electrode on the chip is shown by lines as in part **B)** of the figure. Pt: platinum electrode; Si₃N₄: isolating dielectric layer; SiO₂: inert protective layer consisting of silicone dioxide on the silicon substrate; permeation layer: streptavidin-coupled agarose layer; DNA: biotinylated nucleic acid molecule, coupled to streptavidin. The figure is reproduced with permission from *Proceedings of the National Academy of Sciences of the USA* (Sosnowski *et al.*, 1997).

Nucleic acid arrays are produced by covering the surface with a solution of biotinylated oligonucleotides. These are specifically transported to the streptavidin-containing spots by applying a positive voltage to the each of the electrodes in turn. Immobilisation is achieved by the high binding affinity of the streptavidin/biotin interaction ($K_D = 10^{-15}$ mole). At the time of writing, chips with up to 400 electrodes are under development for genetic analyses while for expression analyses there are plans for arrays with 1,000 to 10,000 electrodes (<http://www.nanogen.com>).

An advantage of these arrays is that the efficiency of hybridisation can also be influenced by applying an electric potential to the electrodes during addition of the sample solution (Heller & Tu, 1993; <http://www.nanogen.com>). Detection is by means of conventional optical systems.

4.2.1.3.6 XNA on Gold™

Interactiva (now Thermo Hybaid GmbH, Ulm, Germany) has developed a biochip suitable for all common biotinylated biomolecules (called 'XNA on Gold™'; figure 4.17). Using thin film immobilisation technology, a self-assembling monolayer of long-chain thioalkanes is covalently bound *via* sulphur atoms to a ten nanometres layer of 24-carat gold that is applied to a glass microscope slide by vapour deposition (Mecklenburg, 1999; Pisano, 2001; <http://www.interactiva.de/>). Biotin is covalently linked to the thioalkanes and saturated with streptavidin. Biotinylated nucleic acids, sugars, peptides or lipids can then be immobilised by binding to the streptavidin layer. To avoid cross contamination, photolithographic thick film tech-

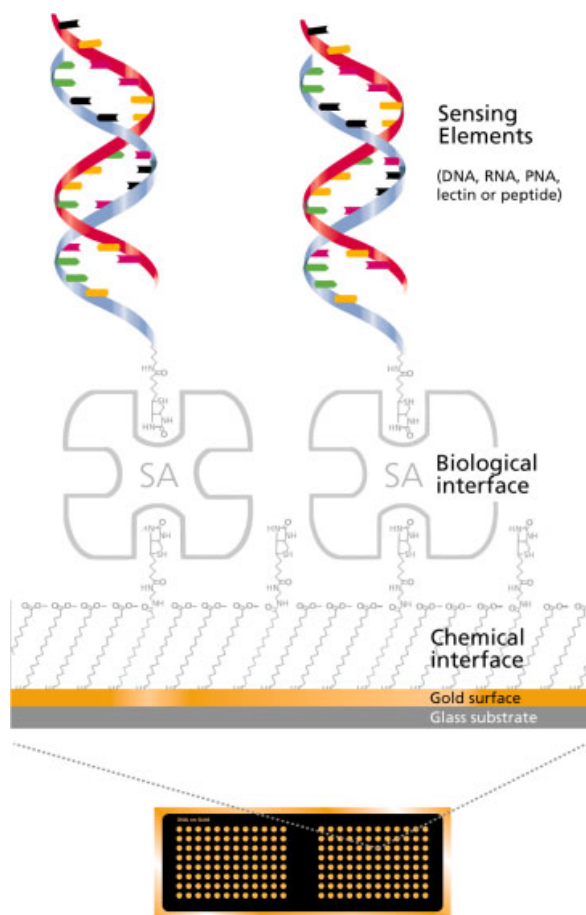


Figure 4.17 Schematic view of a section of an XNA on Gold™ chip. On a carrier with the size of a microscopic slide, two arrays of 96 probes are arranged. The glass substrate is first covered by 24-carat gold. Self-orientating long-chain thioalkanes are then fixed to this layer in order to form a chemical interface. Biotin is then used to tightly bind a streptavidin layer to the thioalkanes and, thus, to form a biological interface. Biological samples such as nucleic acids or peptides can be immobilised to this interface. Figure courtesy of Dr. Michael Mecklenburg (nanoArc Corp., Newport Beach California, USA).

nology is used to cover the surface of the chip with a hydrophobic layer of Teflon™ before immobilisation of the probes. This results in a pattern of cavities 50 millimetres deep and 1.5 millimetres wide. Because of the surrounding hydrophobic array, the aqueous solutions of biomolecules collect in these cavities. At the time of writing, two times 96 or 384 probes can be immobilised using this technology on a slide measuring 75 millimetres × 25 millimetres. Detection is possible by means of chemiluminescence, fluorescence or auto-radiography.

4.2.1.3.7 Photolithography using chrome/glass masks

Affymetrix, Inc. (Santa Clara, California, USA) has developed a photolithographic procedure for solid-phase *in situ* synthesis of oligonucleotides (Lipshutz *et al.*, 1999). This process is based on techniques derived from the semiconductor industry. Glass covered by a photosensitive protective layer is used as a substrate (figure 4.18). Using a mask, areas on the surface are activated by ultraviolet light, causing chemical activation of the surface by elimination of the photosensitive protective groups. The surface is then incubated with nucleotides, which react with free hydroxyl groups at the surface of these activated areas. Using a second mask, other areas are then activated to react with other nucleotides. The nucleotides that are bound to the surface carry protective groups coupled to the 3'-position of the ribose subunit, and are therefore available for binding other nucleotides only after the protective group has been removed by photo-activation. By repeating the cycle of targeted photo-activation and incubation with protected nucleotides, a compact lawn of defined oligonucleotides is produced on the substrate. More than 65,000 different octamers can be produced by this procedure within 32 cycles of synthesis by skilful selection of the 32 masks. To synthesise the 25-mers present on the finished array, up to 80 cycles of synthesis with 80 different chrome/glass masks are necessary. These costly masks can be used only once. At the time of writing, more than 500,000 different oligonucleotides can be applied to a surface measuring 1.28 × 1.28 centimetres (<http://www.affymetrix.com>).

The photosensitive protective groups that have been developed by Affymetrix for the synthesis of oligonucleotides are characterised by the following formula

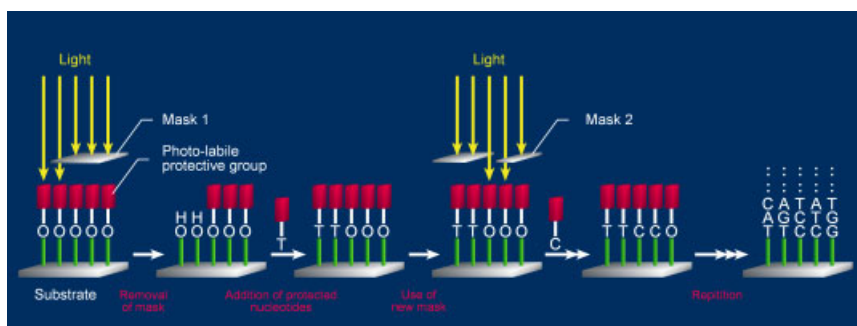


Figure 4.18 Photolithographic production using photosensitive protective groups and photolithographic masks. See text for details.

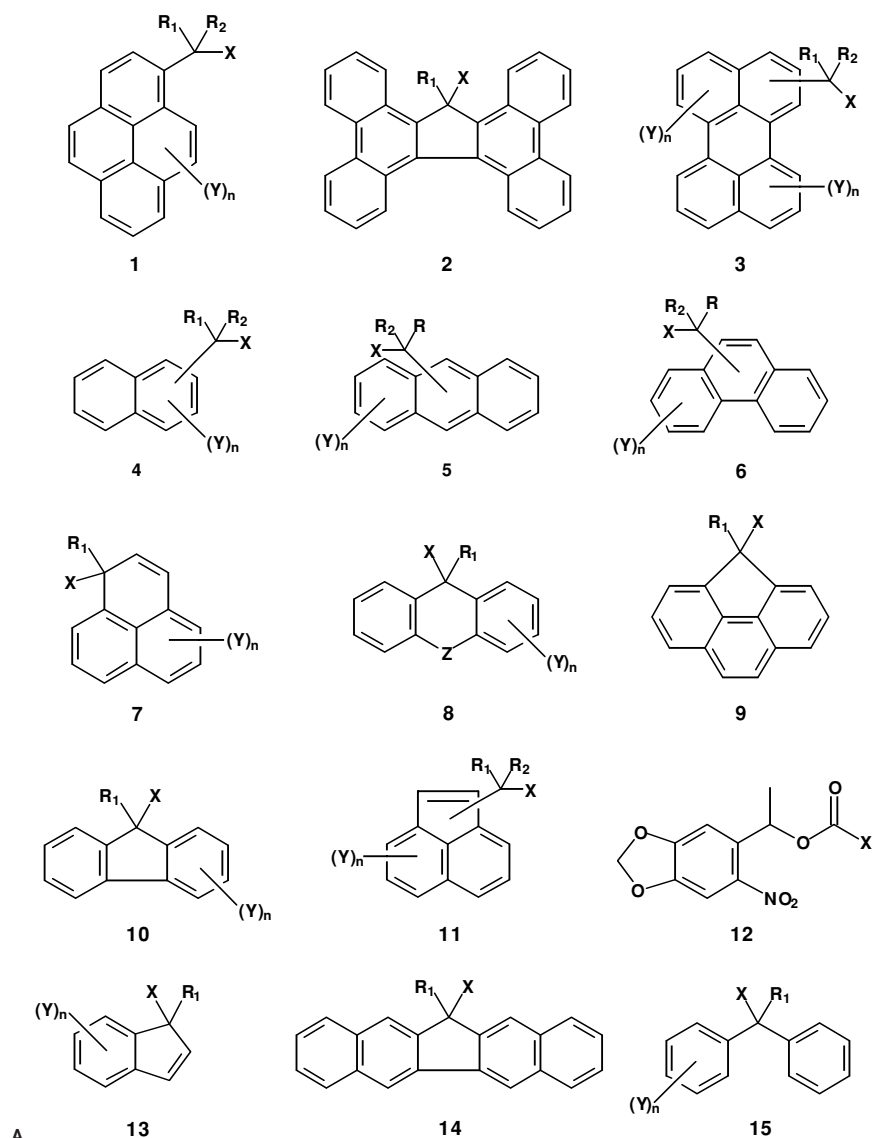
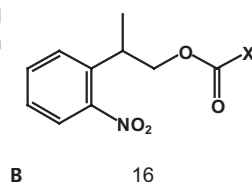


Figure 4.19 Photosensitive protective groups for photolithographic *in situ* oligonucleotide synthesis. **A)** Groups the general formula $Ar-C(R_1)(R_2)-O-C(O)-X$ (Nam *et al.*, 1998). The identities of the 'X' and $R^{1,2}$ moieties are explained in the text. Compound **8** is characterised as follows: $Z = O, S, N$ -alkyl-, $CH_2, CH_2-CH_2, CH=CH$, and Y is generally an alkyl group, a substituted aromatic compound or an electron donor group. **B)** The photosensitive protecting group NPPOC (compound **16**) used by Beier & Hoheisel (1999) for photolithographic oligonucleotide synthesis is more efficient than that obtained using the conventional protecting group MeNPOC (compound **12**) and PYMOC (compound **1**).



Ar-C(R¹)(R²)-O-C(O)-X (figure 4.19). 'Ar' is an optional substituted, condensed polycyclic aryl-group, a heteroaromatic group or a vinyl-like derivative of those groups. R¹ and R² are hydrogen groups that are independent from each other, an optional substituted alky-, alkenyl-, alkynyl-, aryl- group or vinyl like derivative of those. 'X' represents a group to be split off, a chemical fragment to be coupled to the Ar-C(R¹)(R²)-O-C(O)-group by a heteroatom or a solid surface. 'Ar' representing a condensed polycyclic, aromatic hydrocarbon and R¹ and R² electron donors are most suitable compounds for oligonucleotides synthesis. Affymetrix prefers to use the substituent PYMOC (pyrenyl-methyl-oxy-carbonyl-group) with an aromatic pyrenyl-group and hydrogen substituted groups R¹ and R² (compound 1 in figure 4.19.A) and MeNPOC (α -methyl-o-nitro-piperonyl-oxy-carbonyl-group; compound 12) (Nam *et al.*, 1998).

Markus Beier and Jörg D. Hoheisel recently published that the photo-cleavable 2-(2-nitrophenyl)propyloxycarbonyl-groups (NPPOC-groups, compound 16 in figure 4.19.B) represent a new class of rapidly detachable photo-protecting groups of significantly higher coupling efficiencies compared to conventional MeNPOC-groups (Beier & Hoheisel, 1999 and 2000). The enhanced photo-cleavage of the NPPOC-groups takes place *via* a base-assisted β -elimination process.

Disadvantages of these methods are the high costs for the photolithographic masks and the insufficient synthesis recovery of 95 percent per cycle that have a lasting effect on size and quality of the oligonucleotides. Only about 30 percent of the 25 oligomers are synthesised in the desired way.

Affymetrix is offering numerous oligonucleotide arrays for expression analysis or re-sequencing of nucleic acids in the format of 1.24×1.24 centimetres with a maximum density of up to 500,000 oligonucleotides per square centimetre. The oligonucleotides of the Human Genome U133 Array Set represent almost 45,000 probe sets representing more than 39,000 transcripts derived from approximately 33,000 well-substantiated human genes (<http://www.affymetrix.com>).

4.2.1.3.8 Photolithography using digital micro-mirrors

Recently, researchers at the University of Wisconsin reported on a photolithographic technique that avoids the use of expensive chrome/glass masks (Blanchard & Friend, 1999; Singh-Gassom *et al.*, 1999). Using television technology similar to that used in computer beamers (Texas Instruments, Inc., Dallas, Texas, USA), the investigators manufactured a so-called 'virtuous mask' containing up to 480,000 computer-directed micro-mirrors with an edge length of 16 millimetres. By varying the angle of tilt, these mirrors are capable of exactly position sharp spots of ultra-violet light on the surface of an array (figure 4.20). This de-protects discrete areas and allows an oligonucleotide strand to be synthesised in the same way as this is achieved with the conventional photolithographic technique. Using this procedure, ten thousands of 16 millimetres features with a spacing of one millimetre can be produced on an array measuring ten millimetres $\times$ 14 millimetres. Using such high-definition mirror systems, it is theoretically possible to construct arrays containing up to 2,000,000 different features. Since about 40 features are necessary to reliably cover one species of mRNA, this system should be able to deliver an array

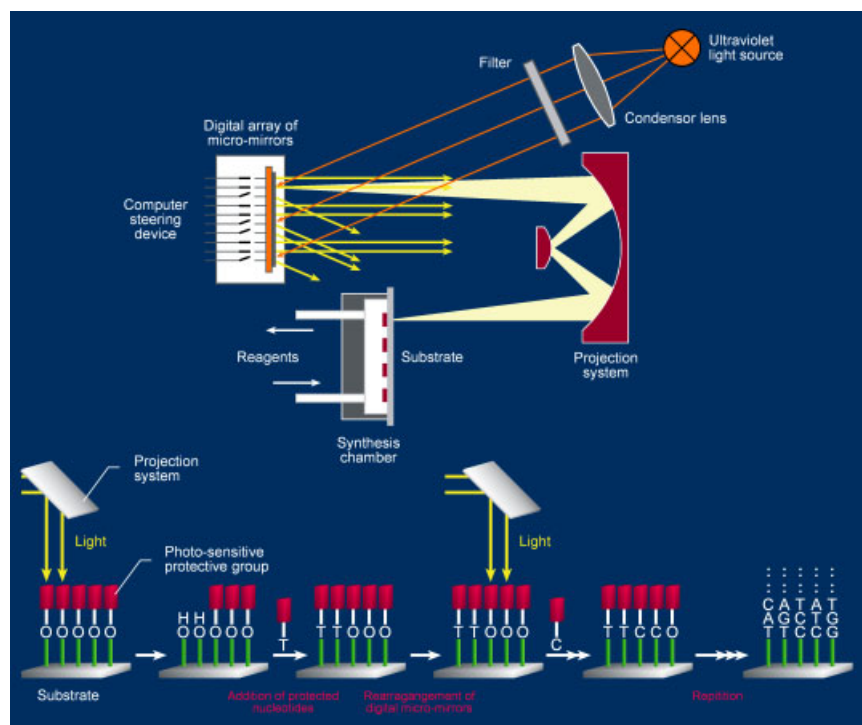


Figure 4.20 Photolithographic production with micro-mirrors. See text for details.

of approximately one square centimetre capable to monitoring the expression of 50,000 different genes. Furthermore, based on the limiting factor of the wavelength of ultraviolet light, it should be possible in the future to reduce the mirror edge-length in such projections systems to 0.35 millimetres.

4.2.1.3.9 Photoresist wet masking

In an attempt to overcome the low efficiency of 95 percent per synthesis step, Affymetrix has entered into a cooperation with the IBM Almaden Research Centre (San Jose, California, USA) to develop a photoresist technique (figure 4.21; McGall *et al.*, 1996) that will use conventional reagents for solid-phase synthesis of DNA while achieving a resolution of the same order of magnitude as optical photolithography (smaller than one millimetre).

With the single resist process (figure 4.21.A), a surface coated with N,N-bis(hydroxyethyl)aminopropyltriethoxysilane is first protected by a phosphoramidite derivative (4,4'-dimethoxytritylhexaethylenglycol(2-cyanoethyl-N,N-diisopropyl)phosphoramidite) and then completely covered by a photoresist layer similar to that used in semiconductor and microsystem technology (figure 4.22; compound SU-8). Just after exposure to ultraviolet light through a photolithographic mask, the resist is developed and removed from exposed areas. This generates a pattern

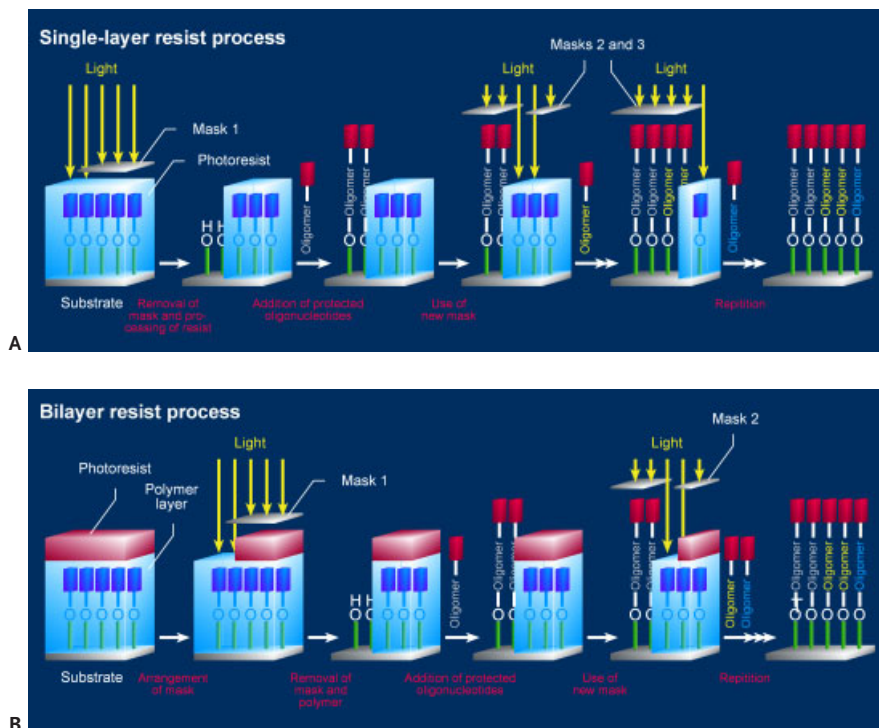


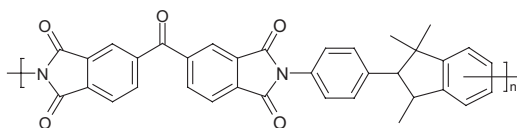
Figure 4.21 Photoresist technique. A) *Single layer resist procedure.* B) *Bilayer resist procedure* (see text for details).

of openings with an edge length of one millimetre, which enable the reagent to reach the surface and to remove the protective phosphoramidite groups. Preformed oligonucleotides containing photo-protective groups are then allowed to bind to the free hydroxyl groups on the array. A second round of light exposure, development and removal of the resist then makes other areas of the array available for reactions with further oligonucleotides. The use of a photoresist increases the contrast between exposed and protected areas, reducing the border effects that occur due to light scattering with conventional photolithographic techniques and increasing the efficiency of oligonucleotide synthesis.

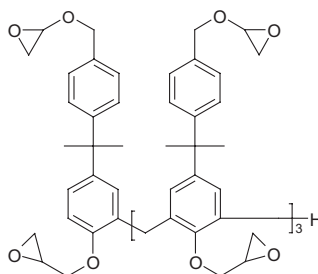
Nevertheless, the photoresist procedures allow only moderate rates of synthetic efficiency. One reason for this is that the reagent used to remove the protection groups reacts with phosphoramidite residues embedded within the resist. In addition, the chemicals used for development and removal of the resist interact with the underlying substrate and, thus, have a negative influence on subsequent chemical reactions.

The use of an inert polymer layer (compound XU-218 in figure 4.22) between the surface of the substrate and the resist layer can reduce these adverse interactions (*bilayer resist procedure*; figure 4.21.B). Optical resolution is impaired by the polymer

Figure 4.22 Structural formulas of the compounds used in the photoresist procedure. XU-218 can be used as protective polymer layer under the photoresist (SU-8).



XU-218



SU-8

layer but nevertheless total recoveries of about 90 percent are achieved. The use of photolithographically-produced masks can be avoided by using a liquid crystal display projector to remove the photosensitive resist layer (Garner, 1999).

4.2.1.3.10 μ WP – Micro wet printing

Micro wet printing (μ WP) was developed by Ermantraut *et al.* (1998) on the basis of the wet masking technique and is shown in figure 4.23. A printing cartridge with a silicon frame is used to position a mask on the surface of the array. The silicon frame is combined with a glass manifold to form a channel system that connects single inputs to the mask in the printing cartridge. The correct alignment of car-

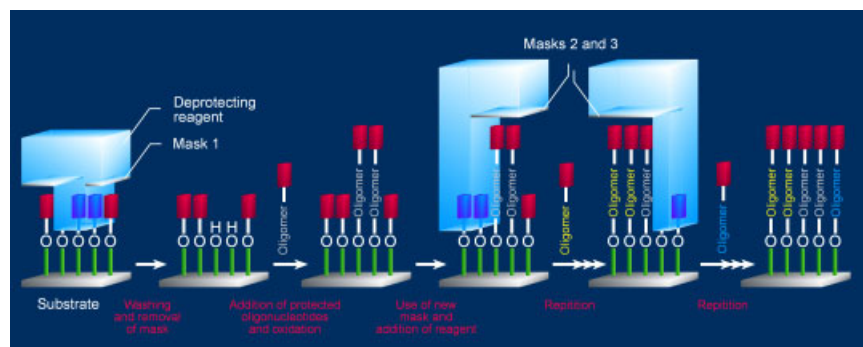


Figure 4.23 Principle of the micro wet printing (μ WP) method. See text for details.

tridge and array surface is guaranteed by complementary position markers on both surfaces and by mask alignment technology. After alignment, only those areas of the substrate located directly below the openings in the mask are accessible to reagent. Once reagents and a washing solution have been passed through the meander to the surface, the mask is removed and a new one is positioned for a further cycle of synthesis. By alternating addition of nucleotides and new masks, short oligonucleotides may be synthesised at defined positions on the substrate with a spot size of about one millimetre. This system can also be used to immobilise preformed oligonucleotides and other substances such as proteins and antibodies.

The phosphoramidite technique used for *in situ* synthesis of oligonucleotides with the μ WP procedure was first described by Robert Letsinger and developed further by Marvin Caruthers as a solid phase technique (Caruthers *et al.*, 1987; Green & Letsinger, 1975). It allows rapid automated oligonucleotide synthesis and is widely used in other applications.

4.2.1.4 Outlook

A major focus of current research is the search for better technologies for low-cost production of high-quality, high-density arrays. Improved production techniques will increase reproducibility with higher rates of synthetic efficiency as a result of the development of new photosensitive protection groups. Techniques for re-use of DNA arrays are also being explored. In the near-term, simple arrays will enter the medical diagnostic field, probably initially for mutation detection. In the medium-term, medical diagnostic applications will be developed based on fully integrated, high-performance and high-throughput systems for analysis of gene expression. Already today, use of DNA arrays have become an indispensable tool in academic and industrial research in the post-genomic era.

4.2.1.5 Transcriptome-based oligonucleotide design

The success of a microarray-based hybridisation experiment depends to a large degree on the accuracy in which the probes on the array represent the mRNAs with which they need to hybridise. In oligonucleotide-based microarray approaches (in contrast to cDNA-based microarray approaches), it is up to the designer of the array to properly select sequences that will guarantee successful and specific hybridisation.

A common approach to the problem of designing oligonucleotides is to consider expressed sequence tag (EST) clusters or mRNAs as representatives of the various genes, and to select the oligonucleotides from specific sequences within a EST cluster or mRNA.

However, EST sequence databases are known to contain several artefacts such as sequencing errors, intron contamination, and chimeric sequences. In addition, the annotation of ESTs in terms of their sense is often erroneous. At a more fundamental level, EST cluster databases do not represent information on the actual mRNA sequences transcribed from a given gene, and their splicing structure. As

such, a probe selected from an EST in such a cluster may turn out to hybridise only with certain splice variants of that gene.

Compugen (Tel Aviv, Israel), for example, developed a computational probe design methodology that alleviates those problems by initially transforming the EST and genome sequence data into a comprehensive collection of mRNA transcript sequences, while taking into account the various phenomena mentioned above. The resulting sequence collection contains mRNA sequences along with their splice variant forms, which forms a more reliable basis for subsequent oligo selection.

Once the transcripts are available, representative oligonucleotides can be selected from regions of the transcripts, which are either shared with as many splice variants as possible, or are specific to each mRNA of each gene. Which one is preferred is determined by the required application. The oligonucleotide selection system chooses a representative sequence of the required length (typically 60 to 70 base pairs) by considering various criteria for each candidate sequence, including:

- (i) Minimal distance from the 3'-end of the gene or gene cluster,
- (ii) Minimal homology to all other genes,
- (iii) Avoiding polymorphic sites, low-quality and repeat regions,
- (iv) Homogenous GC content and melting temperature across the whole library,
- (v) Simple secondary structure.

An application of this approach to splice variant detection in rat was described by Hu *et al.* (2001). More information on synthesised oligonucleotides sets designed by this methodology can be obtained from the Compugen website available at <http://www.labonweb.com/chips/>.

At present, there is a great interest in mathematical algorithms and experimental solutions to select oligonucleotides for microarray-based mRNA expression analyses and several different algorithms are under development or have been published. A complete description of the approaches presented so far goes far beyond the scope of this contribution. The interested reader is therefore referred to the primary literature (*e.g.*, Broude *et al.*, 2001; Kampke *et al.*, 2001; Li & Stormo, 2001; Kane *et al.*, 2000; Raddatz *et al.*, 2001; Relógio *et al.*, 2002; Rouillard *et al.*, 2002; Uhlmann, 2001) and to information or software tools on the world-wide web: for example, Array Designer Software (TeleChem International, Inc./arrayit.com, Sunnyvale, California, USA; <http://arrayit.com/>); BioSap, Blast Integrated Oligonucleotide Selection Accelerator Package (The Institute for Systems Biology, Seattle, Washington, USA; <http://biosap.sourceforge.net/>); DEODAS, DEgenerate Oligo Nucleotide Design and Analysis System (The Institute for Systems Biology, Seattle, Washington, USA; <http://deodas.sourceforge.net/>); Featurama/ProbePicker (The Institute for Systems Biology, Seattle, Washington, USA; <http://probepicker.sourceforge.net/>); OligoArray (Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan, USA; <http://berry.engin.umich.edu/oligoarray/>); Oligonucleotide Probe Sequence Design Software for Genetic Analysis (Olympus Optical Co. Europe, Hamburg, Germany; <http://www.olympus.co.jp/>); Oligos4Array (MWG-Biotech GmbH, Ebersberg, Germany; <http://www.mwg-biotech.com/>).

4.2.1.6 State-of-the-art of nucleic acids labelling and detection

The use of microprobe arrays depends to a significant degree on sample preparation. Here the methodology applied to make the molecular target detectable on the array is crucial. Labelling the target with a detectable marker has become a standard approach in microarray application since the process of attaching a label to the target inherently bears the potential to increase the specificity of the assay. Nevertheless, numerous detection techniques that do not rely on target labelling with a detectable marker have been devised in the literature. These methods include among others surface plasmon resonance imaging (Brockman *et al.*, 1999), direct electrical measurements (Heller & Tu, 1993), surface acoustic wave (SAW) sensors (Barié *et al.*, 1998), quartz crystal microbalance dissipation and various interferometric methods. Particularly, the determination of molecular load on surfaces by reflectometric interference spectroscopy (Gauglitz *et al.*, 1993) appears to be a promising method for label-free detection of target binding on probe arrays due to its robustness in regard to environmental parameters like temperature and buffer effects and the comparably small effort required to establish a parallel read-out of multiple probe/target interactions on a single surface. All these methods that are independent of the physical nature of the methodology have the potential advantage of detecting molecular targets that have not been altered by a labelling procedure. This also implies that no additional labelling is required to make a target detectable. In practice, in complex samples targets are difficult to detect explicitly. The physical variable that is measured by all of these methods (mass, conductivity) is an inherent property of all biomolecules. Thus, additional biomolecules that are usually present in real-life samples contribute significantly to increased background and false-positive signals. This is probably the reason why these methods have not been more popular up to now. Target amplification monitored by detection of the increasing mass at the sensor surface might allow more widespread use of these technologies in diagnostics applications, as assays based on specific amplification of a limited number of targets within complex samples of low concentration are becoming more important in molecular diagnostics.

At the present time, however, the method of choice for detection of molecular targets in microarray experiments is still by specifically labelling of the target. The predominant detection scheme in microarray applications is detection of fluorescence signals emitted from dye labels bound to the target. An alternative is the detection of mass labels by mass spectroscopy (Stomakhin *et al.*, 2000) and the detection of absorbency changes based on the use of appropriate labels with a high specific absorbency.

Labelling and gel-based separation of labelled products are avoided by the use of matrix assisted laser desorption/ionisation time-of-flight mass spectroscopy (MALDI-TOF/MS) as a detector and a separator. Products of enzymatic degradation of oligonucleotides with single strand-specific exonucleases (Pieles *et al.*, 1993) and products of sequencing termination reactions (Köster *et al.*, 1996) have been detected by MALDI-TOF/MS. The method provides a means for *de novo* sequence determination. The disadvantages of this method are the very high equipment prices and limited miniaturisation potential. The method is inherently serial, but exces-

Table 4.1 Methods to label nucleic acids.

Method	Advantage	Disadvantage
Enzymatic incorporation/ primer labelling	• Process easily performed • Reagents readily available • Labelling is carried out in combination with amplification	• Indirect target labelling
Chemical modification	• Direct target labelling	• Labelling and amplification separate • Hybridisation may be influenced by modification
Enzymatic modification/ tailing	• Direct target labelling • Hybridisation is not influenced by base modification	• Labelling and amplification separate • Target length changes

sive process automation allows thousands of analysis to be performed per day. However, with MALDI-TOF/MS, the labelling problem is replaced by difficulties regarding sample preparation with the result that only in a few genotyping applications MALDI-TOF/MS is more convenient than standard fluorescence analysis.

The principal of target labelling is independent of the physical method used to detect the label. In nucleic acids labelling, the label can be added directly by reaction of chemically active groups with the target nucleic acids, by incorporation of labelled nucleotides/by means of labelled primers or by enzymatic modification of the molecular target itself (see table 4.1).

Extensive reviews on this subject have been published over the last few years that provide a good description of this (*e.g.*, Freeman *et al.*, 2000; Hegde *et al.*, 2000).

The predominant label in today's microarray work is without any doubt fluorescent dye. Dendrimers of fluorescent dyes may also be used to increase sensitivity (Shchepinov *et al.*, 1997). Because a long duration of fluorescence is an important factor for result consistency, inorganic fluorescent labels have recently been used that do not show the bleaching characteristic of organic dyes. Examples are the use of up-converting phosphorus (van de Rijke *et al.*, 2001) and quantum dots (Sun *et al.*, 2001).

4.2.1.6.1 Fluorescent labels

The simplest and most widely used approach for labelling of nucleic acid molecules and detection of hybridisation using DNA microarrays is labelling of the sample nucleic acids with a fluorescent molecule (figures 4.1.A and 4.1.B). Several techniques for this purpose have been described including random primed labelling, cDNA synthesis, PCR amplification or *in vitro* transcription (An *et al.*, 1992; Bürgi *et al.*, 1992; Chenal & Griffais, 1994; Emanuel, 1991; Gissing *et al.*, 1997; Hacia *et al.*, 1998; Hölte & Kessler, 1990; Holmstrom *et al.*, 1993; Kath *et al.*, 1993; Kricka, 1999 and 2002; Lanzillo, 1990; McCreery & Helentjaris, 1994; Proudnikov & Mirzabekov, 1996; Reischl *et al.*, 1993 and 1994; Rosemeyer *et al.*, 1995;

Rost *et al.*, 1995; Schmitz *et al.*, 1991; Soldo *et al.*, 1992) and strategies using DNA-binding dyes such as Hoechst 33258 and SYBR[®] green II (Battaglia *et al.*, 2000; Drobyshev *et al.*, 1999).

The principle of fluorescence-based detection of hybridisation using DNA microarrays is simple. Labelled nucleic acids are hybridised to the DNA microarray at temperature and salt conditions that give acceptably strong signals for the desired hybridisation products and much weaker signals for mismatches. Unbound target DNA is then washed away, and the array is ready to be scanned.

Scanning a fluorescently-labelled DNA array is conceptually quite simple. A light source excites the labelled samples and a detector system measures and records the emitted fluorescence. Based on the precise nature of the array, the instrumentation requirements differ. Most image capture instruments use a scanning detector similar to line-scanning detector systems for DNA sequencing instruments. Other instruments make use of a charge-coupled device camera. Another possibility is the use of simple high-resolution microscopes.

Two kinds of different labelling strategies are widely used today: the Cy3/Cy5 system and the streptavidin/phycoerythrin system. The latter system is used for the Affymetrix GeneChip[®] technology (figure 4.1.B) whereas the Cy3/Cy5 system (figure 4.1.A) is widely used as a label for self-made and commercial cDNA microarrays (Duggan *et al.*, 1999).

Typically, labelling of cDNA with Cy3 or Cy5 dyes is performed by reverse transcription using an oligo(dT) primer. As fluorescent labels, Cy3-labelled dUTP and Cy5-labelled dUTP are paired, as they have relatively high incorporation efficiencies with reverse transcriptase, good photostability and yield, and are widely separated in their excitation and emission spectra. This allows highly discriminating optical filtration and simultaneous detection of two different samples on a single array.

A limitation of this labelling technique is the large amount of RNA required per hybridisation. For adequate fluorescence, the total RNA required per sample and array is about 50 to 200 micrograms (two to five micrograms of poly(A) mRNA are required). Even with this amount of RNA, signals of rare transcripts are at the lower limit of fluorescence detection, and are often difficult to distinguish from assay noise (figure 4.24).

Several modifications have been proposed to improve signal detection of limited RNA. Methods that produce multiple copies of mRNA using highly efficient phage RNA polymerases have been developed (Phillips & Eberwine, 1996). The streptavidin/phycoerythrin system mentioned above is such an approach in which biotin-labelled cRNA is made directly from a cDNA having a T7 RNA polymerase promoter site at one end *via in vitro* transcription (Lockhart *et al.*, 1996). Detection of the biotin-labelled molecules is performed using streptavidin-bound phycoerythrin. Post-hybridisation amplification is possible using a further round of incubation with biotin and streptavidin/phycoerythrin conjugates.

Fluorescence detection has become the most widely used detection technique for DNA microarrays. The main reasons for this may be that fluorescent labelling is technically simple and reliable, the availability of a range of fluorescent dyes (figure 4.25), and a choice of commercial detection systems.

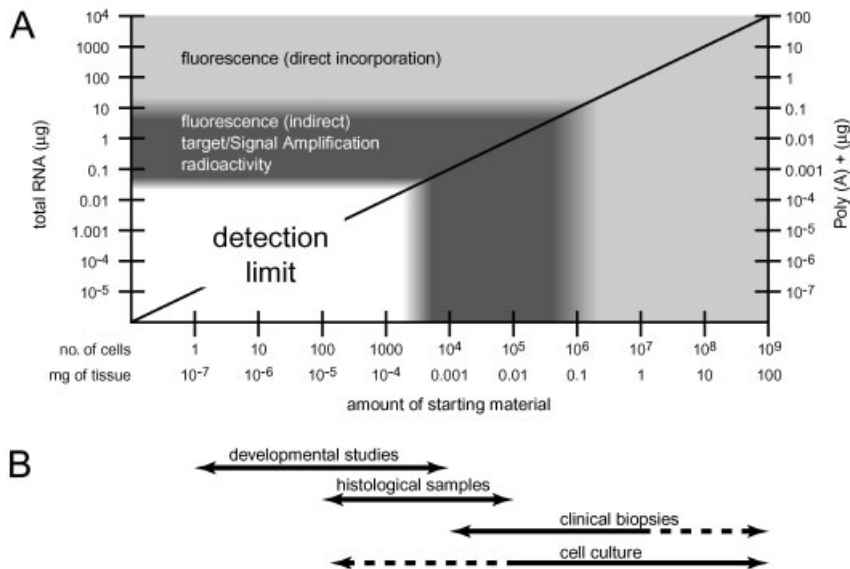


Figure 4.24 Detection schemes and applications of cDNA microarrays. Reproducible quantitative changes in gene expression can be detected using several schemes with varying limits of detection **A**). Direct incorporation of fluorescent nucleotides into a cDNA sample can be used to examine expression profiles generated from ten micrograms or more of total RNA starting material. Indirect fluorescence, target and signal amplification, and radioactivity may also be used to detect expression profiles from as little as 50 nanograms of total RNA. This detection limit allows for the investigation of expression profiles from numerous biological sources including cell culture, clinical biopsies (including autopsy material) and histological samples **B**). Technical improvements are likely to lower the lower limit for the amount of starting material needed for an expression profile to less than 50 nanograms of total RNA. The figure is reproduced from Duggan *et al.* (1999) with permission from *Nature Genetics*.

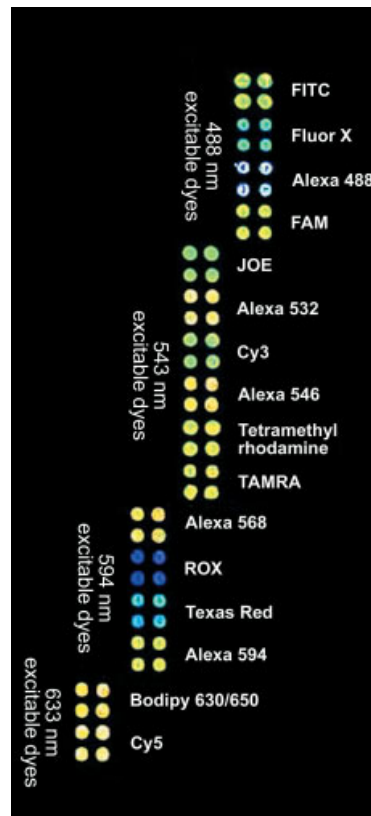


Figure 4.25 Examples of some conventional fluorescent dyes used for nucleic acid labelling and detection. The figure is reproduced from Sinclair (1999) with permission from *The Scientist*.

4.2.1.6.2 Radioactive labelling

Radioactive labelling of nucleic acids is the oldest technique for detecting and quantifying hybridisation events of single-stranded nucleic acid probes with their corresponding single-stranded targets. This kind of label has been widely used in Southern blotting, Northern blotting, DNA filter arrays and *in situ* hybridisation. Usually, DNA or RNA probes are used that are labelled with ^{32}P , ^{33}P or ^{35}S to hybridise to the target of interest.

Several different techniques have been developed to introduce radioactivity into the nucleic acid probes such as random primed labelling or PCR amplification using radioactively-labelled oligonucleotides as PCR primers (An *et al.*, 1992; Bürgi *et al.*, 1992; Chenal & Griffais, 1994; Emanuel, 1991; Gissing *et al.*, 1997; Hölte & Kessler, 1990; Holmstrom *et al.*, 1993; Kath *et al.*, 1993; Kricka, 1999 and 2002; Lanzillo, 1990; McCreery & Helentjaris, 1994; Reischl *et al.*, 1993 and 1994; Rosemeyer *et al.*, 1995; Rost *et al.*, 1995; Schmitz *et al.*, 1991; Soldo *et al.*, 1992). Following hybridisation, the auto-radiographic signal is detected using simple X-ray films, photostimulable storage phosphorus imaging plates (Johnston *et al.*, 1990), or more recently, real-time auto-radiography imaging systems such as the Micro ImagerTM and the Beta ImagerTM (Zinsser Analytic GmbH, Frankfurt, Germany; Biospace Mesures, Paris, France) together with the use of specialised software tools for quantification (*e.g.*, Pelizzari *et al.*, 2000). Figure 4.26 shows an example of a DNA filter array that was labelled with a ^{32}P cDNA sample.

Despite the introduction of a wide variety of methods using fluorescence or chemiluminescence, detection of radioactively-labelled nucleic acids is still the most sensitive labelling technique available. In addition, the linear dynamic range of detection using a radioactive label is ten times larger than that of non-radioactive de-

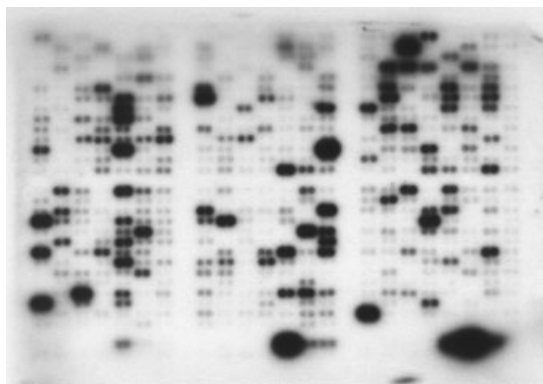


Figure 4.26 DNA filter array hybridised with a ^{32}P -labelled cDNA sample followed by detection using X-ray film. Eukaryotic poly(A) mRNA was reverse transcribed and radioactively labelled using ^{32}P -labelled nucleotides. The labelled cDNA was hybridised to a commercial DNA

filter array (BD Biosciences Clontech, Palo Alto, California, USA) and signals were detected by auto-radiography using conventional X-ray film. The figure was kindly provided by Dr. Bodo Levkau (Institute of Arteriosclerosis Research at the University of Münster, Münster, Germany).

tection systems, in particular if real-time detection systems are used to analyse radioactive signals. Differential screening using probes labelled with two different radioisotopes is also possible providing radio-elements of significantly different energies such as tritium and ^{35}S or tritium and ^{33}P are used (Salin *et al.*, 2002).

Detection procedures using radioactive labelling have two major disadvantages: (i) The use of radioactive labelling is limited mainly to low and medium-density filter systems. Because of the strong lateral radiation of the radioactive label, its applicability for high-density glass microarrays is limited. However, Whitney & Becker (2001) showed that it is possible to circumvent this limitation using photo-stimulable storage phosphorus imaging plates. (ii) The use of radioactivity, however, renders this detection system unsuitable for routine diagnostic laboratory use.

Despite these drawbacks, radioactive labelling offers a fast direct detection system and is not expensive. Simple X-ray films can be used for detection and the equipment required for enhancing X-ray films is available in most clinical institutions.

4.2.1.6.3 Chemiluminescent detection

Chemiluminescent detection is widely used in bioanalytics for analysis and quantification of biomolecules using methods such as *in situ* hybridisation, immunohistochemistry, flow-injection analysis, column-liquid chromatography, capillary-electrophoretic separation, and Southern, Northern, and Western blotting (Pasini *et al.*, 1998; Roda *et al.*, 2000).

Several different systems for chemiluminescence detection have been developed such as the digoxigenin and the biotin labelling and detection system (Emanuel, 1991). In digoxigenin labelling, the sample is labelled by incorporation of digoxigenin-labelled nucleotides using random primed labelling (McCreery & Helentjaris, 1994; Rost *et al.*, 1995), nick translation (Kath *et al.*, 1993), cDNA synthesis (Soldo *et al.*, 1992), PCR amplification (An *et al.*, 1992; Chenal & Griffais, 1994; Emanuel, 1991; Gissing *et al.*, 1997; Holmstrom *et al.*, 1993; Lanzillo, 1990; Reischl *et al.*, 1993 and 1994), *in vitro* transcription (Höltke & Kessler, 1990), 3'-end labelling (Rosemeyer *et al.*, 1995), 3'-tailing (Schmitz *et al.*, 1991), or chemical synthesis (Bürgi *et al.*, 1992). Following hybridisation of the digoxigenin-labelled sample to the probe, the DNA duplex is detected by incubation of a digoxigenin-specific antibody that is coupled with an enzyme (*e.g.*, the widely used alkaline phosphatases) that catalyses the reaction of a substrate to a chemiluminescent product (see figure 4.27). In biotin labelling, the sample is labelled by incorporation of biotin-labelled nucleotides using, in principle, the methods mentioned above. Instead of a specific antibody, streptavidin or avidin is used to detect the biotin-labelled target. In this method, similar enzymes coupled to the (strept)avidin molecule and substrates are used.

For a complete description of enzymes and substrates for chemiluminescence detection, the reader is referred to the specialist literature (Durrant, 1994; Durrant & Fowler, 1994; Rongen *et al.*, 1994). For detailed information and a comprehensive list of references concerning the digoxigenin labelling and detection system, please see also the Roche Diagnostics (Mannheim, Germany) website at <http://>

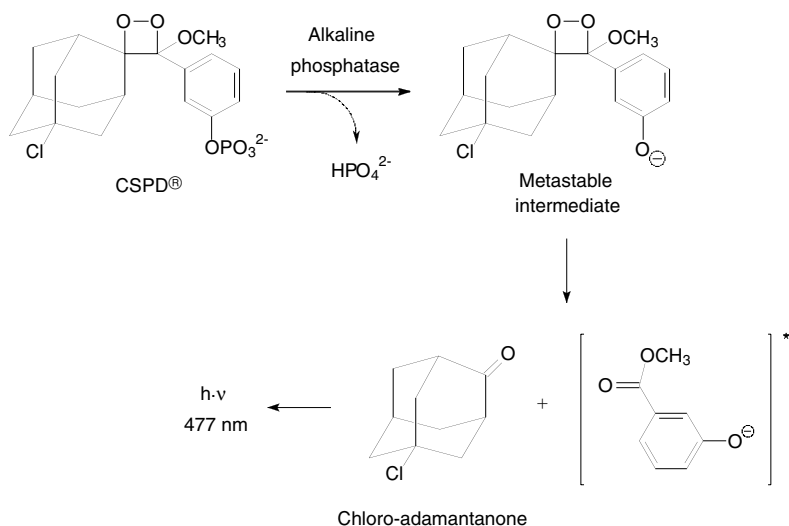


Figure 4.27 Enzymatic dephosphorylation of the chemiluminescence substrate CSPD®.

CSPD® (5-chloroadamantane-2-spiro-3'-[4'-methoxy-4'-(3-phosphonooxyphenyl)-dioxetane]; [3-[5-chloro-4'-methoxyspiro[adamantane-2,3'-dioxetane]-4'-yl]-phenylphosphate or

3-(4-methoxyspiro-{1,2-dioxetane-3,2'-tricyclo[3.3.1.1^{3,7}]decane-4-yl)-phenylphosphate)) is dephosphorylated by an alkaline phosphatase. The resulting intermediate decays to the end product by emitting light of a wavelength of 477 nanometres (Martin *et al.*, 1991).

www.roche-applied-science.com/dig/literature.htm. Further information and protocols are available in the "DIG Application Guide for Filter Hybridization" published by Roche Diagnostics GmbH that is also available in the internet at http://www.roche-applied-science.com/prodinfo_fst.htm?prod_inf/manuals/dig_man/di_g_toc.htm.

The success of chemiluminescence in routine Southern and Northern blotting paved the way for the application of this technique to the analysis of mRNA expression on DNA filter arrays. Rajeevan *et al.* (1999) successfully applied chemiluminescence detection using digoxigenin-labelled cDNA probes to screen for differences in the mRNA expression in human placenta, keratinocytes and peripheral blood lymphocytes derived from patients with chronic idiopathic fatiguing illnesses and corresponding controls (see figure 4.28). Rajeevan *et al.* estimated the detection limit for low-abundance messages to be one to fifteen transcripts per cell, a sensitivity rivalling that reported for microarray formats and exceeding that reported for auto-radiographic filter arrays.

Cheek *et al.* (2001) used the three-dimensional ordered micro-channels of a flow-through DNA array (Benoit *et al.*, 2001) to increase both the sensitivity and spatial resolution required for quantitative measurements on microarrays using chemiluminescent detection. Enzyme-catalysed reactions for the detection of hybridisations on the micro-channel glass were imaged using a charge-coupled device camera (figure 4.29). The relative standard deviation of signal intensities across an array

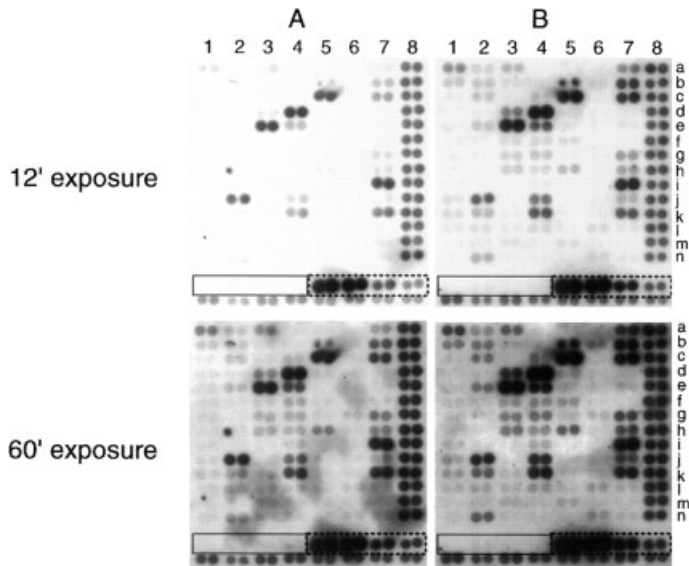


Figure 4.28 Effect of probe concentration and exposure time on chemiluminescent detection of gene expression on filter arrays. Only quadrant F of the six quadrants of an Atlas Human Expression Array I (BD Bioscience Clontech, Palo, Alto, California, USA) is shown for each of the four results. Numbers 1 to 7 and letters a to n are those used by the manufacturer to designate the coordinates of the duplicate spots of target cDNA, a total of 98 different cDNAs. Area within solid box includes the negative control target cDNAs. Area within the dashed box in-

cludes the positive control target cDNAs. The row under the boxes and the doublets in column 8 include genomic DNA for positioning. Digoxigenin-labelled cDNA was prepared from human placental poly(A)⁺ mRNA. The two images beneath each letter are two exposures (twelve *versus* 60 minutes) from the same filter that was hybridised with one third of the total synthesis reaction **A**) *versus* the entire synthesis reaction **B**) as probe. The figure was reproduced from Rajeevan *et al.* (1999) with permission from *Journal of Histochemistry and Cytochemistry*.

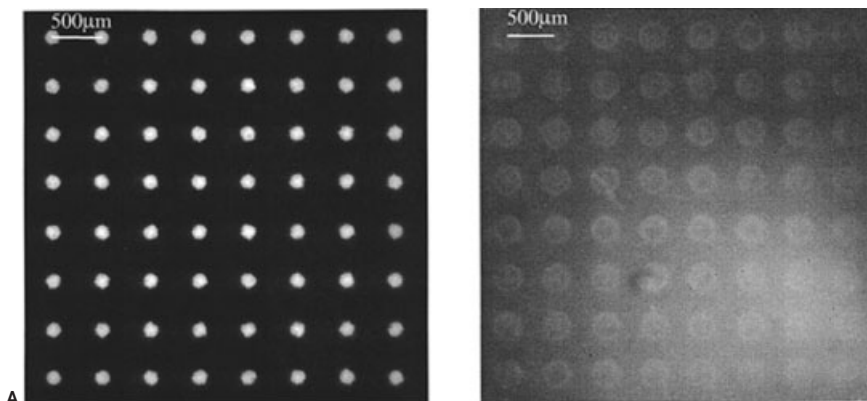


Figure 4.29 (s. next page)

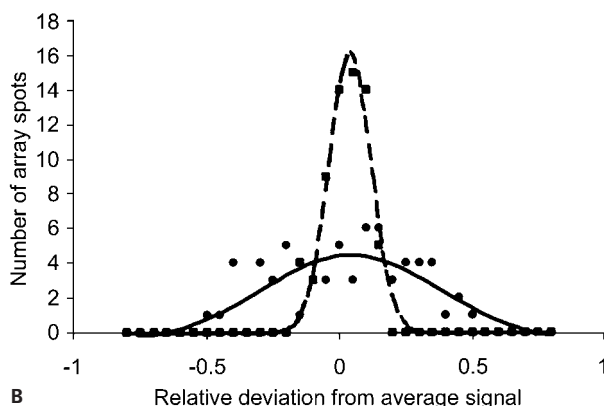


Figure 4.29 Chemiluminescence detection of hybridization assays. **A)** *Left:* Micro-channel glass microarray. *Right:* flat glass microarrays. The micro-channel glass array yielded signal intensities 96 times greater than the flat glass array, consistent with the about 100-fold increase in available surface area for probe immobilisation. The relative standard deviation of spot intensities for the micro-channel glass and

flat glass arrays were 8.1 and 25.3 percent, respectively. The contrast for each image was adjusted for optimal viewing. **B)** Histogram of the chemiluminescence signal uniformity on micro-channel glass arrays (■) and flat glass arrays (●). The lines represent a Gaussian fit to the data. The figure was reproduced from Cheek *et al.* (2001) with permission from *Analytical Chemistry*.

of 64 spots was 8.1 percent and a detection limit of 250 attomole of the target with a linear dynamic range of three orders of magnitude was obtained for a three hour assay.

A further use of chemiluminescent detection for array-based mRNA expression profiling was recently described by Fahnert *et al.* (2002). The authors assessed gene expression data obtained using chemiluminescent detection and employed empirical criteria to identify biologically meaningful expression ratios. These empirical criteria were validated using a knowledge-based system. The authors compared the experience-based and knowledge-based assessment of the expression data from the chemiluminescent and radioactive detection of several experiments with published data in order to evaluate their procedure.

In summary, chemiluminescence detection offers several advantages compared to other labelling and detection techniques: (i) ease of use, (ii) sufficient sensitivity and specificity, (iii) no irradiation or contamination with dangerous hazards, (iv) high dynamic/linear range of detection, (v) low equipment cost (detection with X-ray films or camera systems), and (vi) reusability of the arrays.

The disadvantages associated with chemiluminescent detection are (i) lower sensitivity compared to radioactive labelling, (ii) higher background signals, and (iii) lower specificity. However, although chemiluminescent detection is seldom used for microarray technology, it is a promising alternative to other labelling strategies such as fluorescence or auto-radiography.

4.2.1.6.4 Gold/silver staining

The need for ease of use, reproducibility and affordable detection equipment has led to the adaptation of well established gold labelling/silver enhancement techniques to microarray analysis. Immunogold/silver-staining (IGSS) is one of the most important and widely used methods in the fields of protein chemistry and immunohistology. The process of site-specific silver deposition is best known from photographic techniques. When black and white film is exposed to light, silver ions in a light sensitive layer of silver bromide on the surface of the film are reduced to atomic silver. When the film is developed, the remaining silver bromide is diluted and washed away. The silver ions provided by the development solution are taken up by the surface of the remaining silver particles. Under the influence of a reducing agent in the developer solution, these silver ions are converted into silver atoms precipitating around the initial silver particle. In this way, the silver particle achieves a size that is visible under the light microscope.

As it is described in this section, the process of silver deposition on the surface of gold particles can also be used for the detection of hybridisation events on DNA-arrays.

4.2.1.6.4.1 Principles and basics

Gold crystals and crystals of several other metals or metal sulphides and selenides are able to bind silver ions and reducing agents on their surfaces. After binding, they catalyse the transport of electrons from the reducing agent to the silver ions, thus converting the silver ions into silver atoms. The atomic silver precipitates around the initial crystallisation nucleus and forms a black light-absorbing particle. Hydroquinone, for example, may be used as a reducing agent. Silver acetate, silver nitrate and silver tartrate are suitable silver salts.

As it is known from the use of IGSS in immunohistology, several factors affect the sensitivity and the rate of the reaction. The most important are the size of the gold particles used for labelling of the target, the type of silver salt and the choice of a reducing agent. Also important is the use of a suitable reaction moderator, such as Arabic gum. A good survey of general procedures in immunohistology and protein chemistry can be found in Hayat (1995).

4.2.1.6.4.2 Silver staining on DNA arrays

Two principal systems for detection of hybridisation events on an array may be distinguished (Kessler, 2000): direct and indirect labelling and detection systems. Both differ in the number of reaction steps that are necessary for reaction detection and sample preparation. In direct systems, the target is covalently linked with a signal-generating reporter group. In indirect systems (figure 4.30), there is no direct link between target and reporter group. The reporter group is attached to a mediator molecule such as streptavidin that can bind to a special modification such as biotin within the target. In principle, both direct and indirect labelling techniques can be used for silver-based detection on DNA microarrays.

Direct labelling methods require the incorporation of gold particles into the target molecule. At present, no protocols are available for direct labelling of unmodi-

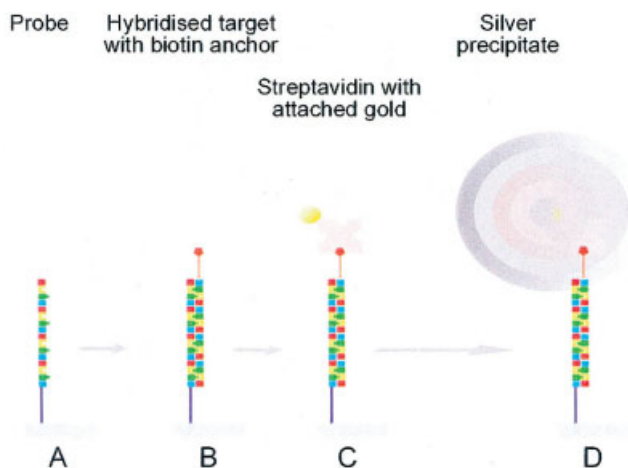


Figure 4.30 Schematic representation of an indirect labelling system. **A)** Covalently to a surface linked probe. **B)** A biotin-labelled target is hybridised to the probe on the array. Unbound target is washed away. **C)** A strepta-

vidin-labelled target with colloidal gold is coupled to the biotin anchor of the hybridised target. **D)** During the final silver staining, silver precipitates around the gold label.

fied DNA or RNA with gold particles of a specific size. It is therefore necessary to introduce a reactive group to the target molecule that in turn can be linked to activated gold particles. This can be a primary amine, which can easily be introduced by means of *in vitro* transcription or cDNA synthesis using aminoallyl-modified nucleotides. In a further step, amino-modified DNA can be coupled to Mono-Sulfo-NHS-Nanogold® (Hainfeld & Powell, 1997 and 2000). A gold label can also be introduced in DNA during PCR using, for example, gold-labelled primers.

Direct detection systems are more straightforward. After labelling the target, the array experiment comprises only two steps: (i) Hybridisation of the target carrying a reporter group to the array, and (ii) signal generation *via* silver enhancing the reporter group.

One disadvantage of labelling the target directly is that the gold label may influence the hybridisation characteristics of the target. This makes it difficult to find the optimal stringent hybridisation conditions that are necessary for such applications as the detection of single nucleotide polymorphisms (SNPs). Nevertheless, direct labelling of PCR or cDNA using gold-labelled primers might be an interesting alternative for specialised applications.

In contrast to direct labelling and detection systems, indirect systems for detecting hybridisation events require an additional step in the array treatment. The assay is carried out after the following scheme: (i) Hybridisation of the target carrying a special modification to the array, (ii) specific but non-covalent interaction between the modification of the target and a mediator molecule carrying the reporter group, and (iii) signal generation *via* silver enhancing the reporter group.

Indirect labelling offers two interesting advantages. First, there are optimised protocols for efficiently labelling DNA or RNA targets with a linker molecule such as biotin or digoxigenin. A selection of binding pairs that can be used for detection of targets on a DNA microarray is shown in table 4.2. Second, because of the universal usability of the detection components, indirect detection systems are more versatile for use on DNA microarrays. Different kinds of targets such as DNA and RNA can be detected using the same detection components. The same target can be examined in parallel experiments using both silver-based and fluorescence-based detection.

A special case of an indirect labelling system is the use of labelled oligonucleotides as mediator molecules (detection probes) in a sandwich type hybridisation assay (Taton *et al.*, 2000). Here, the label is introduced when the detection probes specifically bind to the target molecule in a position different from the one that is complementary to the probe on the array.

Depending on whether the detection oligonucleotide is labelled directly with gold or carries a coupling group such as biotin, the reaction will require three or four steps: (i) Hybridisation of the target to the probes on the array, (ii) hybridisation of a target-specific detection probe to the array-bound target, (iii) (specific but non-covalent interaction between the modification of the detection probe and a mediator molecule carrying the reporter group), and (iv) silver enhancement of the gold label.

This kind of assay is particularly useful for detection of eukaryotic mRNA or poly(A) mRNA. Here, a gold-labelled oligo(dT)₂₀ primer can be hybridised to the poly(A) tail of the hybridised mRNAs. The complex can be directly silver-enhanced making an additional modification of the target unnecessary. In experiments analysing complex samples which lack a common sequence, the situation is more complicated. Either a specific gold-labelled mediator oligonucleotide must be included for every target to be analysed or the targets must be tailed *in vitro* with a common sequence by ligation, polyadenylation, *etc.*

Table 4.2 Selection of indirect labelling and detection systems.

Modification of the target	Labelling procedures (examples)	Labelling substances (examples)	Detection mode
Biotin	PCR, RT-PCR	• Biotin-16-dUTP • 5'-biotinylated primers	• Streptavidin/gold conjugate • Anti-biotin/digoxigenin-antibody gold conjugate
	Chemical modification of DNA or RNA	• Biotin/digoxigenin-ChemLink™	
Digoxigenin	PCR, RT-PCR	• Digoxigenin-11-dUTP • Digoxigenin-16-dATP	• Anti-digoxigenin/antibody gold conjugate
Fluorescein	PCR, RT-PCR	• Fluorescein-12-dUTP • Fluorescein-15-dATP	• Anti-fluorescein/antibody gold conjugate
Bromodeoxyuridine (BrdU)	<i>In vivo</i> modification	• Bromodeoxyuridine	• Anti-BrdU/monoclonal antibody gold conjugate

4.2.1.6.4.3 Advantages of silver staining of DNA arrays

In comparison with the standard detection methods for hybridisation events on DNA arrays, silver staining offers some decisive advantages. The wide dynamic range might be the most interesting of all. By measuring the intensities of the spots on the array at different times during silver staining procedure, it is possible to detect targets that differ in concentration up to five orders of magnitude in the same experiment.

The sensitivity of the silver staining reaction is at least as good as that reached with fluorescence detection. As mentioned above, it is decisively influenced by the composition of the silver staining solution and the components of the detection system used.

For detection of silver-stained arrays, hardware of low complexity is adequate. Silver-stained slides, for example, can be analysed using a conventional flatbed scanner with an optical resolution of at least 1,200 dots per inch. Highly integrated or synthesised DNA arrays can be analysed using a simple transmission light microscope in combination with a charge-coupled device (CCD) camera. The low technical requirements make silver staining of DNA arrays a suitable technique for use in mobile detection systems.

The rate at which the gold particles are enhanced by silver staining may be used to draw conclusions about the hybridisation behaviour and the concentration of the hybridised targets. This is illustrated by the experimental data shown in figure 4.31.

Figure 4.31.A shows the results of an experiment analysing mutations in the human *cyp2D6* gene, a member of the P450 gene family IID. Probes complementary to the wild-type and the mutant sequence of several mutations were spotted in

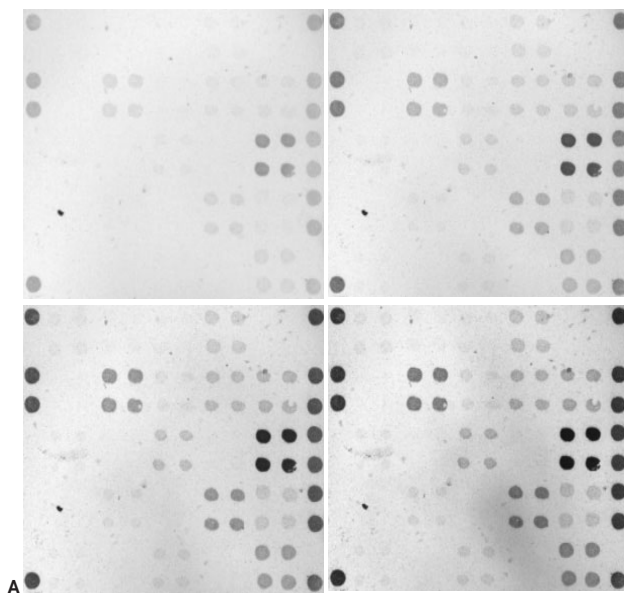


Figure 4.31 (s. next page)

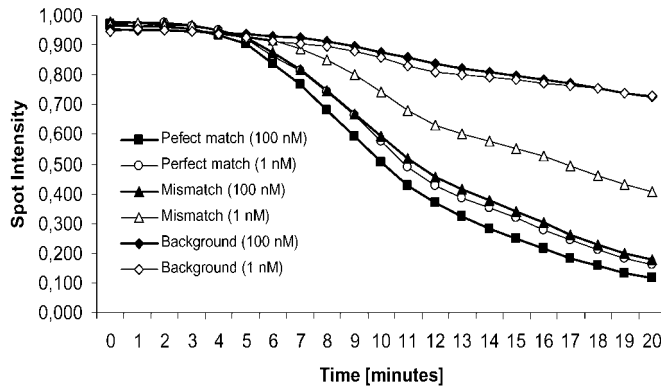


Figure 4.31 Example for silver staining of a DNA array. A) Selection of images taken at different times during silver staining of a hybridised DNA array. The hybridisation was performed using a *cyp2D6* ArrayTube® (Clontia® Chip Technologies, Inc., Jena, Germany; see <http://www.clontia.com> for details). An aliquot of a biotinylated asymmetric multiplex PCR was hybridised against a 100-probe array containing probes for the analysis of twelve mutations of the human *cyp2D6* (cytochrome P450 subfamily IID member 6) gene. After hybridisation and subsequent washing steps, the last washing buffer was removed and a buffer containing gold-labelled streptavidin particles (EM STP5, British Biocell International, Cardiff, Wales, United Kingdom) at a final concentration of 250 picograms per microlitres was added.

After incubation followed by three washing steps using different stringency conditions, a silver staining solution (SEKL15, British Biocell International) prepared corresponding to product information was given into the ArrayTube®. The silver staining was performed at 25 °C. The image series was taken using the ArrayTube® reader (Clontia® Chip Technologies, Inc.). **B)** Analysis of the images obtained from hybridisations of an oligonucleotide at two different concentrations to a DNA array containing a perfect match and a single mismatch sequence for the hybridised oligonucleotide. The hybridisation was performed under the same conditions described above. The obtained image series were analysed using the imaging tool IconoClust® (Clontia® Chip Technologies, Inc.).

fourfold redundancy. The relation between the signal obtained on the mutant and wild-type probe is an indicator of the genotype.

The four images shown were taken from a time series during the silver staining of a DNA array. Some spots are visible only at the end of the development process. At this point, however, the strongest spots have already achieved saturation. The dynamic detection, that is the detection over the complete period of the silver staining process, aids interpretation of the hybridisation results. This dynamic detection procedure is especially advantageous when used for quantitative analyses.

Figure 4.31.B shows an example in which an oligonucleotide was hybridised at two concentrations to an array containing a perfect match and a mismatch sequence for this oligonucleotide. This example shows, that only limited discrimination would have been possible between the different concentrations of the same oligonucleotide if an end point determination had been performed.

Further advantages of silver staining on DNA arrays are: (i) high spatial and detail resolution, (ii) very good signal-to-noise ratio, (iii) no bleaching or quenching of

the silver-enhanced signal, (iv) after silver staining, arrays can be stored for years without any loss of signal, (v) no use of enzyme which can vary in activity, (vi) no risk to health: no hazardous substances are used, (vii) detection components can be stored for long time, and (viii) inexpensive system, when using versatile indirect labelling and detection systems.

4.2.1.6.4.4 Disadvantages of silver staining of DNA arrays

A drawback of silver-based detection of DNA arrays is that the method cannot be applied for real-time detection during the hybridisation reaction. In contrast to fluorescence-based detection, silver signals are seen only after a number of steps. Also, the final silver enhancement prevents further association and dissociation of DNA to the array. For this reason, silver-stained DNA arrays cannot be re-probed with different samples. Since the vast majority of array experiments do not require re-probing or online detection these disadvantages are usually not of great practical importance.

Other disadvantages are as follows: (i) Inhomogeneities over the area to be stained may result in inhomogeneities in the staining intensity. This problem mainly arises when staining arrays with a large surface area – usually glass slides. This can be avoided by minimising the area to be stained. (ii) Multicolour staining like staining of different targets with different colours is not possible. (iii) Reaction rate varies significantly with changing the reaction temperature. The higher the reaction temperature, the faster the reaction. For this reason, temperature control is necessary and critical.

4.2.1.6.4.5 Limits of silver staining of DNA arrays

Although silver staining of DNA arrays usually shows a good signal-to-noise ratio, unspecific deposition of silver on the surface may reduce the sensitivity of the method. A very low concentration of target results in a greatly delayed staining and weak enhancement of the silver spot which finally is no longer distinguishable from the unspecific background signal. Depending on the assay and the detection protocol used, this limit is reached at concentrations between 50 femtomole and ten picomole.

Unspecific silver deposition can be caused by unspecific binding of target, mediator molecule or gold to the surface of the array. Risk factors for unspecific binding include charged surfaces, aldehyde-coated surfaces, as well as surfaces that are extremely hydrophobic or contain sulphur. Efficient blocking of the surface may therefore significantly reduce the lower limit of detection.

The duration of the silver staining step also plays an important role in the generation of background signal. After a certain time, auto-nucleation of silver ions in the staining solution can occur and the silver begins to precipitate unspecifically from the solution. The length of time which elapses before auto-nucleation occurs depends on the composition of the silver staining mixture and in particular on the presence of reaction modifiers.

Some silver staining solutions are sensitive to light. Exposing these staining solution to light during silver staining will increase the auto-nucleation rate of silver

ions. Exposing to light during storage will result in reduced sensitivity and higher background signal.

Initial experiments show a significant dependence of the silver signal on the concentration of the hybridised target (see figure 4.31.B). At the present time, more experimental data are required to establish whether silver staining is equal or superior to fluorescence-based detection in terms of sensitivity and dynamic range for quantitative analyses.

4.2.1.6.4.6 Main fields of application

Due to its advantages, silver staining of DNA arrays is an ideal detection system for experiments requiring qualitative analysis. Therefore, it is ideal for mutation analysis, genotyping and for detecting the presence of bacteria or virus-specific sequences in clinical samples in circumstances where sensitivity, costs and ease of use are important factors. For most of these applications, a simple measurement of the final spot intensity after a distinct time will be sufficient, although for some special applications dynamic detection during staining may be required.

4.2.1.6.4.7 Conclusions

The use of silver staining for detection of hybridisation events on arrays has been a recent advance in DNA array technology. Despite the limitation of silver staining for detection on DNA arrays to a few kinds of assays, it has a great potential and some decisive advantages over the more commonly used detection methods. The challenges of the future will be the optimisation of the silver staining procedure on the array, the improvement of labelling protocols, and the continuous development of quantitative applications.

4.2.1.6.5 AFM – Atomic force microscopy

Atomic force microscopy (AFM) is based on the use of scanning probe microscopes (SPMs). In contrast to optical and electron microscopes which allow one to visualise a sample, the scanning probe microscope produces an image of the structure of the sample surface. The scanning tunnelling microscope (STM) developed by Gerd Binnig and Heinrich Rohrer was the first such probe instrument (Binnig *et al.*, 1982).

Quantum theory suggests that a current will flow between electrical contacts that are very close, but not quite touching, providing the distance between the contacts is of the order of nanometers. A sharp metallic probe is positioned a few nanometres above a sample adsorbed onto a flat conducting substrate. To scan the surface, the tip is pushed toward the sample until the electron clouds of each touch. Application of a voltage between STM tip and sample causes an electron flow, termed tunnelling current. Since the density of an electron cloud falls exponentially with distance, the tunnelling current is extremely sensitive to the distance between the tip and the surface.

Rather than simply monitoring the tunnel current during a scan, it is usually more convenient to control the vertical displacement of the probe by means of a feedback circuit designed to maintain a constant tunnelling current. The motions of the probe are amplified to visualise the surface features of the sample.

The use of STM is limited in biology because the material under study must conduct electricity. In 1985, Gerd Binnig, Calvin Quate and Christopher Gerber presented an atomic force microscope (AFM) that does not require conducting specimens (Binnig *et al.*, 1986). In AFM, the probe is mounted on a spring and is brought into contact with a surface such that it experiences an interaction force of the order of nanonewtons. The probe is scanned across the surface, while maintaining a constant force between the tip and the sample. Movements of the spring caused by changes in surface stiffness or topography allow topographic contours to be recorded.

Forces are governed by the interaction potentials between atoms. The interaction is attractive at large distances due to van der Waals interaction. At short distances, the repulsive forces are caused by the quantum mechanical exclusion principle, which states that no two fermions can be in exactly the same state (*i.e.*, the same spin, angular momentum, and location).

This principle allows different modes of imaging in AFM. The basic mode is known as the contact mode, in which the tip is held a few angstroms above the surface. The surface and the tip interact by repulsive forces. The alternative mode is known as the non-contact or attractive mode. In such measurements, the tip is held some tens of angstroms above the surface and is oscillated at a frequency above its natural resonant frequency.

In 1995, Laura T. Mazzola and Stephen P. A. Fodor described a method for constructing ordered molecular microarrays and for detecting binding of biomolecules to these arrays using atomic force microscopy (AFM). The authors used glass slides that were coated with the phosphoramidite-bound and light-sensitive MeNPOC protecting group (compound 12 in figure 4.19). Using a lithographic mask that can also be used to produce DNA microarrays as described above, the photo-cleavable MeNPOC group was removed by irradiation of the array with light of a wavelength of 365 nanometres to generate a checkerboard pattern of activated and non-activated areas on the array surface (figure 4.32). This pattern was thought to sim-

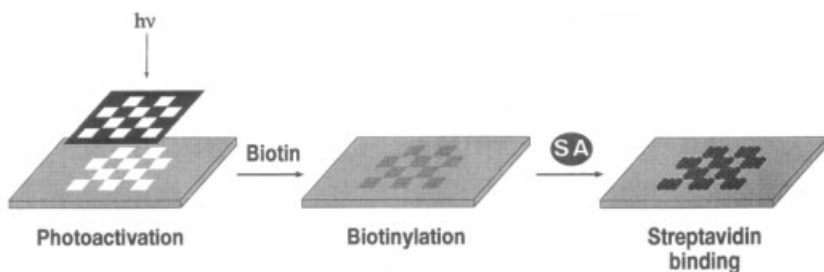


Figure 4.32 Preparation procedure of a protein array in a three-step process: (i) selective photoactivation of the substrate surface, (ii) biotinylation of the activated array, and (iii) incubation of the biotin array with streptavidin. To ensure unambiguous identification of proteins, a

checkerboard illumination pattern was used to activate the substrate surface. This pattern provides a positive and negative control for molecule-binding within an AFM scan. The figure is reproduced from Mazzola & Fodor (1995) with permission from *Biophysical Journal*.

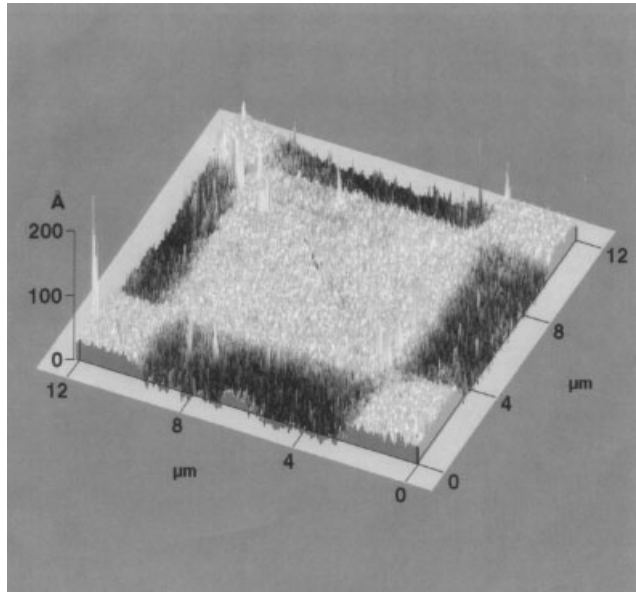


Figure 4.33 Non-contact atomic force microscope (AFM) scan of an eight micrometres biotin/streptavidin array, taken at 470 angstroms pixel resolution. The lighter hue corresponds to taller features of protein. Streptavidin binds with high specificity to the biotin regions,

with a height of about 35 angstroms (?) over the underivatized regions. The following abbreviation is used in this figure: μm , micrometre(s). The figure is reproduced from Mazzola & Fodor (1995) with permission from the *Biophysical Journal*.

plify the discrimination of surface-bound biomolecules from the background substrate by including a negative control area within the AFM scan. Following coupling of biotin-coupled phosphoramidite to the deprotected regions of the substrate surface, a phosphite oxidation was performed to allow deprotection. A final full-surface photolysis was performed to remove all remaining light-sensitive protecting groups. After washing and drying, the biotin-coated arrays were incubated with a solution of streptavidin and scanned using a non-contact atomic force microscope (figure 4.33).

The research team around Stephen P. A. Fodor showed in 1999 (Mazzola *et al.*, 1999) that a modified AFM procedure called chemical force microscopy (CFM) can be used to probe the mechanics of molecular recognition between different oligonucleotides and to detect oligonucleotides on a microarray surface. In CFM, an AFM tip is used that is modified with a biological ligand or receptor to discriminate between its biological binding partner and other molecules on a heterogeneous substrate surface; thus allowing to quantify the strength of the molecular interaction between the two binding partners. In this manner, Mazzola *et al.* used a AFM tip to which a single-stranded oligonucleotide was attached. This DNA-modified tip was used to scan a substrate surface coated with a complementary oligonucleotide (figure 4.34). The strength of the molecular interaction between the two oligonu-

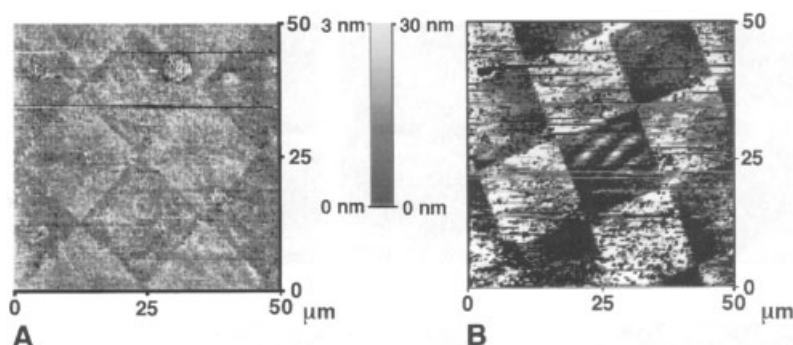


Figure 4.34 Contact atomic force microscopy (AFM) scan of a 16 micrometres checkerboard-patterned DNA oligonucleotide array containing a 20-base complement oligonucleotide alternating with regions of a tetrathymidine base linker, acquired in height mode under ambient conditions. **A)** AFM scan of the oligonucleotide array with a standard silicon nitride tip. The scan has a height contrast of about 15 ang-

stroms. **B)** AFM scan of the oligonucleotide array with the DNA-modified tip. The scan image has an 'apparent' height contrast of about 300 angstroms. Please note the change in scale. The following abbreviations are used in this figure: nm, nanometre(s); μm , micrometre(s). The figure is reproduced from Mazzola *et al.* (1999) with permission from the *Biophysical Journal*.

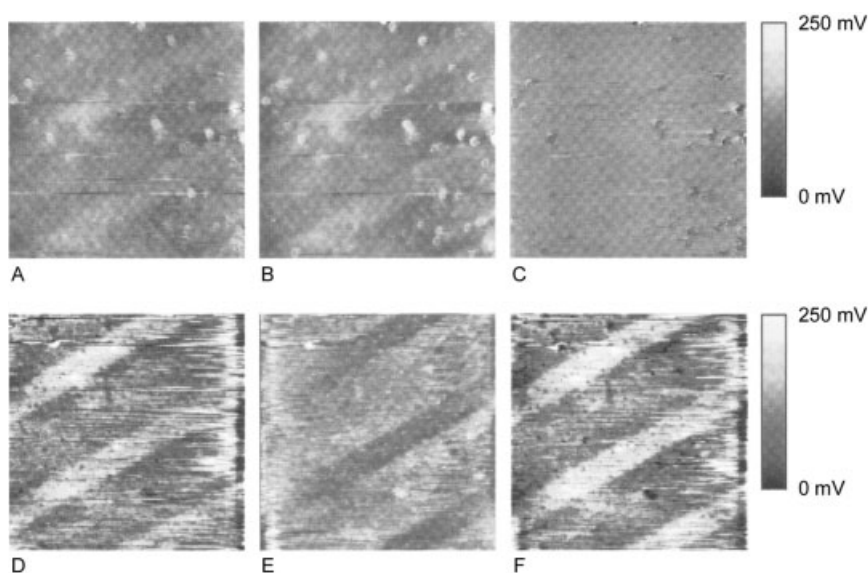


Figure 4.35 Contact atomic force microscopy (AFM) scan of a dual-probe array containing stripes of complementary (ten micrometres) and non-complementary control (five micrometres) oligonucleotides, scanned in deionised water with a DNA-modified AFM tip. **A)** Forward scan. **B)** Reverse scan. **C)** Difference image (forward minus reverse). The same array is

scanned with the same tip in hybridisation buffer. **D)** Forward scan. **E)** Reverse scan. **F)** Difference image (forward minus reverse). The length scale is 50 micrometres and the vertical scale is 250 millivolts (mV) in all images. The figure is reproduced from Mazzola *et al.* (1999) with permission from the *Biophysical Journal*.

cleotides was estimated by using an additional non-complementary oligonucleotide that was also attached to the surface of the microarray (figure 4.35).

The use of atomic or chemical force microscopy is a technically and scientifically demanding approach for the detection of hybridisation events on DNA microarrays. However, it is questionable whether this approach will become a useful tool for the routine detection and analysis of DNA microarrays. Compared to detection procedures based on fluorescence labelling or silver staining, AFM and CFM microscopy require a high technical expertise that in its present form is not suitable in routine laboratory use. In addition, AFM technology is much more expensive than, for example, detection systems based on silver staining procedures. Nevertheless, AFM may be used, for example, as a tool for quality control for DNA or protein microarrays and to check the efficiency of the photolysis reaction during photo-lithographically *in situ* oligonucleotide synthesis on an array surface.

4.2.1.6.6 Electrical/electrochemical detection

Electrochemical devices have also proven to be useful for detection of DNA hybridisation events and a variety of different approaches has been developed during the last years (Mikkelsen, 1996; Wang, 1999 and 2000b). Direct electrochemical detection of DNA hybridisation usually involves monitoring of an electrical current response under controlled voltage conditions. The hybridisation event is detected *via* an increase in the electrical current signal flowing through a redox indicator that recognises the newly-formed DNA duplex or label-free from other hybridisation-induced changes in electrochemical parameters such as conductivity or capacitance (Wang, 2000b).

The team of Susan R. Mikkelsen pioneered the use of redox indicators such as $\text{Co}[2,2'\text{-bipyridine}]_3^{3+}$, $\text{Co}[1,10\text{-phenanthroline}]_3^{3+}$ and $\text{Os}[2,2'\text{-bipyridine}]_2^{2+}$, all of which associate reversibly with immobilised DNA, in detecting the cystic fibrosis F508 deletion that is responsible for about 70 percent of cystic fibrosis cases among Caucasians (Millan *et al.*, 1996). A detection limit of 1.8 femtomole was demonstrated for a 4,000-base DNA fragment using to the redox indicator $\text{Co}[2,2'\text{-bipyridine}]_3^{3+}$.

An important aim of current investigations is the development of new redox indicators that offer greater discrimination between single-stranded and double-stranded DNA for attaining higher sensitivity. For example, in 1998, Takenaka *et al.* introduced a threading intercalator, ferrocenyl naphthalene diimide, that binds to the DNA hybrid more tightly than other intercalators and displays small affinity to the single-stranded probe. This indicator discriminates double-stranded from single-stranded DNA by means of a difference in their dissociation rates.

The use of enzymatic labelling also offers promise for electrochemical detection of DNA hybridisation. The group of Adam Heller demonstrated that direct amperometric monitoring of hybridisation can be achieved using horseradish peroxidase-labelled targets (de Lumley-Woodyear *et al.*, 1996). In this system, the hybridisation event results in the 'wiring' of the enzyme to the transducer, *i.e.* its electrical connection to the electrode, *via* an electron-conducting redox hydrogel, hence leading

to a continuous hydrogen-peroxide electro-reduction current. This approach is similar to the direct enzyme-amplified amperometric affinity assay for the avidin/biotin conjugate described by Vreeke *et al.* (1995). In the latter assay, bonding of biotinylated horseradish peroxidase was used to attach avidin covalently to an electron-conducting redox hydrogel. This also resulted in the wiring of the enzyme and produced a hydrogen-peroxide electro-reduction current proportional to the number of conjugates. The group of Itamar Willner illustrated that multiple amplifications can be achieved by coupling of a peroxidase enzyme label with the surface accumulation of the phenol reaction product (Patolsky *et al.*, 1999).

Similar to the approach described by de Lumley-Woodyear *et al.* (1996), electrochemical sensors have been developed that recognise hybridisation of DNA strands based on electrogenerated chemiluminescence (Kenten *et al.*, 1992; Schutzbank *et al.*, 1995; van Gemen *et al.*, 1994; Xu *et al.*, 1994 and 1995), cyclic voltammetry of intercalated ruthenium and cobalt complexes (Millan & Mikkelsen, 1993; Millan *et al.*, 1994), organic dyes (Hashimoto *et al.*, 1994a and 1994b) and of redox functions covalently bound to single DNA strands (Takenaka *et al.*, 1994).

In the last few years, increased attention has been paid to new label-free electrochemical detection procedures that are promising to offer faster and simpler assays. For example, it is possible to exploit changes in the intrinsic electroactivity of DNA (*e.g.*, the guanine oxidation peak) accrued from the hybridisation event (Johnston *et al.*, 1995; Wang *et al.*, 1998). The inability of this method to detect probes that lack guanine has limited its usefulness. To circumvent this limitation, guanines in the probe sequence can be substituted by inosine that is also able to pair with cytosine. Hybridisation is then detected through the target DNA guanine oxidation signal (Wang *et al.*, 1998). An increased hybridisation response was obtained by an amplified guanine oxidation signal using the $\text{Ru}[2,2'\text{-bipyridine}]_3^{2+}$ redox mediator (Johnston *et al.*, 1995). Direct label-free electrical detection of DNA hybridisation is also possible by monitoring changes in the conductivity of conducting polymer molecular interfaces, for example, DNA-modified polypyrrole films (Korri-Yousoufi *et al.*, 1997; Wang *et al.*, 1999b). Recent developments in this direction are also encouraging (Holmlin *et al.*, 1997; Lee & Shim, 2001).

The electrochemical response of the guanine nucleobase is also sensitive to the DNA structure and can thus be used for probing DNA damage or interactions (Wang, 2000b). Changes in the guanine oxidation or methylation, and of other intrinsic DNA redox signals, have thus been used for detecting chemical and physical damage (Paleček *et al.*, 1998; Wang *et al.*, 1997a). An enhanced protocol for amplifying label-free electrochemical measurements of DNA hybridisation based on the enhanced accumulation of purine nucleobases in the presence of copper ions was recently described by Wang & Kawde (2002). This DNA assay involved hybridisation of the target to inosine-substituted oligonucleotide probes captured on magnetic beads acidic dipurinisation of the hybrid DNA, and adsorptive chronopotentiometric stripping measurements of the free nucleobases in the presence of copper ions. Both amplified adenine and guanine peaks were used for detecting DNA hybridisation. Using this approach, Joseph Wang and his colleague reached a detection limit of 40 femtomoles (250 picograms).

Another interesting approach was recently presented by Huang *et al.* (2002). These authors developed a method for identifying single nucleotide genetic alterations by combining hairpin-forming DNA probes and electrochemical detection of sandwich DNA hybridisation. Incorporation of hairpin-forming competitor probes and the catalysed reporter deposition amplification system further improved the specificity and sensitivity of the assay. The method was used to identify the factor V Leiden mutations from human blood specimens.

Li *et al.* (2000a) introduced for the first time a photo-electrochemical method to characterise the photoelectric behaviour of a single-stranded DNA probe electrode that was prepared with the self-assembly technique and its resulting double-stranded DNA electrode. The obvious decrease in the photocurrent of the double-stranded DNA modified electrode at open potential or a bias voltage indicated that photo-electrochemistry is another useful method for DNA hybridisation detection.

For several years, the development of electrochemical DNA hybridisation sensors used solid electrodes, on which both the hybridisation and detection steps were performed. Paleček *et al.* (2002) and Wang *et al.* (2001b) proposed a new approach in which DNA hybridisations were performed on magnetic beads and electrochemical detection on detection electrodes was used to eliminate non-specific adsorption effects which might otherwise have hampered the label-free detection of DNA hybridisation. Due to low non-specific DNA adsorption at the magnetic beads, very high specificity of the DNA hybridisation was achieved in both cases.

A DNA array detection method was reported by Park *et al.* (2002) in which the binding of oligonucleotides functionalised with gold nanoparticles to the target was detected by changes in conductivity (Service, 2002). The binding events localised gold nanoparticles in an electrode gap; silver deposition facilitated by these nanoparticles bridged the gap and lead to readily measurable conductivity changes. An unusual salt concentration-dependent hybridisation behaviour associated with these nanoparticle probes was exploited to achieve selectivity without a thermal-stringency wash. Using this method, the authors detected target DNA at concentrations as low as 500 femtomolar with a point mutation selectivity factor of approximately 1:100,000. Metal nanoparticle-based electrochemical stripping potentiometric detection of DNA hybridisation is an encouraging approach that is also used by other researchers in this field (*e.g.*, Authier *et al.*, 2001; Wang *et al.*, 2001c).

Although electrochemical detection is a promising tool for the sensitive measurement of DNA hybridisation events, it is still unclear if electrochemical detection can be used as a detection tool for mRNA expression analysis using DNA microarrays of complex nucleic acid mixtures such as the mRNA population of a eukaryotic cell. In addition, it remains to be seen if one of the diverse approaches of electrical/electrochemical detection of DNA hybridisation will meet the requirements of mRNA expression analysis on DNA arrays.

4.2.1.6.7 Mass spectrometry

Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF/MS) and electrospray ionisation (ESI) of proteins are well-established and widely used methods for the identification and analysis of biomolecules such as proteins (see chapter 5). In the last decade of the 20th century, MALDI-TOF/MS modifications were developed that allow identification of DNA oligonucleotides and, as a follow-up, re-sequencing of known DNA sequences to identify single nucleotide polymorphisms (SNPs), mutations and short tandem repeats (STRs).

In the beginning 1990s, first steps to determine the molecular mass of oligonucleotides using MALDI or ESI-based mass spectrometry were presented (*e.g.*, Dunphy *et al.*, 1993; Fitzgerald *et al.*, 1993 and 1995; Huth-Fehre *et al.*, 1992; Kaufmann, 1995; Limbach *et al.*, 1995; Nordhoff *et al.*, 1992; Spengler *et al.*, 1990; Stemmler *et al.*, 1993; Tang *et al.*, 1993 and 1994; Wu *et al.*, 1993). These initial approaches were limited mainly to the identification of uniform oligonucleotides embedded into different kinds of matrices that allow the use of mass spectrometric detection. Since then, several approaches for DNA sequencing using MALDI-based mass spectrometry were developed (for reviews on this application, see Costello, 1999; Crain & McCloskey, 1998; for review articles on mass spectrometry, see Aebersold *et al.*, 2000; Clark *et al.*, 1999; Deforce & van den Eeckhout, 2000; Fei & Smith, 2000; Gatlin *et al.*, 2000; Gevaert & Vandekerckhove, 2000; Griffin *et al.*, 1999; Griffin & Smith, 2000; Griffiths, 2000; Guilhaus *et al.*, 2000; Jackson *et al.*, 2000; Johnston, 2000; Li *et al.*, 2000b, Roepstorff, 2000; Yates, 2000).

A widely used procedure for MALDI-TOF/MS-based DNA sequencing uses primer extension technology. In a first step, DNA of interest is amplified and biotinylated by PCR using a biotinylated oligonucleotide as PCR primer. Amplified single-stranded DNA fragments are isolated using streptavidin-labelled magnetic beads. In the next step, a sequence-specific oligonucleotide is hybridised and a DNA polymerase is used to extend the primer using a mixture of three deoxynucleotides and the corresponding fourth dideoxynucleotide that causes termination of the primer extension reaction (figure 4.36). The resulting elongated DNA strands show allele-specific molecular masses that are detectable by MALDI-TOF/MS.

The resulting DNA is mixed with a matrix solution (*e.g.*, 3-hydroxypicolinic acid in 50 percent acetonitrile) and arrayed on a substrate surface. After drying, laser light is used for structural decomposition of the matrix crystals. Absorption of laser energy by the crystals of the matrix causes spontaneous volatilisation and ionisation of matrix and analyte. Escaping ions are extracted, accelerated and detected (figure 4.37).

For more detailed descriptions on the different DNA sequencing approaches and modifications using mass spectrometry, the reader is referred to the primary literature (*e.g.*, Abdi *et al.*, 2001; Braun *et al.*, 1997; Bray *et al.*, 2001; Buetow *et al.*, 2001; Chicurel, 2001; Fei *et al.*, 1998; Fu *et al.*, 1995, 1996, 1997 and 1998; Garcia *et al.*, 2002; Higgins *et al.*, 1997; Jurinke *et al.*, 2001 and 2002; Leushner & Chiu, 2000; Little *et al.*, 1997; Paracchini *et al.*, 2002; Ross *et al.*, 1998 and 2000; Sauer *et al.*,

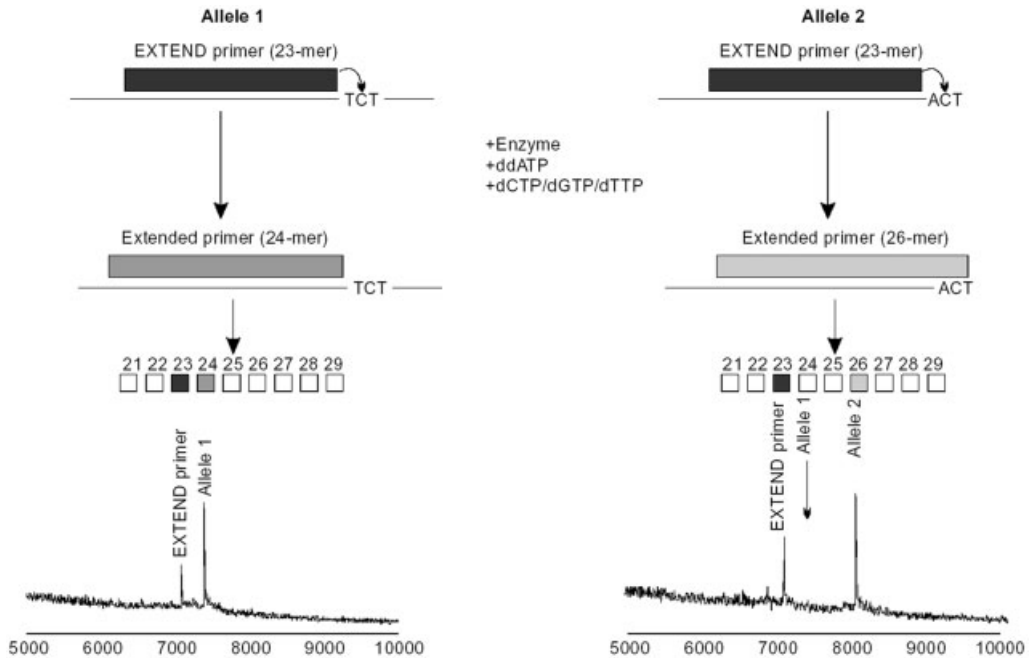


Figure 4.36 Principle of a primer extension reaction used for mass spectrometric analysis of DNA. Illustrated is the principle of the MassExtend™ reaction (Sequenom, San Diego, California, USA). This reaction entails the extension of an oligonucleotide with a combination of deoxynucleotides and dideoxynucleotides with thermosequencase. The left side of the figure is terminated with the ddATP to produce

a 23-mer oligonucleotide and the right side is extended by dTTP, dGTP and terminated with ddATP producing a 26-mer oligonucleotide. The mass spectra show that the non-extended oligonucleotides are easily separated in the mass spectrometer. The figure is reproduced from Leushner (2001) with permission from *Expert Review in Molecular Diagnostics*.

2000 and 2002; Shahgholi *et al.*, 2001; Shchepinov *et al.*, 2001; Stoerker *et al.*, 2000; Stomakhin *et al.*, 2000; Sun *et al.*, 2000 Tang *et al.*, 1995, 1999 and 2002).

MALDI-TOF/MS detection is limited to the analysis of small DNA fragments but has a large potential for high-throughput single nucleotide polymorphism (SNP) analysis, due to a very fast analysis time and possibilities for automation without sample labelling (Larsen *et al.*, 2001). Using sensitive MALDI-TOF/MS detection, it is possible to measure samples in the femtomole range. At present, it is unfortunately not possible to apply MALDI-TOF/MS detection to the analysis of gene expression using DNA microarrays.

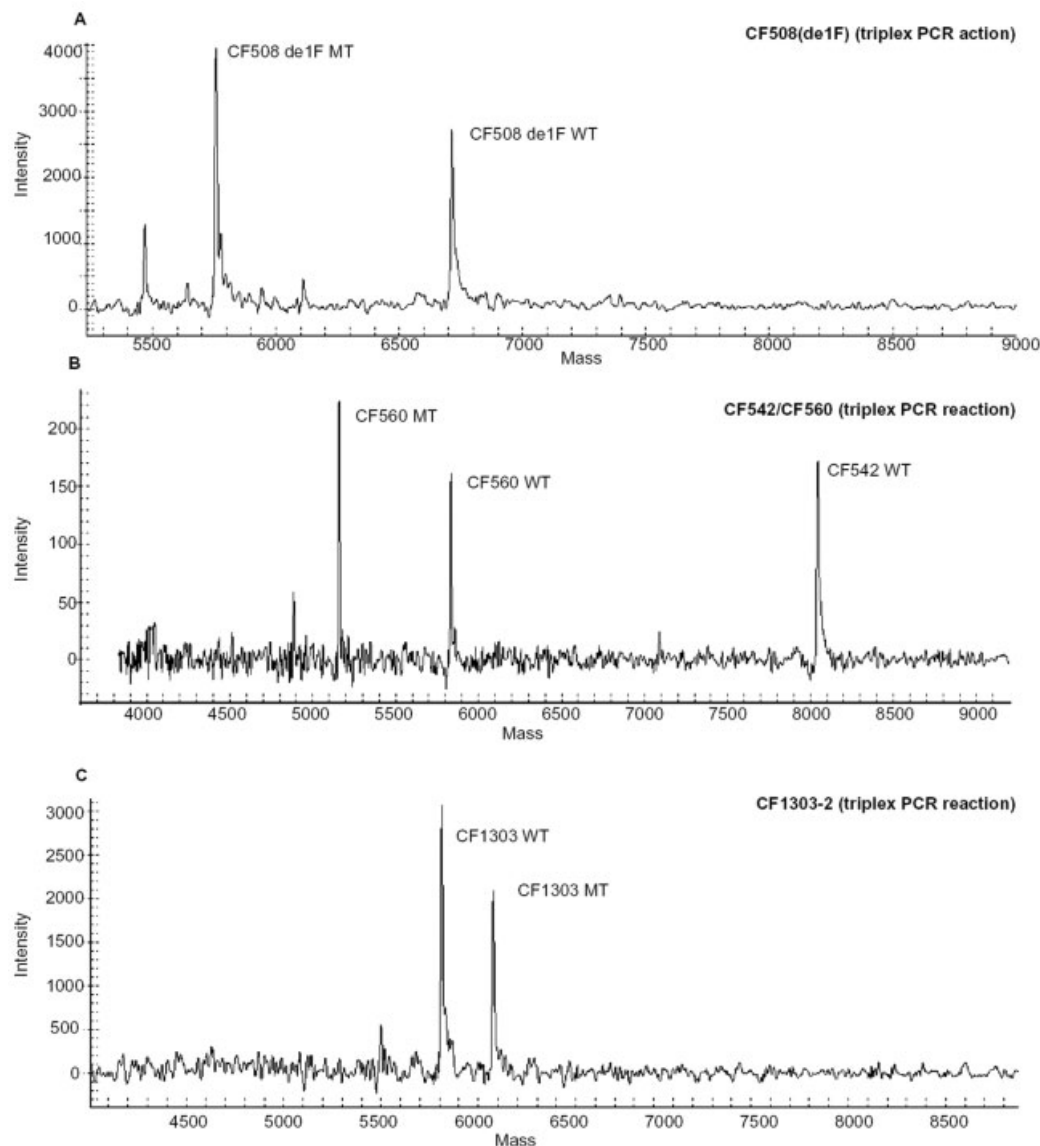


Figure 4.37 Mass spectrometric analysis of mutations. Data from the cystic fibrosis transmembrane conductance regulator (CFTR; ATP-binding cassette transporter, member 7 of subfamily C, ABCC7) gene mutations CF508, CF560, CF542 and CF1303-2 are presented in panels **A**), **B**) and **C**). All expect CF 542 were

heterozygous for the mutant allele. The signal intensity is on the x-axis and mass in dalton on the y-axis. All mutant (MT) and wild-type (WT) alleles were of the expected masses determined in the assay design. The figure is reproduced from Leushner (2001) with permission from *Expert Review in Molecular Diagnostics*.

4.2.1.6.8 QCM – Quartz crystal microbalance

Quartz crystal microbalance (QCM) is an exquisitely sensitive mass-measuring and indicator-free detection principle that allows dynamic monitoring of the interaction between biomolecules including DNA hybridisation events (Okahata *et al.*, 1992; Wang, 2000).

QCM is based on the piezoelectric principle that is widely used in electrochemistry. QCM was invented in 1957 (Sauerbrey, 1959) and first attempts to measure the interaction between biomolecules using QCM transducers were made at the beginning of the 1980s. QCM devices allow detection of a change in mass that occurs during an electrochemical process or the binding of interacting molecules to be converted into a resonant frequency change, which is an easily measurable signal. The basic principle is based on the chronometric properties of the resonant device. A resonator in contact with a solution, in general a quartz crystal, is inserted into an electronic circuit which delivers a large stable signal. This set-up is called an oscillator. Any perturbation of the resonator leads immediately to a measurable change of the oscillation frequency. Through the Sauerbrey relationship

$$\Delta\nu = -2.26 \times 10^{-6} \nu_0^2 \frac{\Delta m}{A}$$

(where $\Delta\nu$ is the change in the frequency of the oscillator, Δm is the mass introduced onto the crystal surface, and A is the crystals surface area), the mass change is calculated from the corresponding change in frequency.

QCM sensors for monitoring DNA hybridisation consist of an oscillating quartz crystal with the single-stranded DNA probe immobilised on its surface. The increased mass, associated with the formation of double-stranded DNA during the hybridisation reaction, results in a decrease of the crystal's oscillating frequency.

Figure 4.38 shows an illustration of a QCM hybridisation measurement. The high sensitivity of QCM transducers has been coupled to the specificity of peptide nucleic acid (PNA) probes in order to detect single-base changes in the *p53* gene (Wang *et al.*, 1997a). A further highly-sensitive micro-gravimetric device was developed by Bardea *et al.* (1998) for detecting the Tay-Sachs genetic disorder. QCM transducers have been used for investigating DNA interactions, protein/DNA interactions in real-time (Nikura *et al.*, 1996) and for the monitoring of enzymatic cleavage reactions (Wang *et al.*, 1999a). Analogous acoustic wave devices have been developed for monitoring the binding of platinum-based anticancer drugs to double-stranded DNA (Su *et al.*, 1995).

Although QCM is a promising tool for the sensitive measurement and indicator-free detection of DNA hybridisation events and for the analysis of protein/DNA interactions, it is still unclear if QCM can be used as a detection tool for mRNA expression analysis using DNA microarrays of complex nucleic acid mixtures.

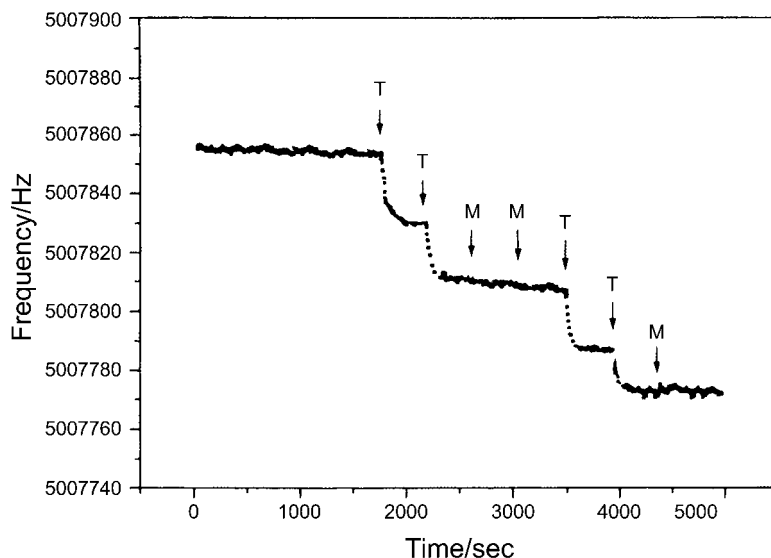


Figure 4.38 Frequency-time response of a peptide nucleic acid/quartz crystal microbalance (PNA/QCM) transducer to the addition of the perfect match target (T) and mismatch (M) oligonucleotides. The hybridisation events result in decreased frequency, reflecting the

increased mass of the crystal. The following abbreviations are used in this figure: Hz, hertz; sec, second(s). The figure is reprinted from Wang *et al.* (1997b) with permission from *Analytical Chemistry*.

4.2.1.6.9 RfS – Reflectometric interference spectroscopy

Reflectometric interference spectroscopy (RfS) directly determines the effective thickness and, thus, the surface concentration of an analyte. The physical principle of RfS is that the reflectivity of a thin transparent film of molecules bound to a surface is modulated by the wavelength of incident light (Janshoff *et al.*, 2000). When a layer of molecules attached to a surface is illuminated with white light through the substrate, the light is reflected at the solution interface and the substrate interface of the layer. If monochromatic light is used, either constructive and destructive interference occurs, dependent on the phase shift. The interference obtained using white light varies with the wavelength, so that a periodic modulation of the reflecting light intensity, the so-called interference spectrum, results. The positions of the maxima and minima depend only on the thickness of the molecule layer at a given refractive index and on the angle of incidence. The constructive interference obtained by illumination along an axis coincident with the surface normally is given by the following equation,

$$2nd = m\lambda$$

in which n is the refractive index, d is the thickness of the molecule layer, m is the order of the maxima, and λ is the light wavelength.

An increase in the thickness of the transparent molecule layer results in a shift of the interference spectrum to larger wavelength. Strictly speaking, this principle holds true only for transparent films with a thickness in the range of the wavelength of the incoming light. This condition does not apply to high-molecular weight molecules such as proteins or to low-molecular weight analytes. However, a simple experimental modification allows defined interference spectra to be obtained. For example, a glass substrate may be covered with a thick interference layer of a polymer material such as gels or oxides such as silicon oxide. This layer exhibits an interference pattern and is functionalised by receptor molecules. Adsorption of the binding partner causes a change in the thickness of the layer that is far below the wavelength of radiation. This modification allows measurements of changes in effective optical thickness of as little as one picometre.

Such RIfS modifications have been used to quantify protein/protein interactions, binding of pesticides and DNA/DNA interactions (Barzen *et al.*, 2002; Birkert *et al.*, 2000; Coille *et al.*, 2002; Dieterle *et al.*, 2001; Janshoff *et al.*, 2000; Jung *et al.*, 2001; Piehler *et al.*, 1997; Schobel *et al.*, 2000).

Although proof-of-principle has been obtained that RIfS may be used for quantification of DNA hybridisation, further enhancement and optimisation is required before this method can be used for the analysis of more complex nucleic acid mixtures.

4.2.1.6.10 SPR — Surface plasmon resonance

While the vast majority of array-based studies of bioaffinity interactions employ fluorescently labelled biopolymers, there is a need for the continued development of sensitive analytical methods that can be used to detect affinity interactions in biological samples without the need of molecular labels or tags. This is especially true for the case of protein/protein interactions, where labelling can interfere with protein function and is more difficult than that of nucleic acids.

Surface plasmon resonance (SPR) reflectivity imaging is a surface-sensitive, label-free optical technique that can be used to characterise the thickness and/or index of refraction of ultra-thin organic or biopolymer films on chemically modified noble metal (gold, silver, or copper) surfaces by measuring the change in the local index of refraction that occurs upon adsorption of a molecule to the layer. SPR is based on the collective excitation of electron oscillation in ultra-thin films of noble metals. Photons of the near-infrared light range striking on the metal's surface cause a resonance oscillation of the electrons. The resonance oscillation is influenced by the properties of the boundary layer between the metal surface and a covering solution of molecules such as proteins or nucleic acids. Chemical and physical changes in the boundary layer cause a change of the resonance oscillation of the electrons that is measurable with a high sensitivity. Thus, SPR imaging allows monitoring of the adsorption of various solution species into patterned molecular arrays that have been formed on chemically modified metal surfaces; in particular the detection of hybridisation of nucleic acid strands or protein/DNA, protein/RNA and protein/protein interactions at metal interfaces in a spatially resolved manner is possible (Brockman *et al.*, 2000).

An SPR imager is based on a collimated, polychromatic light source. The emitted light passes through a polariser and impinges on a prism/thin metal film sample assembly at an incident angle above the critical angle. The reflected light then passes through a narrow-band interference filter and is detected with a charge-coupled device (CCD) camera that allows real-time video imaging of the occurring binding reactions.

SPR imaging allows to detect the change in the molecular thickness of a molecular layer of less than one angstrom. The spectral used ranges from 600 to 1,100 nanometres and the incident angle ranges from 40 to 70 degrees for differing prism and sampling systems.

For measuring hybridisation events of nucleic acids using SPR, gold surfaces deposited on special glass slides are usually used. Oligonucleotides can be attached to self-assembled monolayers on the gold surface. When an oligonucleotide hybridises to its complement on the surface, the refractive index of the solution adjacent to the gold changes. This causes a measurable change in reflectivity of the surface at each site of hybridisation.

Surface plasmon resonance spectroscopy has become widely used in the fields of chemistry and biochemistry to characterise biological surfaces and to monitor binding events. In particular, several research groups established the experimental set-up to measure the interaction between DNA, RNA and PNA molecules. Although first attempts were made in the late 1980s (Pollard-Knight *et al.*, 1990; Schwarz *et al.*, 1991), several years passed before SPR imaging became a useful tool to investigate nucleic or peptide acid hybridisation. Recently, several approaches have been published using SPR imaging for mutation analysis (Brockman *et al.*, 2000; Burgenner *et al.*, 2000; Feriotto *et al.*, 2001; Gotoh *et al.*, 1995; Guedon *et al.*, 2000; Heaton *et al.*, 2001; Jin-Lee *et al.*, 2001; Kambhampati *et al.*, 2001; Kukanskis *et al.*, 1999; Nakatani *et al.*, 2001; Nelson *et al.*, 2001; Peterson *et al.*, 2001; Smith *et al.*, 2002).

The success of SPR is primarily due to three factors: (i) the kinetics of biomolecular interactions can be measured in real-time, (ii) the adsorption of unlabelled analyte molecules to the surface can be monitored, and (iii) SPR has a high degree of surface sensitivity that allows weakly bound interactions to be monitored in the presence of excess solution species.

Because of its high surface sensitivity and its ability to measure in real-time the interaction of unlabelled biological molecules with arrays of surface-bound species, SPR imaging has the potential to become a powerful tool in biomolecular investigations. However, SPR imaging has not yet been applied to the analysis of gene expression using DNA microarrays.

4.2.1.7 Conclusions

Over the past decade intense activity aimed at developing DNA microarrays. Such devices offer considerable promise for obtaining sequence-specific information in a faster, simpler and cheaper manner compared to traditional assays. These DNA microarray technologies are rapidly advancing; applications ranging from genetic testing to measuring gene expression and drug discovery have been widely demon-

strated. Innovative efforts, coupling fundamental biological and chemical sciences with technological advances in the fields of micro-machining and micro-fabrication should lead to the development of even more powerful devices that will accelerate the realisation of large-scale functional genomic analyses. A wide range of new DNA microarrays are thus expected to reach the market in the coming years.

Developments of stains, labels and strategies for detecting and measuring nucleic acids and related molecules are surveyed based on the literature in the final decade of the 20th century (Kricka, 1999 and 2002; Wang, 2000). During this period, important families of cyanine dyes were developed for sensitive detection of double-stranded DNA, single-stranded DNA, and oligo(deoxy)nucleotides in gels and in solution. The continuing quest for improved labels for hybridisation assays has produced a series of candidate labels including genes encoding enzymes, microparticles (*e.g.*, quantum dots, nano-crystals, phosphors), and new examples of the fluorophore (*e.g.*, cyanine dyes), intercalating agents (*e.g.*, ethidium bromide, SYBR[®] green I and II) and light-emitting enzymes (*e.g.*, firefly luciferase mutants). Label detection technologies for use in Northern and Southern blotting assays have focused on luminescent methods, particularly enhanced chemiluminescence for peroxidase labels and adamantyl 1,2-dioxetanes for alkaline phosphatase labels. Micro-analytical high-throughput devices such as DNA microarrays present new challenges in nucleic acid detection. At the present time, fluorescent methods dominate. However, several new approaches have been presented in the last years that are not yet competitive to fluorescence, but that hold a great promise for the future.

4.2.2

OFP – Oligonucleotide fingerprinting

4.2.2.1 Introduction

Although ever more genomes will be fully sequenced over the coming years, the complete genome sequence on its own is not sufficient to understand the biology of an organism. Many of the potential coding sequences within these completed genomes are not readily identifiable by gene prediction programs. Furthermore, correct splice patterns and splice variants are even more difficult to predict by computer analysis. By contrast, cDNA libraries ideally represent all expressed genes specific for the biological source from which they are derived (*i.e.*, a cell type, a tissue, or a developmental stage). They provide direct access to the final splice variants and therefore correct protein products and give valuable insight into the abundance levels of different mRNAs in the starting material. To guarantee that low-abundance mRNAs or very rare transcripts (with about one molecule per cell) are represented in a cDNA library, often consisting of several cell types, it is necessary to construct and screen libraries containing up to several million independent cDNA clones (Sambrook *et al.*, 1989).

One way of analysing such libraries is the oligonucleotide fingerprinting (OFP) approach which was conceptually developed more than a decade ago (Poustka *et al.*, 1986) and has been successfully implemented in numerous applications (Clark

et al., 2001; Craig *et al.*, 1990; Drmanac *et al.*, 1996; Hoheisel *et al.*, 1991; Maier *et al.*, 1994; Meier-Ewert *et al.*, 1993 and 1998; Milosavljevic *et al.*, 1996; Poustka *et al.*, 1999; Radelof *et al.*, 1998). In principle, the method is based on the concept of sequencing by hybridisation (Drmanac *et al.*, 1989, Lysov *et al.*, 1988) where a complete target sequence is ultimately deduced from independent hybridisations with a set of all possible oligonucleotides of a given length. However, OFP differs from that concept in that only partial sequence information (a fingerprint) is gained. Through a series of up to 250 successive hybridisations of short, radioactively labelled oligonucleotides of known sequence, vectors of hybridisation signal intensities – oligonucleotide fingerprints – are generated for each clone of an arrayed library. According to their individual fingerprints, clones are grouped into clusters. Ideally, each cluster represents one gene with the size of the cluster indicating its relative level of expression (Drmanac *et al.*, 1996; Meier-Ewert *et al.*, 1998). Via database comparisons, known genes are identified and unknown candidate genes are discovered. A representative clone of each cluster is selected for subsequent experiments, such as the generation of gene expression arrays. Thus, a highly normalised almost non-redundant clone set is created representing the expressed gene repertoire of the analysed biological source.

4.2.2.2 Elements of oligonucleotide fingerprinting

OFP is a complex process consisting of several experimental and evaluation steps. It can be subdivided into a molecular biology and a bioinformatics section which are, however, intertwined. The former section comprises all practical steps starting from cDNA/genomic library construction up to the detection of hybridisation signals. The latter section includes the oligonucleotide probe design as well as all data analysis elements. A concise graphic overview of the OFP pipeline is given in figure 4.39. Comprehensive descriptions of the procedure and detailed protocols of the laboratory process are available in Clark *et al.* (1999), Meier-Ewert *et al.* (1998) and at http://www.molgen.mpg.de/~ag_onf/. In the following, all elements of the OFP pipeline are briefly described.

4.2.2.2.1 Probe selection

The choice of oligonucleotide probes is essential for the identification of clone sequences by hybridisation experiments because it has a strong influence on the clustering quality of the resulting fingerprints. The probe selection algorithm that is routinely used optimises probes with respect to a training set of sequences based on Shannon entropy as a quality criterion (Herwig *et al.*, 2000). This has been shown to be superior to the selection of the most frequent probes and to randomly chosen probe sets.

4.2.2.2.2 Laboratory processes

All individual steps starting from cDNA/genomic library construction up to the detection of hybridisation signals are shown and described in figure 4.39.

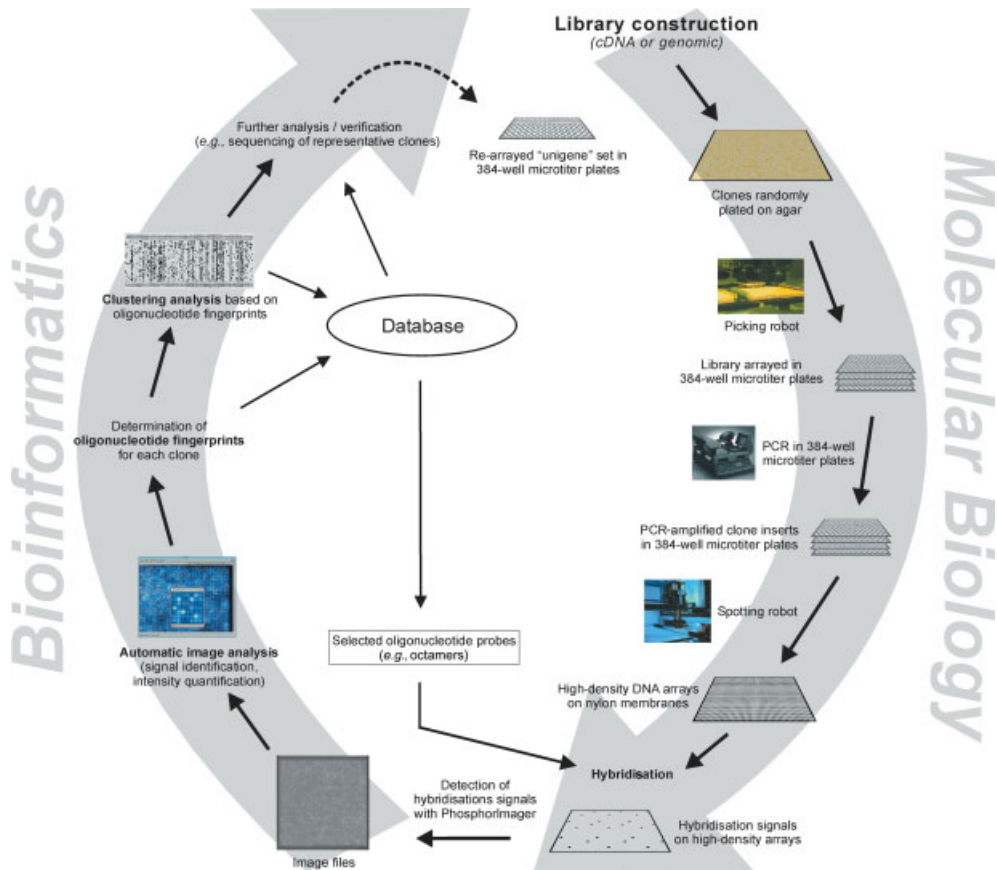


Figure 4.39 Overview of the oligonucleotide fingerprinting process. Genomic or cDNA libraries are seeded out on agar plates in the form of bacterial suspensions. By the use of a self-devised picking robot (Maier *et al.*, 1994) well separated grown bacterial colonies (genomic or cDNA clones) are automatically transferred into 384-well microtiter plates. In order to generate pure and concentrated target DNA, clone inserts are amplified by PCR carried out in the same plate format by means of a Peltier thermal cycler. PCR products are spotted in duplicates onto nylon membranes employing a spotting robot developed in-house (Maier *et al.*, 1994). Using this procedure, high-density DNA arrays of twice about 28,000 individual PCR products are routinely produced. Subsequently, up to 250 radioactively labelled octamer oligonucleotide probes of optimised sequence (Herwig *et al.*, 2000) are in turn hybridised onto the high-density DNA arrays. Hybridisation

events are detected by means of phosphorus storage auto-radiography and digitally stored as image files. Individual signal intensities are quantified for each spot in every hybridisation image using an automatic image analysis program (Steinfath *et al.*, 2001). By conversion of the respective signal intensities, oligonucleotide fingerprints are determined for each clone across all hybridisations. The fingerprints obtained are subjected to a clustering analysis (Herwig *et al.*, 1999) by which similar fingerprints are assigned to the same cluster whereas dissimilar ones are separated. For each cluster a consensus oligonucleotide fingerprint is computed. The clone showing highest similarity to this consensus fingerprint is chosen as the representative clone for that cluster. The set of representative clones of all clusters embodies a non-redundant 'unigene' set of the library under investigation.

4.2.2.2.3 Image analysis and data normalisation

Scanned images are analysed as described in figure 4.40. After quantification, raw signal intensities are normalised using a double-ranking method (Milosavljevic *et al.*, 1995) through which, for each clone, all intensities are replaced by their ranks – first within each and second across all experiments – as described in Herwig *et al.* (1999).

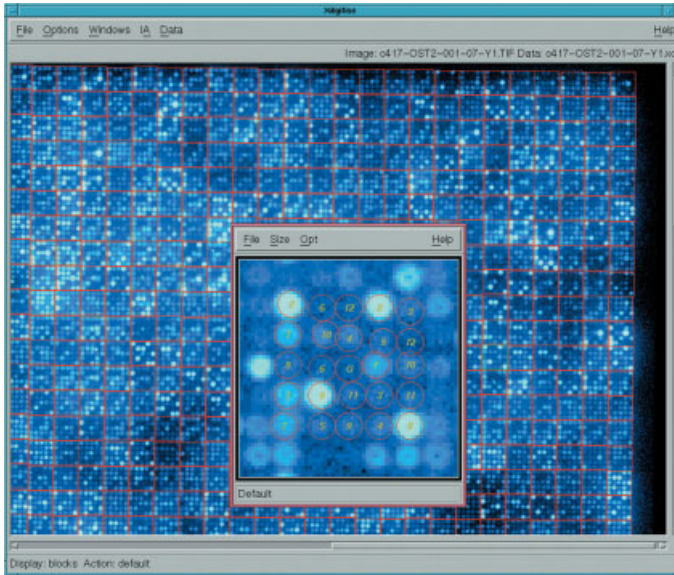


Figure 4.40 Automated image analysis. In a two-step procedure (Steinfath *et al.*, 2001), the location of each spot is first computed by placing a rectangular grid on the spotted array incorporating local corrections ('grid finding'). In a second step, each spot is assigned an intensity corresponding to the respective amount of hybridisation ('quantification').

4.2.2.2.4 Clustering procedure

After image analysis and normalisation, each clone is represented by a vector of normalised signal intensities – its so-called oligonucleotide fingerprint. Through a clustering procedure based on a sequential K-means algorithm (see chapter 7) these fingerprints are grouped according to a well-defined pairwise similarity measure (Herwig *et al.*, 1999). Similar fingerprints are grouped together while dissimilar ones are separated (figure 4.41). The clustering result reflects the number of different expressed genes (number of clusters) and their relative abundance (size of clusters).

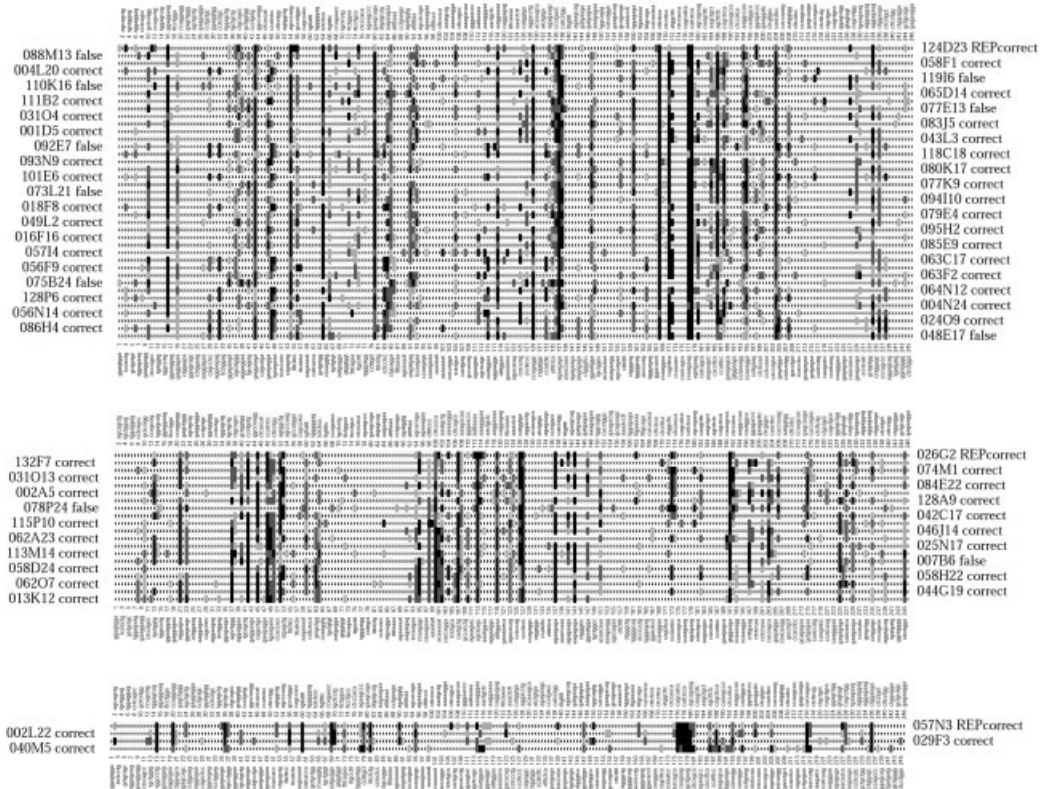


Figure 4.41 Visualisation of three clusters of different size of a cDNA library derived from sugar beet developing root. The clone names (internal abbreviations) are listed on the left and right, whereas the probe sequences are listed on top of and underneath each illustration, respectively. Clones that showed a signal on control hybridisations (see 4.2.2.2.5) are marked 'correct' the remaining ones were false.

sely clustered and are marked 'false'. The representative clone of each cluster that was selected for sequencing is marked 'REP'. The results demonstrate the robustness of the oligonucleotide fingerprinting method. Even for small clusters, experimental noise randomly spread among all 246 hybridised probes could be compensated by data analysis.

4.2.2.2.5 Verification of cluster analysis

Clustering results must be verified by control experiments, such as tag sequencing and hybridisations of single cDNA clones to the DNA arrays (Clark *et al.*, 2001). In the case of the latter, all clones derived from the same gene give a positive hybridisation signal validating the clustering results.

4.2.2.3 Applications of oligonucleotide fingerprinting

OFP has been successfully employed in a wide range of applications, such as the identification of overlapping clones in order to construct ordered clone libraries (Craig *et al.*, 1990) and the establishment of clone maps of large insert clones (Hoheisel *et al.*, 1991). More recent applications comprise the reduction of redundancy within large-scale genomic sequencing projects (Radelof *et al.*, 1998) and various expression profiling studies examining different model organisms or tissues (Clark *et al.*, 2001; Eickhoff *et al.*, 2000; Meier-Ewert *et al.*, 1998; Poustka *et al.*, 1999).

In the following, the most significant fields of application are presented. These are the generation and characterisation of normalised cDNA libraries to create 'unigene' sets (a unigene set is a non-redundant set of cDNA clones; each cDNA clone of such a set represents a different gene), differential cDNA library screening, and pre-selection of shotgun clones.

The main application of OFP is the normalisation and characterisation of cDNA libraries of a certain developmental stage, tissue or cell type of an organism under investigation. In this process, a highly normalised 'unigene' cDNA clone set is derived, which represents the almost complete gene inventory of the organism or tissue. For several reasons 'unigene' sets are highly valuable resources. They are well-defined and represent a non-redundant clone set which is directly accessible, *e.g.*, in the form of bacterial clones. In addition to global sets, tissue-specific and development-specific subsets can be arrayed and used for various downstream applications such as expression profiling *via* global or tissue-specific gene expression and protein arrays. A variety of projects has taken advantage of the normalisation effect achieved by OFP and the specific 'unigene' set which resulted from there. These projects comprise the normalisation of embryonic mouse cDNA libraries (Meier-Ewert *et al.*, 1998) and their subsequent gene expression profiling by complex hybridisation (Eickhoff *et al.*, 2000), the use of OFP as a means of pre-screening in the generation of a partial sea urchin gene catalogue (Poustka *et al.*, 1999) and for a large-scale zebrafish EST project (Clark *et al.*, 2001) as well as the creation of 'unigene' sets for sugar beet, bovine brain and human and murine T-lymphocytes and natural killer cells. Figure 4.42 shows the superior normalisation performance of OFP in comparison with the reassociation kinetics technique which was applied for the most extensively used human cDNA libraries (Soares *et al.*, 1994). More than two thirds of the generated ESTs of the normalised MPMGp609 (Max-Planck Institute for Molecular Genetics plasmid library number 609) library represent unique sequences. In recent projects, resulting EST sets contain 90 percent unique sequences.

The suitability of OFP as a means of differential cDNA library screening compared to other differential expression technologies, such as EST analysis (Adams *et al.*, 1995) and SAGE (Velculescu *et al.*, 1995), was first shown in a pilot study by Meier-Ewert *et al.* (1998). In that study, the gene expression of two stages of embryonic mouse development was compared by clustering analysis of two different oligofingerprinted cDNA libraries ('co-clustering'). Several hundred clusters were identified that showed significant differential representation. Furthermore, in their large comprehensive OFP study on an embryonic and an adult-liver zebrafish

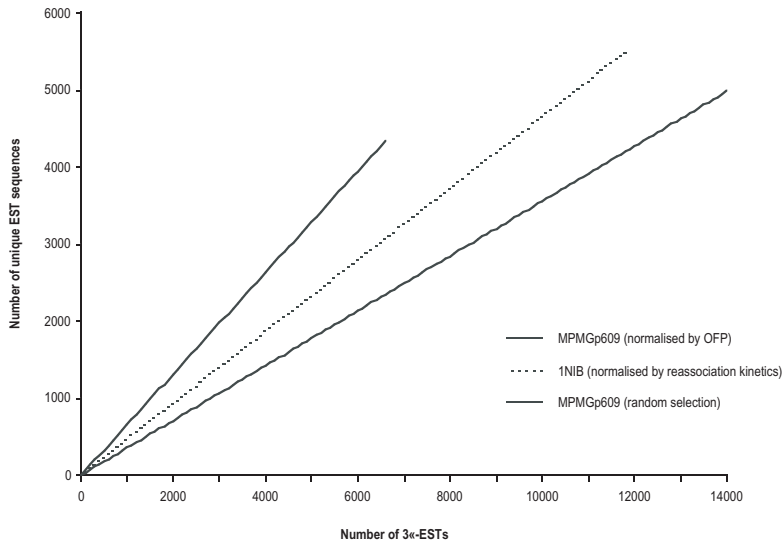


Figure 4.42 Impact of normalisation on the redundancy of expressed sequencing tag (EST) projects. The number of unique EST sequences as a function of the overall number of 3'-ESTs is shown. *MPMGp609 library (Max-Planck Institut for Molecular Genetics plasmid library number 609; normalised by oligonucleotide fingerprinting)*: More than 6,000 3'-sequences of the normalised MPMGp609 zebrafish library (Clark *et al.*, 2001) were assembled. Starting from 200 sequence reads, all unique EST sequences were counted by following the chronological order during sequencing. *1NIB library (normalised infant brain library 1)*: The assembly of roughly

12,000 3'-ESTs of the infant brain library normalised by reassociation kinetics (Bonaldo *et al.*, 1996, Hillier *et al.*, 1996) was carried out in the same way as for the normalised MPMGp609 library. *MPMGp609 (random selection)*: on the basis of the clustering analysis for the whole oligofingerprinted MPMGp609 library (about 21,000 clones) a random selection was simulated. Out of all selected clones, only those representing singletons were considered, thus allowing to simulate a non-normalised library. The diagram clearly shows the higher degree of normalisation accomplished by oligonucleotide fingerprinting.

cDNA library, Clark *et al.* (2001) were able to isolate and identify those transcripts specific to the respective library. Their results strongly underpin the findings of the pilot study in that OFP is confirmed as an effective system for the identification of differentially expressed genes and the quantification of their expression levels.

The aim of pre-selection by OFP (Radelof *et al.*, 1998) is to reduce sequencing redundancy experienced in traditional shotgun sequencing projects. Usually, completed shotgun projects show an eight to twelve-fold average coverage per base final sequence, more than necessary for consensus sequence data of sufficient quality. Through pre-selection by OFP, a set of minimally overlapping shotgun clones, derived from a single original genomic clone, is selected prior to sequencing on the basis of maximally dissimilar hybridisation patterns (fingerprints). By this efficient combination of OFP and shotgun sequencing, the number of shotgun clones to be sequenced is significantly reduced (at least two-fold) and substantial cost savings are made.

4.2.2.4 Technology development

As a result of its multiple robotic steps and automated bioinformatics element the OFP technology can be employed in numerous large-scale genomic applications. Nevertheless, the current technology shows some inherent limitations which need to be overcome in order to further increase the degree of automation and, hence, throughput. Bottlenecks are the high number of oligonucleotide probes needed, the confined array density and the restriction to serial hybridisations.

For probe to target hybridisations, the optimal amount of information is gained if the hybridisation probability of the probe is 50 percent. Since the hybridisation frequencies of the DNA octamer probes are in practice much lower, the overall number of hybridised oligonucleotides must be increased in order to achieve a sufficient partitioning and, thus, clustering quality of the analysed library (Herwig *et al.*, 1999 and 2000). One possible way to statistically increase hybridisation frequency is the use of shorter probes which would be particularly advantageous for the characterisation of short insert (cDNA or shotgun) libraries. Although the use of DNA oligonucleotides as short as hexamers has been reported (Drmanac *et al.*, 1990), very short DNA probes predominantly show poor hybridisation stability. This drawback can be overcome by employing peptide nucleic acids (PNAs) (Nielsen *et al.*, 1991). Peptide nucleic acids possess a neutral pseudo peptide backbone and show more favourable hybridisation properties than their DNA counterparts in terms of both stability and mismatch discrimination (Egholm *et al.*, 1993). In fact, PNAs can be used as short as six nucleotides to generate reliable hybridisation results and to reduce the overall number of probes needed.

Another restricting factor is the DNA array density which is limited to about 125 spots per square centimetre. Higher densities impair the data quality due to spot-to-spot over-shining effects of the radioactively ³³phosphorus-labelled probes. By the combination of fluorescent probes with laser-scanner detection, signal interferences can be excluded (as long as the laser spot diameter is smaller than the spot-to-spot distance) and higher spotting densities can be achieved.

Taking advantage of PNA properties and the convenience of fluorescence detection a novel oligofingerprinting protocol has been developed (Guerasimova *et al.*, 2001). It provides a comprehensive new environment for OFP by combining the superior hybridisation properties of PNAs and the fluorescent characteristics of phycoerythrin with a very powerful, but simple laser-scanning system and our automated image analysis software (Guerasimova *et al.*, 2001). Signal detection rates have been increased and signal interference effects have been strongly reduced, thus allowing an increase in spotting density of at least two-fold.

Hybridisation of probes in pools and their parallel detection comprise a further potential for improvement of the OFP technology. Although the detection of 27 different colours by combinatorial multi-fluor fluorescence *in situ* hybridisation (FISH) has been described (Speicher *et al.*, 1996), the number of fluorescence labels to be detected simultaneously is limited. An alternative possibility towards a higher degree of multiplexing is to detect probes by virtue of their molecular mass. Matrix-assisted laser desorption/ionisation-based time-of-flight mass spectrometry (MALDI-TOF/MS) (Karas *et al.*, 1988) offers the simultaneous detection ca-

capacity necessary for multiplexing which allows the hybridisation of many probes in parallel. This alternative is being pursued in an ongoing project where the use of PNAs is combined with MALDI-TOF/MS. In addition to their favourable hybridisation properties, PNAs also show a high ion stability compared to DNA rendering them ideally suited for MALDI-TOF/MS detection conditions (Butler *et al.*, 1996; Tomac *et al.*, 1996). This innovative approach offers the potential to further automate the OFP procedure employing DNA microarrays as well as automatic mass spectrometry detection and substantially increase the throughput by at least one order of magnitude (please refer to <http://www.molgen.mpg.de/~ag-onf/> for further details).

4.2.3

Quantikine® mRNA assay

4.2.3.1 Introduction

Quantification of gene-specific mRNA is a common, yet extremely important tool in studies of gene expression and function. Methods such as Northern blot analysis, ribonuclease protection assay (RPA), and RT-PCR have been and continue to be routinely used for these types of studies. In this chapter, a novel, microplate-based mRNA detection assay, the Quantikine® mRNA assay from R&D Systems, Inc. (Minneapolis, Minnesota, USA), is described (product literature) which offers notable advantages, including ease of use and sample preparation, as well as high-throughput capabilities. The Quantikine® mRNA assay can be completed in less than five hours, allows the direct use of cell lysate samples, and is performed in a 96-well microplate format. Furthermore, this assay does not make use of radioactivity or target amplification and results can be obtained using a conventional microplate reader. Other microplate-based mRNA quantification assays based on different methods than that used in the Quantikine® mRNA assay have also been developed, such as the QuantiGene™ Expression assay (product literature) and the VERSANT® RNA assay from Bayer Diagnostics (product literature), the Invader® RNA assay from Third Wave Technologies, Inc. (Allawi *et al.*, 2001; Kwiatkowski *et al.*, 1999, and product literature), the Xpress-Screen® mRNA Detection assay from Applied Biosystems (Aristarkhov *et al.*, 2000, and product literature), and the Array-Plate™ Multiplexed Molecular Profiling system from High Throughput Genomics, Inc. (Felder & Kris, 2001, and product literature). The Quantikine® mRNA assay will be discussed in detail here and compared to the other commercially available assays developed within the last few years.

4.2.3.2 Principles and basics

Quantikine® mRNA assays use an ELISA-like format with colourimetric detection to quantify gene-specific mRNA in less than five hours. They provide versatility in sample requirements by allowing the use of cell lysates, total RNA, or poly(A)⁺ selected mRNA. The assay first involves hybridising samples with gene-specific biotin-labelled capture probes and digoxigenin-labelled detection probes in a micro-

plate well (figure 4.43). The hybridisation solution is then transferred to a streptavidin-coated microplate where the mRNA/probe hybrid is captured through biotin/streptavidin interactions. Following washes to remove unbound material, the captured mRNA/probe hybrid is incubated with an anti-digoxigenin, alkaline phosphatase-conjugated antibody. After washing away unbound antibody, a substrate solution is added, followed by an amplifier solution, to allow colour development. The intensity of the colour that develops is proportional to the amount of gene-specific mRNA in the test sample. Colour development is stopped, and the intensity of

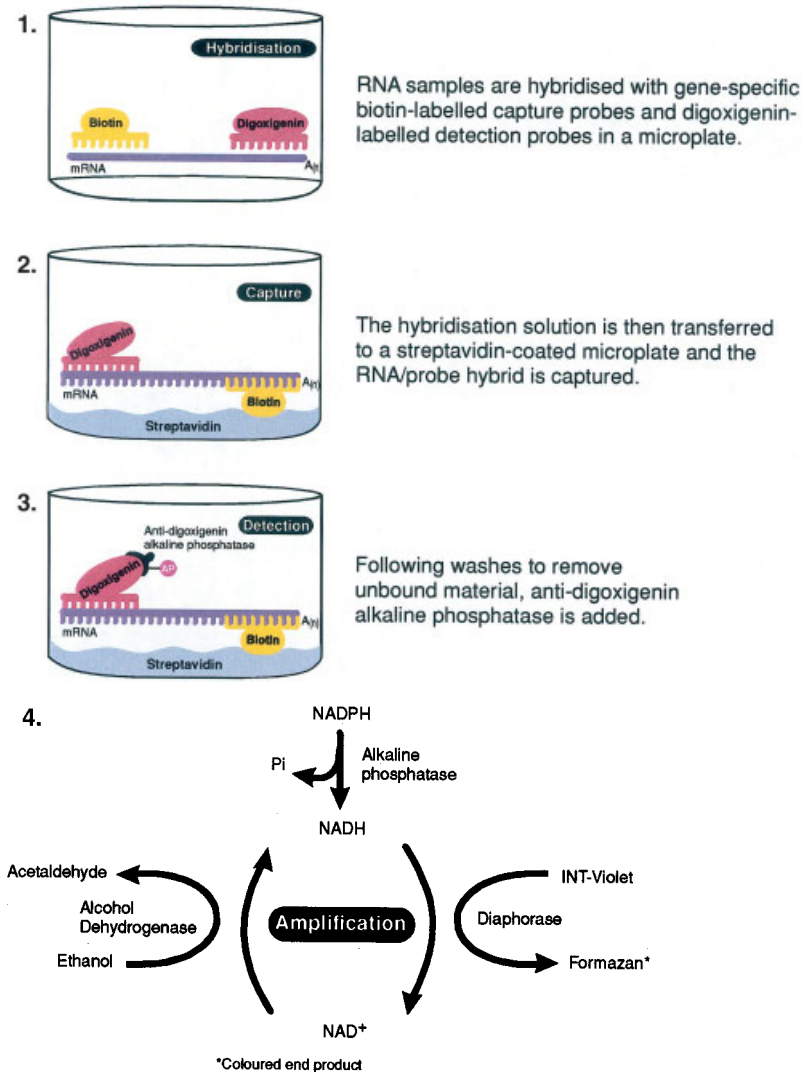
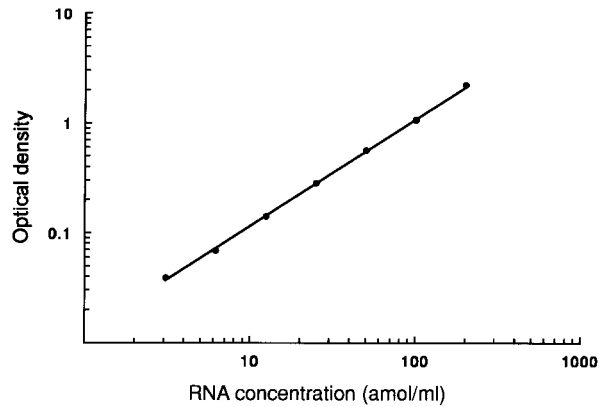


Figure 4.43 Schematic presentation of the principle of the Quantikine® mRNA assay. The following abbreviation is used in this figure: INT, *p*-iodonitrotriazolium.



Amount of RNA (amol/ml)	Optical density	Average of optical density	Corrected optical density
0.0	0.0059 0.0059	0.0059	-
3.1	0.097 0.099	0.098	0.039
6.2	0.128 0.129	0.129	0.069
12.5	0.203 0.198	0.201	0.141
25	0.333 0.350	0.342	0.283
50	0.639 0.604	0.622	0.563
100	1.157 1.095	1.126	1.067
200	2.304 2.265	2.285	2.225

Figure 4.44 Human interleukin-2 calibrator curve. The curve was generated using the Quantikine® interleukin-2 calibrator mRNA and the human interleukin-2 Quantikine® mRNA kit. (R & D Systems, Inc., Minneapolis, MN, USA).

the colour is measured spectrophotometrically on a colourimetric microplate reader. A standard curve is generated for each mRNA target using a gene-specific calibrator RNA, which is synthesised by *in vitro* transcription. A representative standard curve for a Quantikine® mRNA assay is shown in figure 4.44. A standard curve should be generated each time an assay is performed to correctly calculate the concentration of mRNA in test samples.

4.2.3.3 Typical results

Linearity of the Quantikine[®] mRNA assay was assessed by quantifying human interleukin-2 mRNA levels in cells of the HuT-78 lymphoma cell line stimulated with phorbol 12-myristate-13-acetate (PMA) (table 4.3). First, the amount of human interleukin-2 mRNA present in the test sample was measured using human interleukin-2 Quantikine[®] mRNA assay. Then, expected mRNA values for a two-fold dilution series were calculated. Linearity of the assay was determined by comparing the observed mRNA results from the serially diluted sample with the calculated, expected results. As shown, excellent linearity is achieved with this assay and observed mRNA concentrations correlate well with expected values.

In order to assess the precision of the Quantikine[®] mRNA assay, mRNA levels were quantified from samples containing known concentrations of a gene-specific mRNA (tables 4.4 and 4.5). For intra-assay precision, three samples containing known concentrations of a human interleukin-8 calibrator mRNA were assayed twenty times on one microplate (table 4.4). Inter-assay precision was also evaluated by assaying three samples containing known concentrations of human interleukin-8 calibrator mRNA in twenty separate assays (table 4.5). The percents coefficient of variation for both intra and inter-assay precision were all below six percent.

Table 4.3 Human interleukin-2 mRNA levels from HuT-78 lymphoma cells stimulated with phorbol 12-myristate-13-acetate (PMA). Cells were stimulated with ten nanograms PMA per millilitre culture medium for three hours. Total RNA was isolated using a modification of the guanidine isothiocyanate-based phenol/chloroform extraction described by Chomczynski & Sacchi (1997) prior to quantifying the interleukin-2 specific mRNA levels with the human interleukin-2 Quantikine[®] mRNA kit. The following abbreviations are used in this table: amol, attomole; ml, millilitre; ng, nanogram(s).

<i>Sample</i>	<i>Sample amount (ng)</i>	<i>Expected amount (amol/ml)</i>	<i>Observed amount (amol/ml)</i>
Stimulated HuT-78 lymphoma cells	2,000	—	69.27
	1,000	34.64	37.34
	500	17.32	17.66
	250	8.66	8.11
	125	4.33	4.58

Table 4.4 Intra-assay precision for the human interleukin-8 Quantikine[®] mRNA kit. The following abbreviations are used in this table: amol, attomole; ml, millilitre.

<i>Sample number</i>	<i>1</i>	<i>2</i>	<i>3</i>
Number of repetitions	20	20	20
Mean mRNA amount (amol/ml)	128	209	373
Standard deviation (SD)	3.6	4.3	6.4
Coefficient of variation (%)	2.8	2.0	1.7

Table 4.5 Inter-assay precision for the human interleukin-8 Quantikine® mRNA kit.

The following abbreviations are used in this table: amol, attomole; ml, millilitre.

<i>Sample number</i>	<i>1</i>	<i>2</i>	<i>3</i>
Number of repetitions	20	20	20
Mean mRNA amount (amol/ml)	126	206	367
Standard deviation (SD)	7.1	7.9	15.7
Coefficient of variation (%)	5.6	3.8	4.3

4.2.3.4 Examples

Determining the expression, or fluctuations in the expression, of specific mRNA species during certain disease states, treatments, or stresses can increase our understanding of specific molecular interactions, as well as biological systems in general. This is exemplified in the study by Borovikova *et al.* (2000), where the molecular mechanism of acetylcholine inhibition of human interleukin-6 and 1β expression in macrophages was investigated by measuring mRNA levels. In this study, mRNA expression of both interleukin-6 and 1β increased when macrophages were stimulated with lipopolysaccharide or lipopolysaccharide and acetylcholine. In contrast, the protein levels for both cytokines decreased in a dose-dependent manner when macrophages were treated with lipopolysaccharide and acetylcholine. The discrepancy between the induced levels of mRNA expression and the amount of secreted protein led to the conclusion that acetylcholine inhibition must be occurring by a post-transcriptional mechanism, rather than at the level of transcription. Similar experimental designs were used to ascertain the modulatory effects of serotonin on human interleukin-6 production in vascular smooth muscle cells (Ito *et al.*, 2000) and, also, to examine the effects of cystatins on the expression of interleukin-6 mRNA in human gingival fibroblasts (Kato *et al.*, 2000). These varied examples demonstrate how determining mRNA levels for particular molecules can provide valuable information when trying to clarify the intricacies of biological systems.

4.2.3.5 Discussion

As biotechnology has developed and diversified, several new and unique methods have emerged for quantifying mRNA. In addition to the Quantikine® mRNA assay, a wide variety of other commercially available kits also utilise a microplate-based mRNA detection format. Some of these include the QuantiGene™ Expression kit and the clinically applicable VERSANT® RNA assays from Bayer Diagnostics (Leverkusen Germany), the Invader® RNA assay from Third Wave Technologies, Inc., (Madison, WI, USA), the Xpress-Screen® kit (based on Hybrid Capture® 2 technology licensed from Digene Corporation, Gaithersburg, MD, USA) from Applied Biosystems (Foster City, CA, USA), and the Array-Plate™ Multiplexed Molecular Profiling system from High Throughput Genomics, Inc. (Tucson, AZ, USA).

Despite the development of these newer microplate-based techniques for mRNA quantification, the traditional methods for mRNA detection, such as Northern blot analysis, ribonuclease protection assay, and RT-PCR, are still employed by many researchers. A number of disadvantages, however, are associated with each of these techniques when compared to the newer microplate-based assays. Northern blot analysis only allows semi-quantitative determination of mRNA levels, is labour-intensive, takes several days to perform, and is unsuitable for high sample-throughput or for multiple mRNA analyses. Ribonuclease protection assay and RT-PCR technologies have been improved in recent years, yet these methods still have drawbacks. For example, ribonuclease protection assays are now commercially available that allow the use of cell lysate samples (*e.g.*, Direct Protect™ Lysate RPA from Ambion, Inc., Austin, Texas, USA). However, even with the elimination of the RNA isolation step, ribonuclease protection assay technology still requires the use of polyacrylamide gel electrophoresis and typically utilises radioactively labelled probes. Real-time RT-PCR now makes use of fluorescent probes, such as TaqMan® probes or molecular beacons, which allow the simultaneous amplification of an internal control, as well as the target of interest. One clear advantage of these fluorescent-based mRNA reporter systems is that they facilitate direct sample to sample comparisons, which could not be accomplished by using traditional PCR. Some concerns in these assays, however, are the efficiency of the reverse transcription step, the lack of consistency in exponential amplification, and whether both the control and mRNA targets have the same amplification efficiency. Additionally, real-time RT-PCR requires the use of special instrumentation, which may prevent this technology from being accessible to all labs. Furthermore, even though RT-PCR offers the benefit of increasing the copy number of the target to be detected, this method is susceptible to false-positive results due to contaminating amplicons from prior amplifications. In short, these traditional mRNA quantification techniques have some drawbacks that were addressed by the development of microplate-based assays.

One of the most distinct advantages of the new, microplate-based mRNA quantification technologies is the utilisation of signal amplification instead of target amplification, which as indicated can give misleading results. The Quantikine® mRNA assay is unique in its use of a chemical reaction that both amplifies the signal generated from the alkaline phosphatase-conjugated detection antibody and leads to a colourimetric end product. The Xpress-Screen® kit also makes use of an alkaline phosphatase-conjugated antibody as a detector; however, the antibody recognises RNA/DNA hybrids and the end result is luminescent signal detection (Aristarkhov *et al.*, 2000, and product literature). Bayer Diagnostics' kits utilise branched DNA, in conjunction with alkaline phosphatase (Warrior *et al.*, 2000, and product literature), and a luminescence-generating substrate for signal amplification (Urdea, 1994). The branched DNA reporter probes consist of a tree-like structure and each branch of the tree can bind up to three alkaline phosphatase reporter molecules (see chapter 3). This design allows the signal from each mRNA molecule to be significantly enhanced. The Invader® RNA assay uses an alternate chemistry and is based on the cleavage of an oligonucleotide that contains

a fluorescence resonance energy transfer, or FRET, probe (Allawi *et al.*, 2001; Kwiatkowski *et al.*, 1999). Once the FRET probe is cleaved, the fluorophore and quencher are separated, and a fluorescent signal is generated. Finally, the Array-Plate™ technology is based on ribonuclease protection assay principles and uses peroxidase-conjugated detection probes and luminescence to quantify analytes arrayed in a well (Felder & Kris, 2001; Seligmann, 2001). One drawback of many of these assays is the requirement for specialised fluorescent or luminescent detection equipment. In contrast, Quantikine® mRNA assay results can be read on a conventional microplate reader; thus reducing the cost to perform these assays. This characteristic perhaps makes the Quantikine® mRNA assay affordable to a greater number of labs not set up for molecular biology research.

Another feature of the microplate-based mRNA assay platform is its suitability to high-throughput screening. Each of the described assays utilise either a 96 or 384-well microplate, with most having the ability to measure multiple analytes on the same plate, or even within the same well. Furthermore, all of the assays are validated for use with cell lysate samples. The Array-Plate™ Multiplexed Molecular Profiling assay stands out among the rest of the mRNA detection assays by being able to measure up to 16 different mRNA species in one well (Seligmann, 2001). Invader® RNA assays can measure two analytes per well by selecting probes that fluoresce at different wavelengths. In contrast, the Quantikine® mRNA assay, the branched DNA assay, and the Xpress-Screen® kit can measure multiple analytes on the same plate, but not within the same well.

Perhaps of greatest concern, however, to any investigator is the sensitivity of these various microplate-based mRNA assays. The Quantikine® mRNA kit, for instance, has a sensitivity of 0.1 attomole, which is equivalent to 60,000 copies of mRNA. Similarly, the Array-Plate™ Multiplexed Molecular Profiling assays have the ability to detect 0.25 attomole, or 150,000 copies, of mRNA (product literature). Even though these levels of sensitivity will accommodate the measurement of mRNA in a variety of circumstances, other assays are available with higher detection sensitivities. For example, the assays using branched DNA are able to detect 0.005 attomole (3,000 copies) of a particular mRNA species (Shen *et al.*, 1998a), while, the Invader® RNA assay and the Xpress-Screen® assay can detect 0.01 attomole (6,000 copies) (Kwiatkowski *et al.*, 1999, and product literature). Also, the Invader® RNA assay has the added advantage of being able to differentiate between splice variants, as well as mRNA species that exhibit a high degree of homology (Kwiatkowski *et al.*, 1999, and product literature).

In summary, Quantikine® mRNA assay shares many of the same features, such as signal amplification, utilisation of cell lysate samples, and high-throughput capability, as other currently available microplate-based mRNA quantification assays. However, when compared to other methods, the Quantikine® mRNA assay offers some clear advantages by obviating the use of radioactivity, polyacrylamide gels, electrophoresis, and RT-PCR and by including a colourimetric detection signal requiring commonly available instrumentation. Even though the Quantikine® mRNA kit has many advantages over other technologies, the assay could no doubt be improved in two specific areas: (i) increased sensitivity and (ii) the ability

to multiplex within the same well. Whether this technology would be based on fluorescence or other available tools remains to be determined.

Investigators today have access to a variety of mRNA detection and quantification technologies and, of course, each technology offers both advantages and disadvantages, making the selection of suitable reagents for a particular application at times difficult. However, from the perspective of ease of use and lost entry cost, the Quantikine® mRNA kit offers an affordable and reliable approach to mRNA quantification.

4.3

PCR-based techniques

4.3.1

ADGE – Amplified differential gene expression

4.3.1.1 Introduction

Methods for analysing gene expression use PCR, sequencing or hybridisation-based approaches to measure inherent differences in mRNA expression. None of them magnifies the magnitude of expression differences prior to displaying them. Amplified differential gene expression (ADGE) is designed to magnify the ratio of a gene between two samples before displaying them and was described for the first time in 2001 (Chen *et al.*, 2001). Recent efforts have been sought to advance the utility of ADGE by coupling it with DNA microarray analysis. Overall, ADGE either alone or in series with DNA microarray seeks to improve the accuracy and sensitivity of gene expression profiling.

4.3.1.2 Principles and basics

The scheme of the ADGE technique is shown in figure 4.45. Two different nucleic acid samples are selected, one as control, the other as tester. After synthesis from total RNA or mRNA, the control and tester cDNA are cut with the restriction enzyme *Taq* I. This generates three types of DNA fragments depending on the number of *Taq* I sites on any particular cDNA: a type A fragment with a *Taq* I site at one end, a type B fragment with *Taq* I sites on both ends and a type C fragment without a *Taq* I site at either end. The *Taq* I fragments of control DNA and tester DNA are ligated to the CT adapter and the TT adapter oligonucleotides, respectively. The sequences of the CT adapter and the TT adapter are listed at the bottom of figure 4.45. The adapter-linked control and tester DNA samples are mixed in equal amounts, denatured and annealed. After filling in the ends, the reassociated DNA is amplified using a pair of CT primers complementary to the CT adapter or a pair of TT primers complementary to the TT adapter. The sequences of the CT primer and the TT primer with the selective nucleotides CCAC are also listed at the bottom of figure 4.45. The PCR products are separated on gels or hybridised onto DNA microarrays. The bands on gels displaying differences between

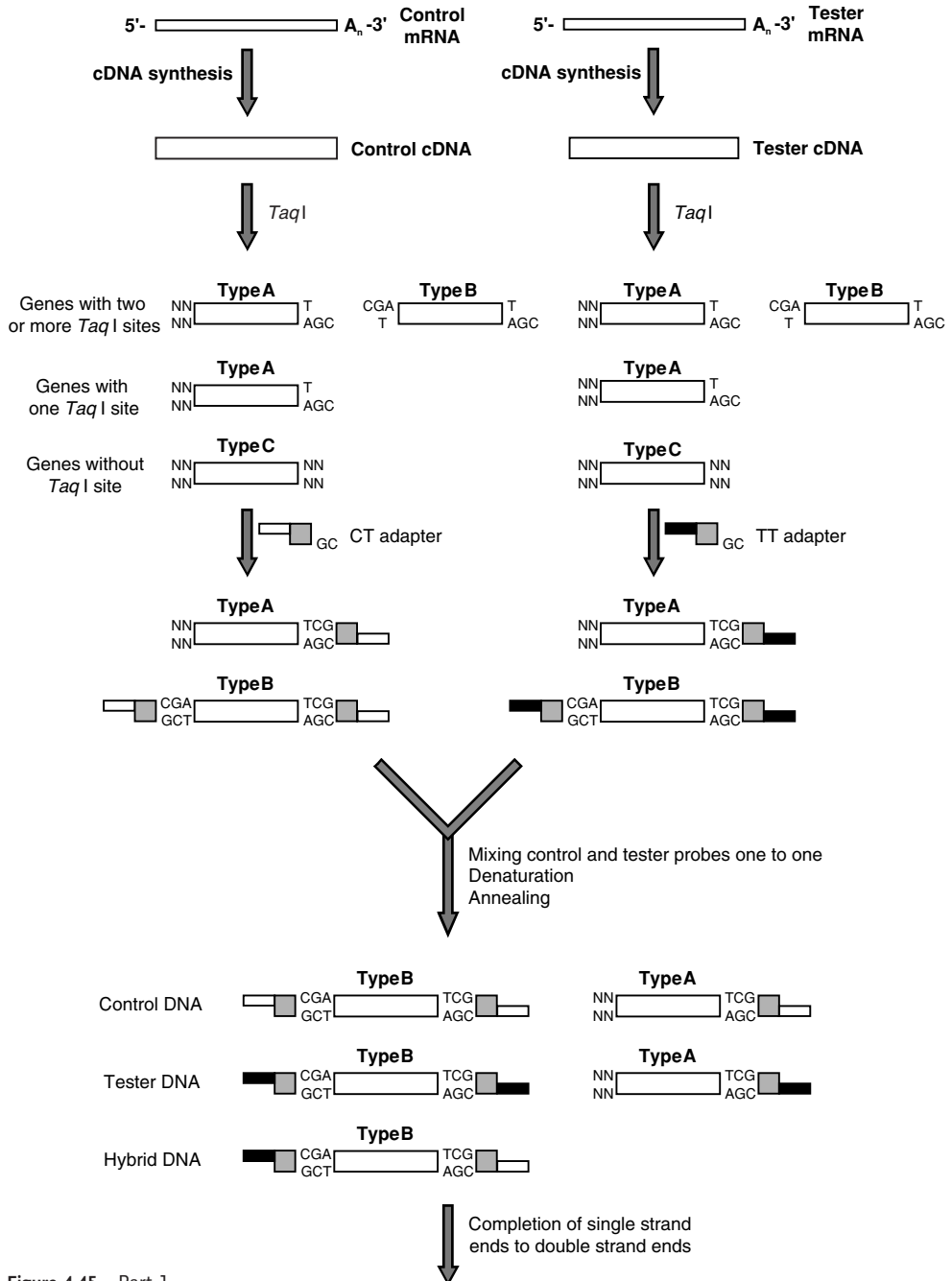
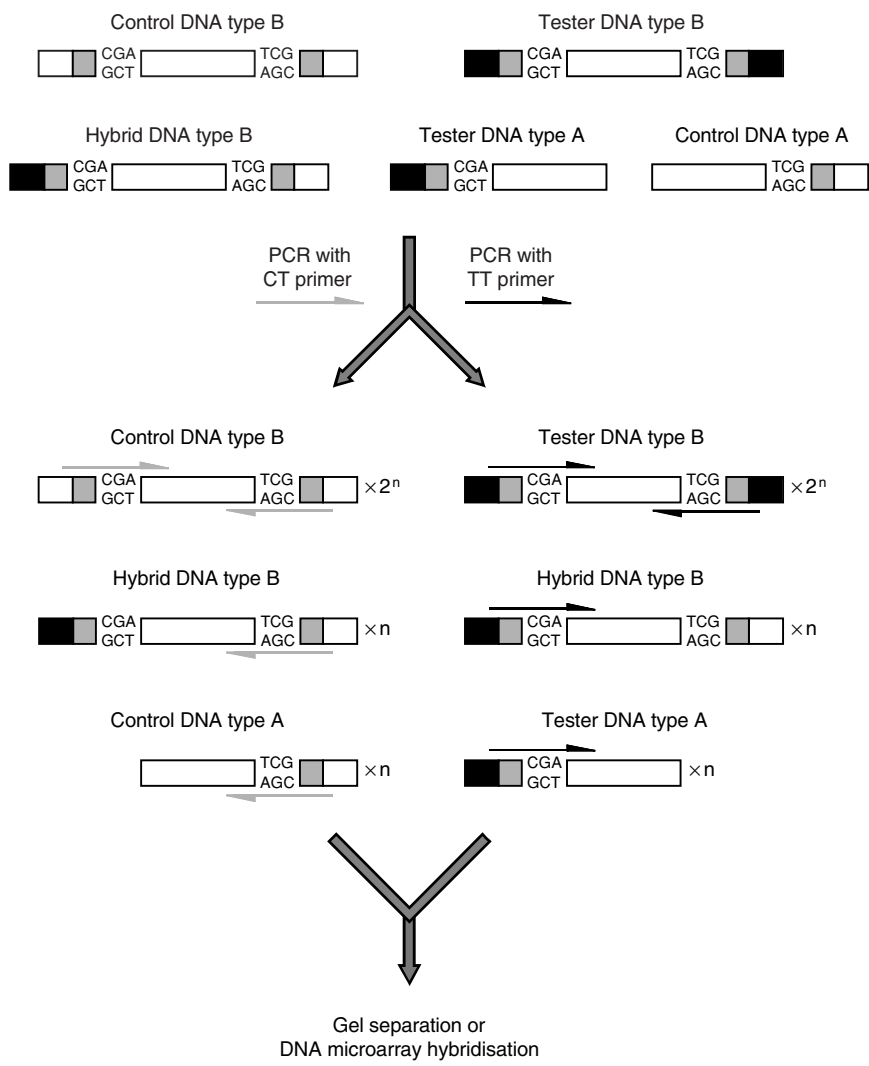


Figure 4.45 Part 1



CT primer: G TCA GTT GAA GGA AGC **ACG ACC AC**

CT adapter: TGA GGG TCA GTT GAA GGA AGC **A**
CCT TCG **TGC**

TT primer C AGT TCA AGC GGA AGC **ACG ACC AC**

TT adapter GTG TCC AGT TCA AGC GGA AGC **A**
CCT TCG **TGC**

Figure 4.45 Part 2

- ◀ **Figure 4.45 Schematic representation of the amplified differential gene expression (ADGE) method.** Double-stranded cDNA of control and tester samples is cut with the *Taq* I restriction enzyme. Digestion with the restriction enzyme *Taq* I generates three types of fragments: type A with a *Taq* I site at one end, type B with *Taq* I sites at both ends, and type C without a *Taq* I site at either end. The *Taq* I fragments of control DNA and tester DNA are ligated to the CT adapter and TT adapter oligonucleotides, respectively. After reassociation of the adapter-linked control and tester DNA, three types of duplexes for the type B fragments are formed: control DNA, tester DNA and hybrid DNA. The ratio of control and tester DNA is quadratically magnified. After filling in the single-stranded ends of the adapters, reassociated DNA is amplified using a pair of CT primers and a pair of TT primers. CT primers amplify control type B DNA exponentially while hybrid type B DNA and control type A DNA are amplified in a linear fashion. Equally, TT primers amplify tester type B DNA exponentially while producing linear amplification of hybrid type B DNA and tester type A DNA. The PCR products are separated on gels or hybridised to DNA microarrays. The sequences of the adapters and the primers are listed at the bottom of this figure. The *Taq* I restriction site is highlighted and the selective nucleotides are underlined.

the two samples are isolated and the corresponding genes are identified by sequencing.

The integration of DNA reassociation and PCR amplification makes it possible to obtain quadratic magnification of the ratio of an mRNA between two samples. DNA reassociation occurs when the control and tester DNAs are mixed in equivalent amounts, denatured and annealed together. After DNA reassociation, three different duplexes of type B fragments are formed: control DNA with the CT adapters on both ends, tester DNA with the TT adapters on both ends, and hybrid DNA with the CT adapter on one end and the TT adapter on the other end. The relative amount of the three types of duplexes for each gene is theoretically governed by the algebraic formula

$$(a + b) \times (a' + b') = a \times a' + b \times b' + a' \times b + a \times b'$$

where a is the number of sense strands of one mRNA in the control sample, b is the number of sense strands of the mRNA in the tester sample, a' is the number of antisense strands of one mRNA in the control sample, b' is the number of antisense strands of the mRNA in the tester sample, $a \times a'$ is the number of double strands of the mRNA in the control sample, $b \times b'$ is the number of double strands of the mRNA in the tester sample and $a' \times b$ or $a \times b'$ respectively are the number of double strands of the hybrid DNA.

For example, for a gene overexpressed five-fold in tester over control, $(b \times b') / (a \times a') = 5$. Thus, the formula is

$$(a + 5 \times b) \times (a' + 5 \times b') = a \times a' + 25 \times b \times b' + 5 \times a \times b' + 5 \times a' \times b.$$

After DNA reassociation, the ratio of $(b \times b')$ to $(a \times a')$ increases from five to twenty-five. If expression of another gene is, for example, reduced ten times in the tester, then the formula is

$$(10 \times a + b) \times (10 \times a' + b') = 100 \times a \times a' + b \times b' + 10 \times a \times b' + 10 \times a' \times b.$$

In this case, the ratio of $a \times a'$ to $b \times b$ increases from ten to one-hundred after reassociation.

Although the ratio of control DNA and tester DNA has been magnified quadratically after DNA reassociation, control and tester DNA have not yet been separated from each other or from hybrid DNA. This is achieved by PCR. The CT primers amplify control DNA exponentially and hybrid DNA linearly since control DNA has the CT adapter on both ends while hybrid DNA contains the CT adapter on one end only. Equally the TT primers amplify tester DNA exponentially and hybrid DNA linearly. After at least 20 cycles of PCR, the exponentially amplified control or tester DNA is a million times more abundant than the linearly amplified hybrid DNA. Thus, the PCR step both represses the hybrid DNA and separates control DNA and tester DNA from each other. The ratio of type A control DNA to type A tester DNA does not change after DNA reassociation. Type A DNA is also linearly amplified and, thus, repressed by PCR.

4.3.1.3 Examples

After screening the human leukaemic cell lines HL60 and HLA60/ADR (a HL60 cell line resistant to adriamycin) with three pairs of the CT and TT primers, four up-regulated genes and one down-regulated gene were identified (Chen *et al.*, 2001). Among the four up-regulated genes, the human DNA-dependent protein kinase catalytic subunit (*PRKDC*) and the human multidrug resistance-associated protein (*MRP*) were known from previous studies to be over-expressed in HLA60/ADR cells (Shen *et al.*, 1998b). The ratios of the HLA60/ADR to HL60 cells for *PRKDC* and *MRP* were detected as 3.8 and 47.6, respectively, by ADGE, while compared to 2.2 and 9.8 by reverse transcription-based PCR (Chen *et al.*, 2001).

The scatter-plot of the combination of the ADGE method with the microarray technique has a wider distribution on both sides of the diagonal line representing a one-to-one ratio than that of regular microarray (figure 4.46, unpublished data). The ratios of gene expression were magnified both for up and down-regulated genes. As a result of quadratic magnification, 535 genes with two-fold or greater alterations in expression were detected on an array of 10,000 genes using this modified ADGE technique, compared to 152 such genes with regular microarray experiments. The scatter-plots also show that the signal intensities of most genes were elevated in ADGE combined with the microarray compared to their intensity in a regular microarray experiment. There were 7,552 spots with signal intensities over 1,000 on both fluorescence channels (Cy3 and Cy5) in ADGE-combined microarray, compared to 4,040 such spots in regular microarray.

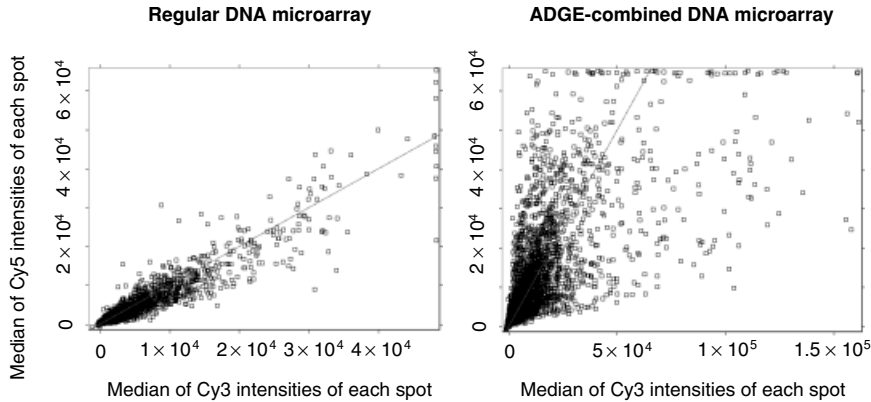


Figure 4.46 Scatter plots of regular microarray (left panel) and ADGE-combined microarray (right panel). Cell lines HL60 and HLA60/TLK286 (a HL60 cell line resistant to the pro-drug TLK286) were used in the experiment shown in the figure. The diagonal lines represent the one-to-one ratio of the two fluorochromes Cy5 and Cy3. The expression of genes above the line is up-regulated while those below the line is down-regulated. The FairPlay™ microarray labelling kit (Stratagene, La Jolla, California, USA) was used for regular microarray analyses with Cy3 dye incorporated in HL60 cDNA and Cy5 dye in HLA60/TLK286 cDNA. The ADGE-combined microarray procedure was

implemented following the scheme shown in figure 4.45. After the CT adapter and TT adapter oligonucleotides were ligated to the HL60 cDNA and HLA60/TLK286 cDNA, respectively, the HL60 and HLA60/TLK286 cDNA were re-associated. The reassociated DNA was used as a template for the PCR reaction. The CT primers amplified HL60 DNA with incorporation of Cy3-labelled dCTP while the TT primers amplified HLA60/TLK286 DNA with incorporation of Cy5-labelled dCTP. The labelled DNA was hybridised onto a DNA microarray with 10,000 genes. The microarray data were analysed and normalised with the software ImaGene 4.1 and GeneSight 3.0 (Biodiscovery, Los Angeles, California).

4.3.1.4 Discussion

The structure of the adapters and the primers is critical to ensure the quadratic amplification of a DNA target ratio. The adapters are composed of long and short oligonucleotides. In order to form hybrid DNA molecules, the short oligonucleotides are the same in both CT and TT adapter oligonucleotides. The optimum complementary region is seven nucleotides. If it is too short, the adapters may not be stable. If it is too long, cross annealing becomes possible. The adapters have cohesive ends complementary to the restriction site sequence of the restriction endonuclease *Taq* I. The CT and TT primers consist of regions complementary to the adapter oligonucleotides, regions corresponding to *Taq* I sites and selective nucleotides. The length of regions with different sequences between CT and TT primers should be sufficient to prevent cross annealing (at least ten nucleotides long). Although their sequences can be changed, the CT and TT adapter oligonucleotides and primers are designed so that CT primers amplify control DNA exponentially and hybrid DNA linearly, while TT primers amplify tester DNA exponentially and hybrid DNA linearly.

The ADGE scheme in figure 4.45 may be modified. For example, other four-base restriction enzymes other than *Taq* I may be used and are recommended for those mRNAs with only one *Taq* I site or no *Taq* I site. The PCR products can be visualised with either isotope or fluorescence labels. The number of selective nucleotides can be varied from zero to four, depending on the number of genes that are to be amplified in one PCR reaction.

The major advantage of ADGE is that it quadratically magnifies the ratio of a mRNA between two samples before displaying them, thus improving detection accuracy. ADGE thus raises the magnitude of the gene ratios above the threshold of detection error before displaying them, thus increasing the number of differentially expressed genes that can be identified. Therefore, ADGE identifies small differences in mRNA expression that might elude detection. It also renders more reliable detection of those genes with greater differences in expression.

4.3.2

DEPD® – Digital expression pattern display

4.3.2.1 Introduction

In contrast to so-called closed systems such as quantitative PCR or DNA microarrays that require prior sequence knowledge of the genes to be analysed, DEPD is an ‘open architecture’ system, which is not limited by a lack of prior sequence information. Most open architecture methods start with digestion of cDNA using restriction enzymes. DNA adapters are ligated to the resulting cDNA fragments, thereby creating binding sites for specific PCR primers. Finally, PCR reactions with specific complementary primers are performed and the resulting fragments are analysed.

All open architecture methods have limitations in determining the gene expression profiles of complex tissues such as the mammalian brain. It is estimated that 90 to 95 percent of all mRNA species may be present at a frequency of five or fewer copies per cell (Alberts *et al.*, 1989; Bonaldo *et al.*, 1996; Zhang *et al.*, 1997) and represent only 35 to 50 percent of the total cellular mRNA mass. The main difference between brain and other tissues resides in the number of very rare transcripts (less than one copy per cell), which is much higher in brain than in any other tissue (Chaudhari *et al.*, 1983; Hastie *et al.*, 1976; Kaplan *et al.*, 1978; Young *et al.*, 1976) (figure 4.47). In brain, this group actually comprises the majority of all mRNA species, even though it represents less than five percent of brain mRNA mass. Since it also contains most of the genes involved in neuronal signalling, a complete analysis of the transcriptome of the brain requires a technology that can detect and reliably quantify mRNAs present at abundances of less than one in 500,000 mRNA molecules. In addition, the method should be capable of resolving more than 50,000 different transcripts.

With this in mind, researchers at Biofrontera Pharmaceuticals AG developed a method, DEPD® (digital expression pattern display) (Hoffmann *et al.*, 1998), which meets the above requirements for profiling very rare transcripts expressed in the mammalian brain. DEPD® is currently being used to support projects

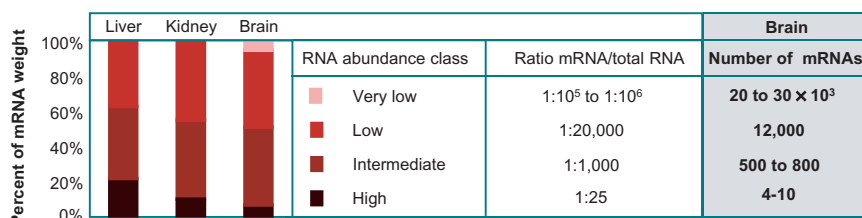


Figure 4.47 Concentration of mRNA molecules differs in various tissues, but can generally be divided into four different classes, from high to very low abundance. More than 80 percent of the distinct transcripts in the brain belong to the class of low or very low abundant mRNAs, like receptors for neurotransmitter, ion-channels, uptake carriers or other genes involved in neuronal transmission that have in the past been proven to be suitable drug targets.

aimed at discovering potential new drug targets, or new uses for known drugs, for the treatment of central nervous system (CNS) disorders. It is also used to profile new drugs with the goals to predict their action, side effects, and toxicity in the brain *in vivo*. The capabilities of the DEPD® technology now make this a realistic alternative to more conventional strategies in many fields of drug discovery research.

4.3.2.2 Principles and basics

DEPD® is an automated PCR-based procedure, its principle is outlined in figures 4.48.A and 4.48.B. DEPD® can be applied to any eukaryotic species and any tissue, and detects both known and unknown gene transcripts. Based on the very high sensitivity and exceptional resolving power of the method, its major advantages become most apparent when dealing with the highest complexity of gene expression found in biology – the mammalian brain.

In the performance of a typical DEPD® experiment, two to ten micrograms of total RNA are processed through several steps, including cDNA synthesis, restriction enzyme digestion, adapter ligation, and PCR amplification. High resolution is achieved by subdividing the cDNA species present in the sample into over 3,000 groups arrayed in microtiter dishes. Each group is then analysed by capillary electrophoresis, allowing the separation of up to 300,000 distinct DNA fragments.

After the extraction of total RNA, double-stranded cDNA is synthesised. Through enzymatic digestion of the cDNA with so-called class IIS restriction enzymes, short DNA fragments are generated, containing single-stranded overhangs in all possible sequence combinations. Afterwards, specific DNA adapter molecules are ligated and, in two subsequent steps, 3,072 PCR reactions are performed. Improved results are obtained if the 3'-poly(A) sequences are removed during the process and replaced with a mixed-sequence adapter. Since PCR primers are complementary to the adapters, it is possible to perform the PCR reactions at high temperatures, resulting in high reproducibility and reliability. This depicts a crucial difference in comparison to the original differential display RT-PCR method invented by

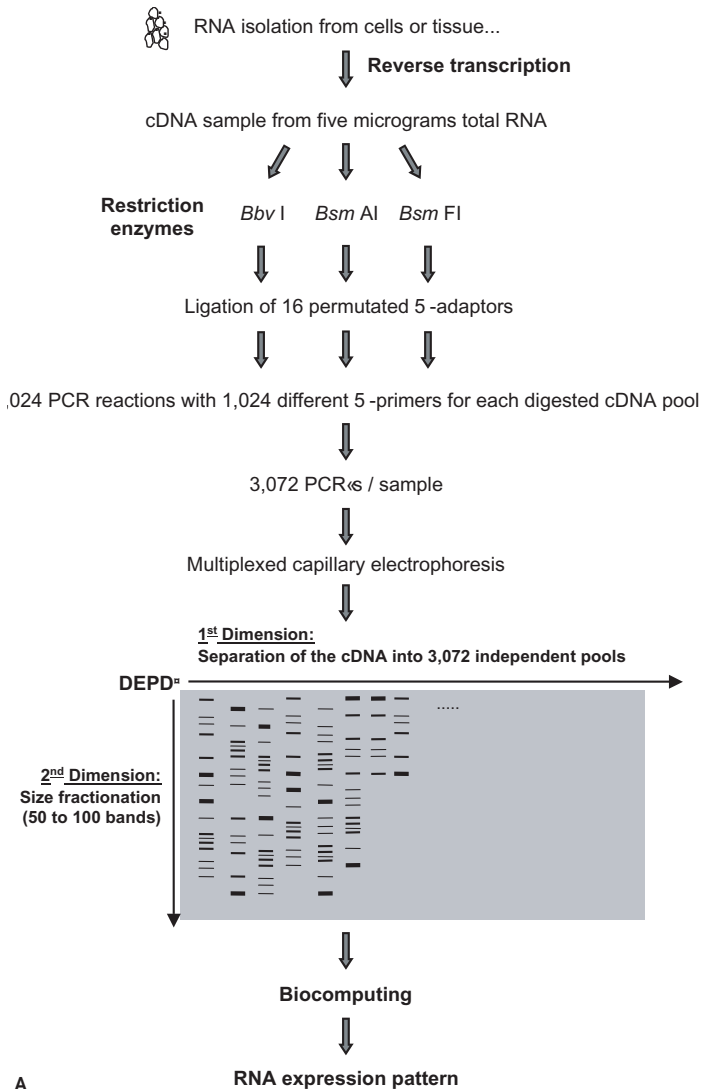


Figure 4.48 The principle of a typical DEPDP[®] experiment is schematically illustrated.

A) Flow-chart of the DEPDP[®] process. Total RNA is isolated from the test tissue or cells, then transcribed into cDNA, and digested with class IIS restriction enzymes. The next steps of the DEPDP[®] technique involve the ligation of specific adaptors to the DNA fragments and the generation of 3,072 different PCR reactions. Hereby, a separation of the original cDNA into 3,072 different sub-pools is achieved. PCRs are usually performed with fluorescent dye-labelled

primers and afterwards separated by capillary electrophoresis in a 96-well format. Fifty to one hundred fragments can be separated by capillary electrophoresis in each of the 3,072 PCR reactions, thus resulting in a separation power of almost 300,000 cDNA fragments. This high resolution is a necessary requirement for the complete profiling of gene expression in the brain. Sequence analysis of all fragments obtained with mouse brain cDNA revealed a redundancy of approximately 2.5.

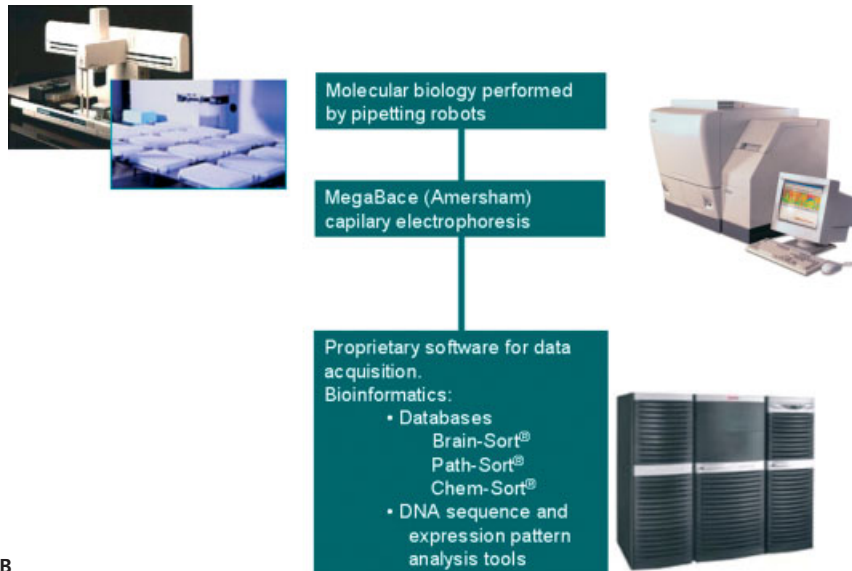


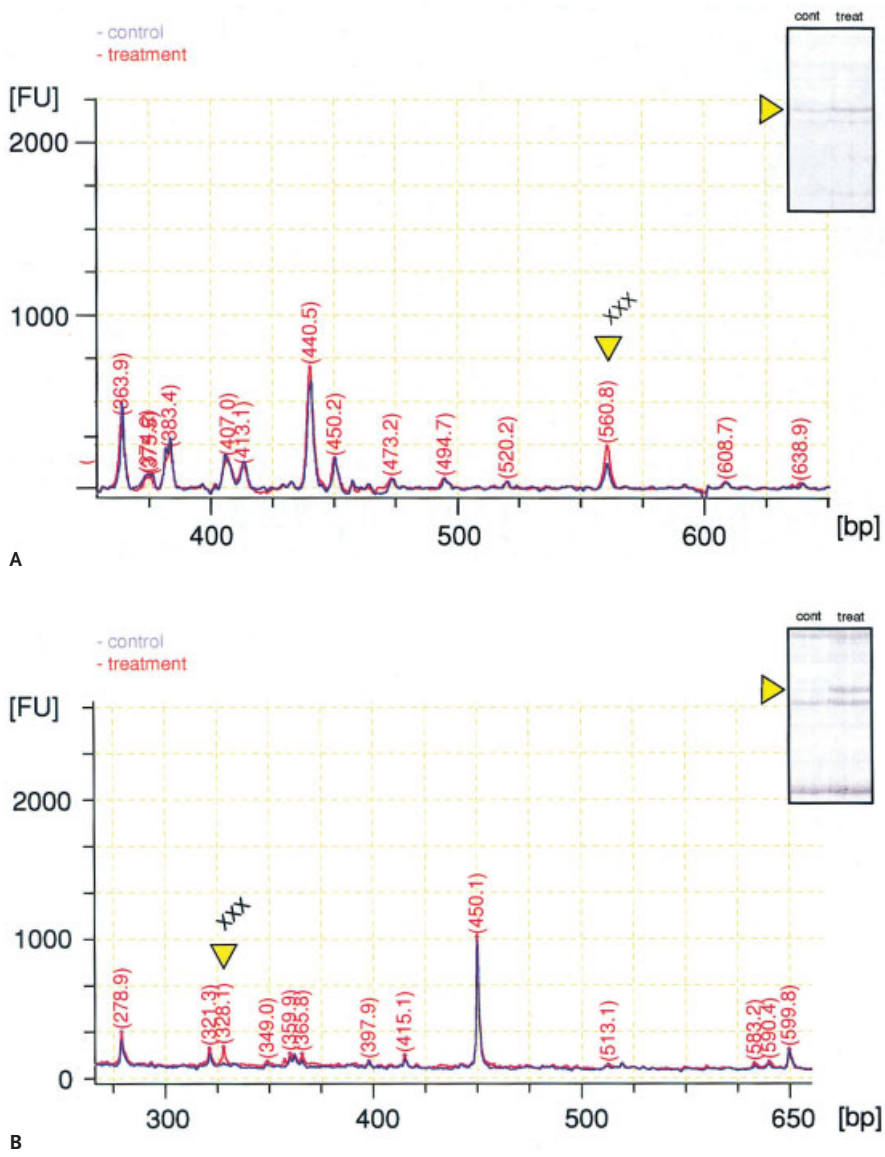
Figure 4.48 B) The automated, industrial scale mRNA expression analysis workflow. After the isolation of mRNA, pipetting robots perform most of the experimental steps. The PCR pools are analysed with 96 capillary sequencers, and raw data are captured automatically. Further data analysis results in constantly growing databases, linking the gene expression patterns to sequence and functional information.

Liang & Pardee in 1992, in which the use of randomised PCR primers requires low temperatures, leading to a high error rate.

Following PCR amplifications, the samples are separated by electrophoresis with 96 capillary sequencers. Capillary electrophoresis allows the separation of up to 100 fragments per lane, resulting in a separation power of about 300,000 fragments (figure 4.48.A). Some typical results produced with DEPD® are shown in figures 4.49.A and 4.49.B. In these examples, total brain from a group of drug treated animals is compared with a control group, demonstrating the high reproducibility of the method. Each of the many thousand distinct fragments that are produced by DEPD® represents an expressed sequence tag (EST), which can be identified by reference to Biofrontera's proprietary database, Brain-Sort®, containing the sequence information of all transcripts already analysed in one of Biofrontera's prior experiments. In practice, a DEPD® experiment of mammalian brain tissue produces over 100,000 different fragments.

Isolation and sequence analysis of all fragments obtained in a DEPD® experiment of brains from the mouse inbred strain C57BL/6 showed that DEPD® produces a redundancy of about 2.5. Therefore, 100,000 identified fragments correspond to 40,000 different transcripts in the cell.

The roughly 40,000 genes of the mammalian genome are assumed to produce at least twice the number of distinct transcripts, of which at least half are expressed in



the brain. Therefore, the 40,000 transcripts detected with DEP^D are likely to represent the majority of the transcripts that exist in brain.

Sequence analysis of all transcripts seen in the rodent brain revealed further that, at the time when this contribution was written, about 20 percent of all transcripts seen were not represented in public cDNA sequence databases. This illustrates a clear advantage of DEP^D over closed architecture systems like DNA microarrays, where such unknown genes cannot be analysed.

- ◀ **Figure 4.49 Drug-profiling using DEPD®.** Two typical results of a DEPD® experiment are shown. Rats were injected daily for five days with an anti-psychotic drug. Five days after the last injection, brain total RNA was isolated and a gene profile of the brain was generated. Animals that were given saline in the same way served as controls. Typical DEPD® PCR reactions were separated by capillary electrophoresis. Two short segments of the gene profiles, ranging from 300 base pairs to 650 base pairs are shown. The signals from the control animals are shown in blue, those of compound-treated animals in red. In both examples shown, a total of 13 peaks are visible. In twelve cases, the peaks from controls and treated animals are identical. Thus, the expression of the genes that correspond to the peaks are not changed by treatment with the drug. The two other peaks, one at 560 base pairs in **A**), and one at 328 base pairs in **B**), appear to be 1.8-fold and about 50-fold up-regulated in the animals treated with the compound, respectively. In the case that Biofrontera's database Brain-Sort® does not contain the corresponding sequence of a differentially regulated peak, this peak is confirmed on an additional polyacrylamide gel. The band of interest is then isolated from the gel and sequenced subsequently. FU means relative fluorescence units and bp stands for base pairs.

A relevant feature of DEPD® is the separation of a cDNA population into over 3,000 pools that needs to be achieved with a minimal redundancy between the pools. Thus, several thousand PCR reactions need to be performed to obtain a complete data set. This effort provides the unique separation power and sensitivity of the method, but it is also the reason why DEPD® cannot easily be done in a standard laboratory setting. As it is a proprietary technology that belongs to Biofrontera Pharmaceuticals AG, this company has made a considerable effort to transform the principle of DEPD® into a fully automated, industrial scale process, as illustrated in figure 4.48.B.

4.3.2.3 Power of DEPD® and statistical data analysis

In principle, DEPD® can detect every transcript, which is cut by one of the three type IIS enzymes used at a distance from the 3'-end of the RNA such that the resulting 3'-end fragment has a size that can be separated by capillary electrophoresis. This is the case for about 98 percent of all transcripts. Thus, within this group of 98 percent of all fragments, detection is limited only by the sensitivity of the method. To analyse the sensitivity, synthetic mRNAs were mixed into a brain RNA sample at increasing dilutions. The result was that DEPD® is capable of detecting mRNA species that are present at a molar ratio of less than one molecule in 750.000 molecules (data not shown), corresponding to less than one copy in every third cell.

The high resolution of DEPD® renders it possibly to identify about 40,000 distinct transcripts in the mammalian brain, considerably more than any other published method. The industrial set up (Figure 4.48.B) uses fluorescent labelling of each cDNA fragment, allowing automated, high-throughput data generation. Usually, the biological experiments providing the tissue for DEPD® analysis are performed twice independently at different times of the year. Each of the resulting two mRNAs is then analysed in duplicates, thus providing a basis for the statistical

appreciation of the results. Depending of the peak size, peaks typically display a size variance of only ten to twenty percent (figure 4.50). This deviation from the mean follows a Gauss distribution, thus allowing the number of false-positive peaks to be calculated for each experiment. If the cut-off for differential gene expression is set at a statistical significance of three times the sum of the standard deviations of two peaks that are being compared, less than 0.2 percent false-positive peaks will be obtained, and peak size differences can be detected with high statistical significance even if peak differences are smaller than a factor of 1.5. Biofrontera has developed a software, called DEPDstat, that automatically calculates all results including their statistical significances, plus the number of false-positives expected in each particular experiment. The program is also used to perform simulation studies on each data set in order to determine the number of false-negative peaks. The result is that more than 90 % of all transcripts that are differentially expressed in a data set will be detected in a single experiment, a series of experiments will not miss any data.

To achieve fast and efficient data handling and mining, a proprietary bioinformatics software platform has been established. For this purpose, Biofrontera has entered into a strategic partnership with a bioinformatics software provider, Infor-

Determine the precision of the expression measurement by comparing signal intensities from peaks that are expected to be identical

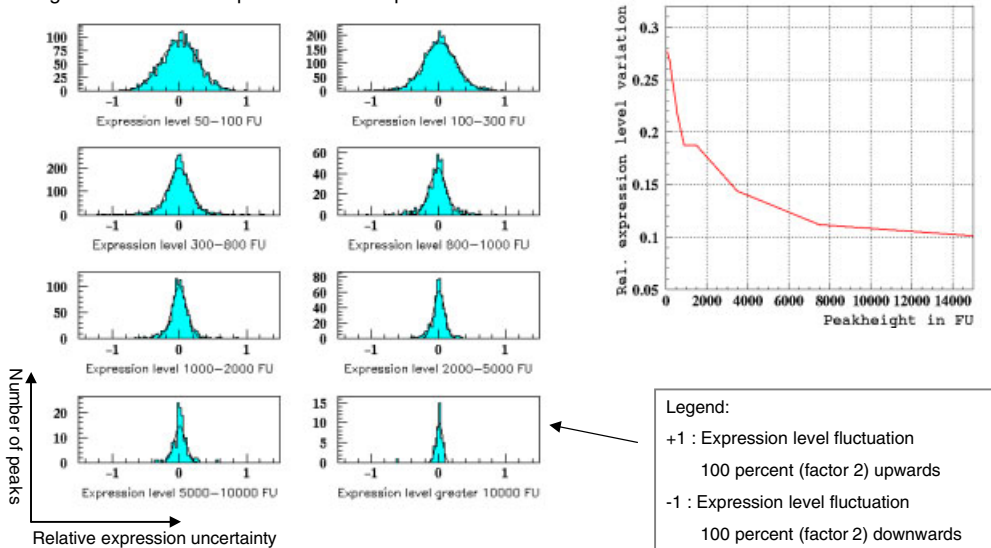


Figure 4.50 The variance in peak size determination in a typical DEPD[®] experiment is illustrated. Processing the same RNA twice and comparing the resulting peaks with one another recorded the data. Even though, all peaks should be identical in principle, minor deviations were present and were more pronounced

for small peaks. Therefore, such deviations are shown on the left for peaks within defined height ranges. The graph on the right displays the dependence of the mean variance from peak height. FU means relative fluorescence units and bp stands for base pairs.

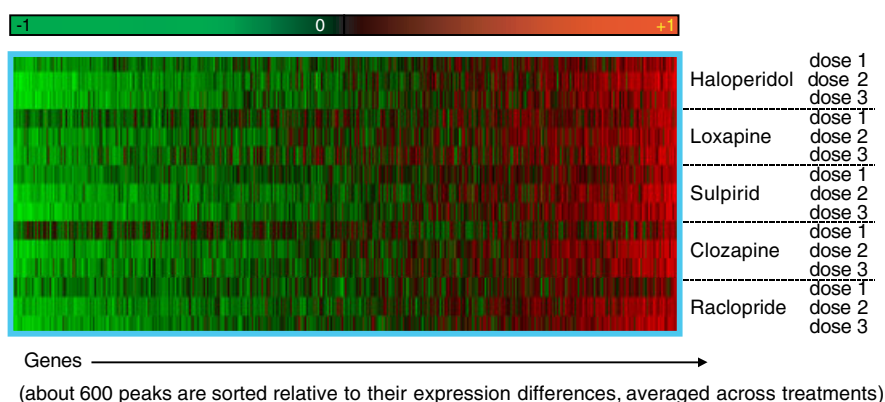
max, Inc., in Bethesda, USA. Informax has developed a software system called Genomax, which complements Biofrontera's in-house developments. This set-up, linked with a computing system consisting of a cluster of high-end computers, makes it possible to perform DEPD® and data analysis in an automated procedure at a high-throughput scale. This includes quality and process control of the raw data, as well as the quantitative analysis, and interpretation of the very large amount of gene expression data generated by DEPD®.

4.3.2.4 Examples

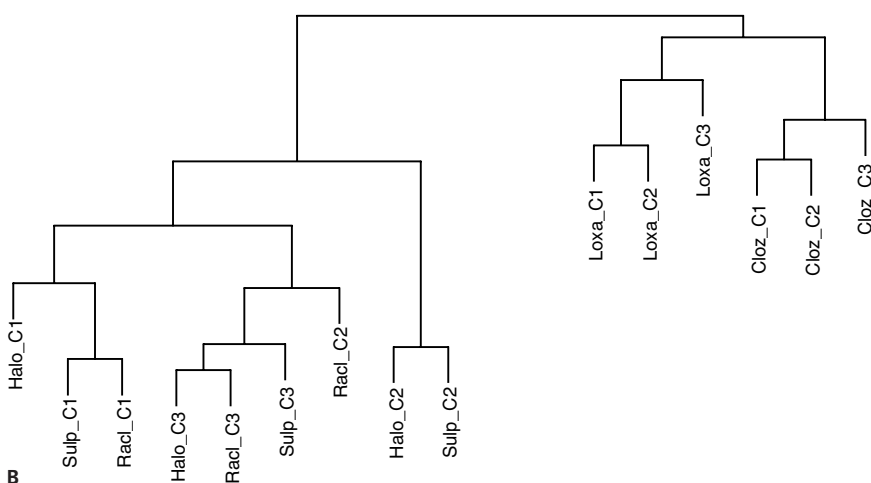
DEPD® is applied to projects aimed at three goals: (i) the discovery of potential new drug targets for CNS disorders, (ii) finding new uses for known drugs in CNS indications, and (iii) profiling new drugs in order to predict their actions, side effects and toxicity *in vivo*. It is believed that a thorough understanding of transcriptional regulation during a disease process or after drug treatment, together with a uniquely large data set for comparison, will provide insights into the pathological processes underlying brain diseases.

The analysis and the interpretation of the resulting gene expression profiles is the next crucial step following the DEPD® analysis itself. Rather than homing in on particular interesting genes early on, large data sets are first compared by cluster analysis using different clustering algorithms. This is done across multiple data sets, in order to compare different disease states with one another, different animal models for the same disease, or the action of different drugs.

As an example, the effects of five different anti-psychotic drugs on brain gene expression are shown in figures 4.51.A and 4.51.B. Figure 4.51.A depicts dose-dependent expression profiles, registered after daily injections of rats for eight days with three different doses for each drug. Shown is a set of 600 differentially expressed transcripts, ordered according to their expression differences. Each gene is represented as a thin vertical line, the colour indicates its status of differential regulation. The same genes are shown for all drugs and doses, and always arranged on top of one another. Dose 1 represents the lowest and dose 3 the highest dose of the drugs administered to the animals. Increasing doses result in increasing strengths of gene expression differences. At the next level of analysis, a hierarchical clustering of the patterns was performed (figure 4.51.B). This experiment showed that chemically diverse drugs like haloperidol, sulpiride, and raclopride, which are all used as typical anti-psychotic drugs in humans (dopamine 2 receptor antagonists), produce patterns of differential gene expression that closely resemble each another. Clozapine and loxapine, two drugs with a high degree of structural similarity but different action in humans (clozapine acts as an atypical anti-psychotic drug while loxapine is a typical one), by contrast, cause differences in gene expression differences that are distinct from those produced by the above drugs, but that are also not closely related to each another. This example may demonstrate the power of DEPD® to distinguish functional and structural similarities or differences between neuroactive compounds.



A



B

Figure 4.51 Dose-dependent expression profiles of rat brains after treatment with anti-psychotic drugs and hierarchical clustering of the data. A) Rats were injected daily for eight days with three doses of each drug, or with saline for the controls. For each drug, transcripts that were differentially expressed compared to the controls were determined. A set of 600 differentially expressed transcripts was ordered according to their expression differences. For each injection drug or concentration, the same transcripts are always shown on top of one another, the colour illustrates differential expression, with green illustrating down-regulation, red up-regulation. Dose 1 represents the lowest and dose 3 the highest dose of the drugs administered. The level of differential gene expression is

expressed as $[(h_{tr} - h_{co}) / (h_{tr} + h_{co})]$ with h as peak height, tr for signals of drug treated samples and co for signals of control samples, resulting in a scale that ranges from plus one to minus one, with zero being unchanged gene expression. **B)** A hierarchical clustering algorithm was applied to the data shown in **A)**. The result of the clustering illustrates the close relationship between typical anti-psychotic drugs with different chemical structures, and the difference seen between the chemically related drugs clozapine and loxapine, which display very different actions in humans. The further down on the graph the crossing point between two lines is, the closer is the similarity between the gene expression effects of the test drugs.

4.3.2.5 Discussion

DEPD® is a comprehensive, sensitive and highly reproducible display of the genes expressed in a tissue or cell, independent of prior sequence information. The above description illustrates its power and some potential applications. It may be the most sensitive method that currently exists. Its high sensitivity is based on the use of PCR on multiple pools of cDNA, each of which contains only a limited number of fragments that are co-amplified in parallel. Compared to array-based methods, DEPD® is about an order of magnitude more sensitive, probably because hybridisation-based detection cannot match the sensitivity of PCR. This, together with the advantages of an open architecture system, and the limitation in the number of genes that are present on current DNA microarrays, renders DEPD® clearly superior to array-based techniques for many applications.

A disadvantage of DEPD® is the fact that a very specialised laboratory setting is required to perform it routinely. Therefore, its application will always be limited to special questions in research, where its unmatched sensitivity and comprehensiveness are valuable. DEPD® will most likely never be able to match the speed and simplicity of DNA microarrays, which will therefore be superior for routine applications such as diagnosis.

DEPD® shows similar advantages when compared to other open architecture technologies, performed by other specialised companies, such as restriction enzyme analysis of differential sequences (READS™; Gene Logic, Inc., Berkeley, California USA), total gene expression analysis (TOGA; Digital Gene Technologies, Inc., La Jolla, California, USA), and GeneCalling (CuraGen Corporation, New Haven, Connecticut, USA). All these methods are based on splitting the cDNA into noticeably fewer pools (for READS™ and GeneCalling about 100, for TOGA between 256 and 768), which leads to greatly reduced resolution and sensitivity. Due to the principles of these methods, an increase in pool number would, for these techniques, merely increase redundancy, without reducing the complexity of each pool, and would therefore not lead to an increase in sensitivity and separation power.

DEPD® detects differences in mRNA abundance of less than 1.5-fold with high statistical significance. Even though the biological significance of such small changes in individual cell populations may be questionable, it should be remembered that transcription profiling experiments are frequently performed with tissues which contain rather complex mixtures of different cell types, such that differential gene expression effects may be diluted.

In this context, the crucial importance for the results of a DEPD® experiment (or any other transcriptional profiling) of the quality and reproducibility of the tissues analysed should be stressed. A major source of variability in the results stems from differences between the biological sources, even though they may be thought to be similar. Therefore, highly controlled animal maintenance and experimental conditions are crucial. For this reason, at Biofrontera every experiment is repeated at a different time of the year.

In the coming years, because of the increasing amount of data, the importance of bioinformatics and statistics will continue to grow. The interpretation of the

mass of gene expression data poses a major challenge of the future, and the examples shown above may give an impression about the bioinformatics power that is required to make effective use of gene expression analysis.

4.3.3

FDD – Automated fluorescent mRNA differential display

4.3.3.1 Introduction

How can a fertilised egg containing a complete set of genes unique to a species give rise to so many different cell types that will ultimately organise into the different tissues and organs that define each specific organism? This has been one of the most elusive questions in biology, as, to date, even complete sequencing of many genomes, from a few thousand base pairs for bacteria to over 3 billion base pairs for human, has yet to provide clues to the mystery of life. Of the estimated 30,000 to 40,000 genes embedded in our genome, only ten to fifteen percent are ‘turned on’ (expressed as mRNAs for protein synthesis) at any given time in each of our cells. Thus, interpretation of the genomic instruction will have to rely, at least in large part, on tools which can allow us to determine when and where a gene is to be turned on or off in a cell as it divides, differentiates and ages. Obviously, such tools are also important for the detection of when and where a seemingly precise interpretation of genomic instruction goes awry, which underlies many disease states such as cancer. RT-PCR-based differential display technology is one of the major tools that has already helped thousands of researchers in the world to interpret the genomic information (Liang *et al.*, 1992).

Differential display RT-PCR technology has been, and continues to be, one of the most reliable methods for gene expression analysis available to biomedical researchers. Since its invention in 1992, the number of publications using differential display RT-PCR has exploded to over 3,800, outnumbering the combined total of publications using competitive methodologies such as DNA microarrays, serial analysis of gene expression (SAGE) and representational difference analysis (RDA) (table 4.6). It is clear that the rapid and successful adoption of differential display RT-PCR has been largely attributed to the simplicity of the method. Simplicity ensures a higher probability of success and few artifactual differences caused by experimental errors.

Table 4.6 Impact of major technologies in gene expression analysis.

<i>Methods</i>	<i>Year invented</i>	<i>Medline hits*</i>
Differential display RT-PCR	1992	3,862
DNA microarrays	1995	1,196
Subtractive hybridisation	1985	886
Serial analysis of gene expression (SAGE)	1995	644
Representational difference analysis (RDA)	1993	288

* Medline searches conducted on January 14th, 2002, by PubMed at NCBI website (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>) using keywords depicting each methodology.

Essentially, starting from the RNA samples being compared, only two steps, reverse transcription and PCR, are needed before signals generated are analysed on a gel matrix. No second strand DNA synthesis, purification of cDNA, restriction enzyme digestion or adapter primer ligation, probe labelling and normalisation, hybridisation or washing steps are required, as each of these steps could introduce and amplify errors or lead to the loss of mRNAs being detected.

4.3.3.2 Differential display RT-PCR methodology

Differential display RT-PCR was developed by integrating three of the most simple, powerful and commonly used molecular biological methods, namely RT-PCR, DNA sequencing gel electrophoresis and cDNA cloning (Liang *et al.*, 1992). The principle of differential display RT-PCR is depicted in figure 4.52. First, mRNAs from cells are converted to first strand cDNAs using three individual anchored oligo(dT) primers that differ from each other at the last 3'-non-thymidine base. The use of these anchored primers enables the homogeneous initiation of cDNA synthesis at the beginning of the poly(A) tail for any given mRNA. The resulting three subpopulations of cDNAs are further amplified and labelled with either isotopes or fluorescent dyes by PCR in the presence of a set of second primers that are short and arbitrary in sequence. The length of an arbitrary primer is so designed that by probability each will recognise 50 to 100 mRNAs under a given PCR condition (Liang *et al.*, 1992). As a result, mRNA 3'-termini defined by any given pair of anchored primer and arbitrary primer are amplified and displayed by denaturing polyacrylamide gel electrophoresis. Side-by-side comparisons of such cDNA patterns between or among relevant RNA samples would reveal differences in gene expression. Differentially expressed cDNA bands can be retrieved, cloned and sequenced for further molecular characterisation. Based on the finding that each arbitrary primer would recognise its corresponding mRNA targets with a minimum of a seven bases match, mathematical models have been proposed to predict the relationship between the number of arbitrary primers and the coverage of expressed genes in any given eukaryotic cells (Liang *et al.*, 1995). These models have yet to be validated by saturation differential display RT-PCR screening using a large number of primer combinations. One of the major advantages of differential display RT-PCR is that it does not require any prior knowledge in mRNA sequences, making the gene screening systematic and non-biased.

4.3.3.3 Automation of fluorescent differential display (FDD) with digital readout

Although differential display RT-PCR is the most widely used method (based on the number of publications) for isolating differentially expressed genes, most differential display screenings take a shot-gun approach by using only a limited number of primer combinations, with which a few genes are identified and characterised. Theoretically, a genome-wide comprehensive differential display RT-PCR screening is possible (Liang *et al.*, 1994 and 1995), but would require hundreds of PCR reactions for each RNA sample being compared. Clearly, a high-throughput

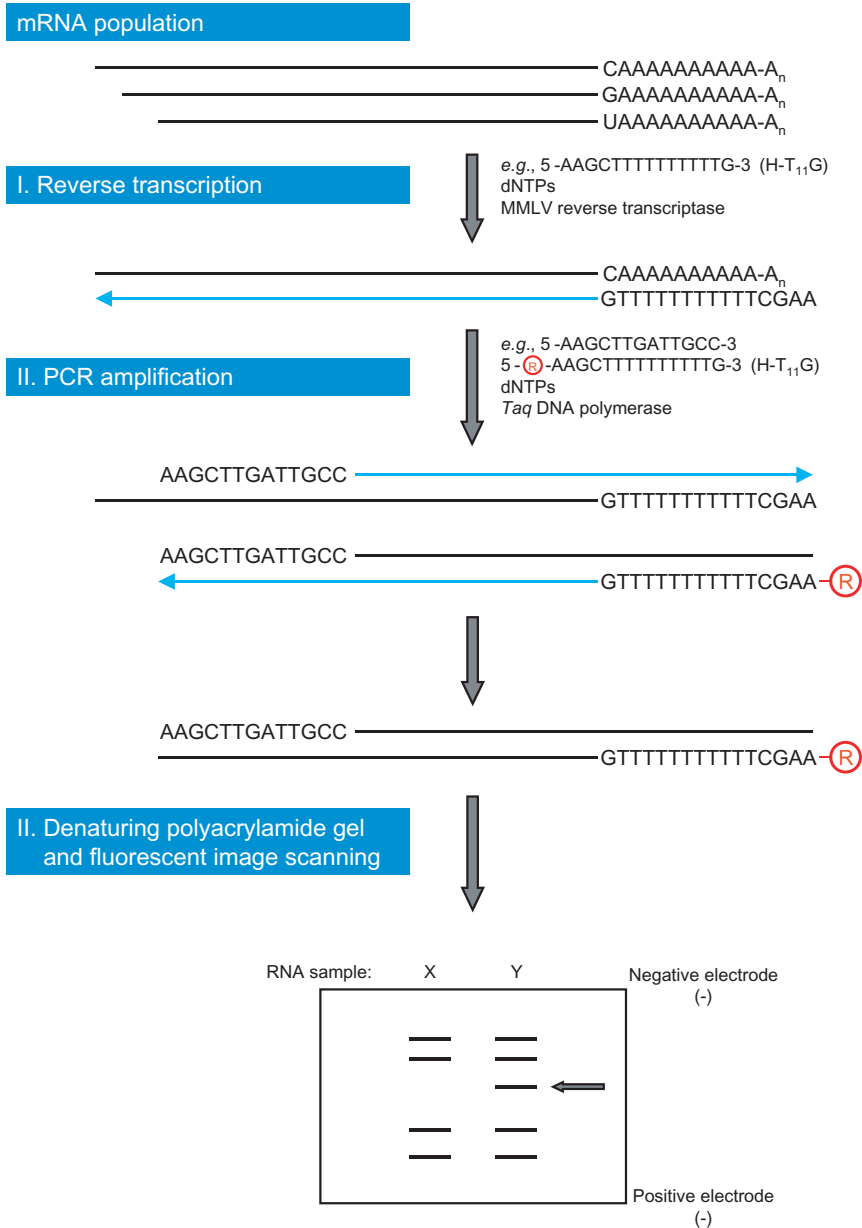


Figure 4.52 Schematic representation of fluorescent mRNA differential display (FDD). Three one-base anchored oligo(dT) primers with 5'-Hind III sites (H-T₁₁G, H-T₁₁C and H-T₁₁A) are used in combination with a series of arbitrary 13-mer oligonucleotides also containing 5'-Hind III sites to reverse transcribe and amplify the mRNAs from a cell. Fluorescently labelled anchor primers are used in the PCR amplification step.

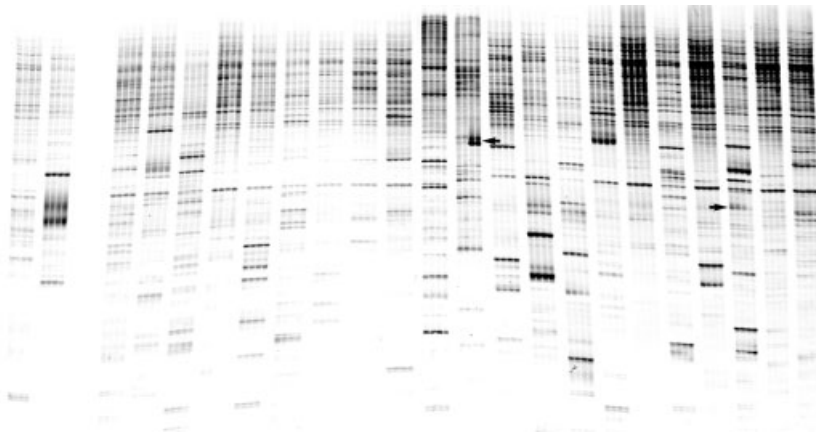


Figure 4.53 Digital fluorescent mRNA differential display (FDD) image. Comparison of four RNA samples (before and six, nine, and twelve hours after drug treatment) with automated denaturing polyacrylamide gel. Each fluorescent band is a cDNA amplified by a specific anchor primer and arbitrary primer combination. The arrows indicate potential differentially expressed cDNAs.

and high-precision differential display RT-PCR platform, which incorporates robotic liquid handling alongside digital data acquisition and analysis, is highly desirable. To this end, fluorescent differential display (FDD) was developed (Ito *et al.*, 1994) and optimised to have a similar sensitivity (Cho *et al.*, 2001) (figure 4.52). Combined with robotics and digital data analysis, automated FDD was shown to be accurate and to allow high throughput (Cho *et al.*, 2001; Liang, 2000). Data analysis tools such as spectra overlay, which allows digital data presentation and quantification, have been developed by Hitachi Genetics Systems, Inc. (San Francisco, California, USA), for use with the FMBIO® series of fluorescent scanners (Hitachi Genetics Systems, Inc.) (figure 4.54). Further streamlining the FDD process, any differentially expressed cDNAs of interest could be reamplified and sequenced directly with a given arbitrary primer without sub-cloning (Buess *et al.*, 1997). Furthermore, computer programs were developed to automatically allow positive band identification from an FDD image (Aittokallio *et al.*, 2000 and 2001; Qin *et al.*, 2001). The most sophisticated attempt in FDD downstream automation may have been the development of a prototype computer-controlled system for positive band identification and retrieval by the Hitachi Genetics Systems' system (Irie *et al.*, 2000). This approach employed capillary array gel electrophoresis coupled with fraction collection using sheath flow technology. Automation in PCR set-up and fluorescent data analysis for the total gene expression analysis (TOGA) and GeneCalling methods has also been described (Green *et al.*, 2001; Lo *et al.*, 2001).

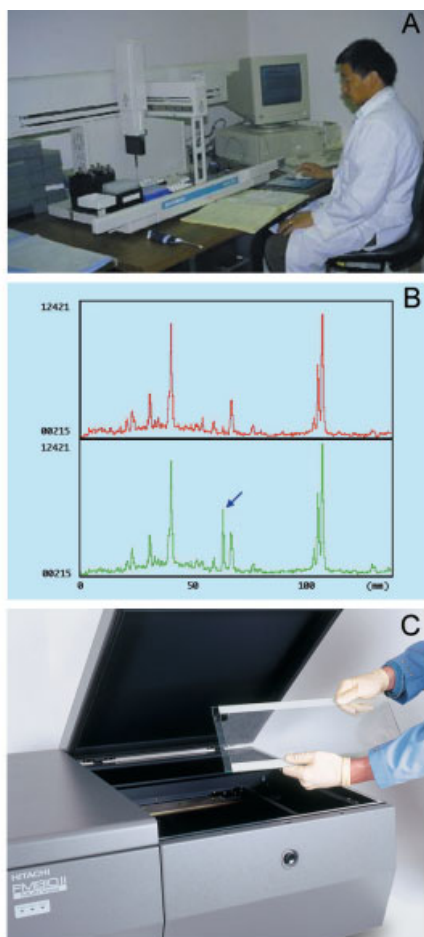


Figure 4.54 Workflow of fluorescent mRNA differential display (FDD). **A)** Beckman Biomek 2000 robotic liquid dispenser (Beckman Coulter, Inc., Fullerton, California, USA). Complete automation in FDD set-up by robotic liquid dispensing ensures high-throughput and reproducibility by reducing pipetting errors by laboratory personnel. **B)** Digital FDD electropherogram analysis. Digital quantitative analysis of an FDD gel electropherogram accurately reveals a drug-inducible gene (arrows). The individual spectra represent different RNA samples being compared. **C)** Image scanning with a fluorescent imager. The digital FDD image from figure 4.53 can be acquired using a fluorescent imager such as the FMBIO® II Fluorescence Scanner Laser from Hitachi Genetic Systems (Alameda, California, USA).

4.3.3.4 Comparison of FDD with RFLP-based differential display strategies

One of the assumptions for the false-positives in differential display RT-PCR has been that the short arbitrary primers (ten to thirteen bases) hybridise promiscuously to the cDNA templates, in sharp contrast to experimental data showing that these primers can hybridise predictably to their target templates with seven to ten bases matches at their 3'-ends (Liang *et al.*, 1994 and 1995). To this end, many attempts have been made to replace the selection of the mRNAs by arbitrary primers with restriction enzyme digestions (Matz *et al.*, 1998). There have been close to a dozen such restriction fragment length polymorphism (RFLP)-based differential display methods described so far, each with its own name such as restriction fragment length polymorphism-coupled domain-directed differential display (RC4D) (Fischer *et al.*, 1995), ordered differential display (ODD) (Matz *et al.*, 1997), GeneCalling (Green *et al.*, 2001), amplified differential gene expression

(ADGE) (Chen *et al.*, 2001), total gene expression analysis (TOGA) (Lo *et al.*, 2001; Sutcliffe *et al.*, 2000) and amplification of double-stranded cDNA ends restriction fragments (ADDER) (Kornmann *et al.*, 2001) (see also chapter 3 and the corresponding section within this chapter). These methods, in principle, are essentially the same with a variation on a theme. The strategy is to reverse transcribe the mRNA with oligo(dT) primers with or without an anchor, followed by the second strand cDNA synthesis. The double-stranded cDNAs are then digested with four-base cutter restriction enzymes, which creates a population of cDNA ends to which adapter primer-binding sites can be ligated. The cDNAs are then amplified with either the anchored oligo(dT) primers in combinations with upstream-ligated primers, or with the ligated primers alone. In the case of ODD, the amplifications of cDNA fragments other than the 3'-termini is suppressed by self complementary primers ligated to the cDNAs, whereas for ADDER and TOGA, the cDNA fragments corresponding to the 3'-termini of mRNAs are purified and released by the use of a specially designed anchored primer attached to a solid phase, such as magnetic beads (Kornmann *et al.*, 2001; Lo *et al.*, 2001). Unlike the traditional differential display RT-PCR method which requires only two steps, namely reverse transcription and PCR, the RFLP-based differential display strategies all require many more experimental steps, such as second strand cDNA synthesis, restriction digestion and ligation of adapter primers before mRNA samples are compared. Since each of the additional steps inevitably exposes a method to variability between or among RNA samples before they are compared, the improved accuracy of these strategies over that of traditional differential display RT-PCR has yet to be substantiated. At the time of this publication, in contrast to traditional differential display RT-PCR which has been cited several thousand times (ISI data base search), none of the RFLP-based differential display strategies have been cited more than ten times. Clearly, it can not be emphasised enough that if a method is to be developed, or an existing method is to be improved upon, one must first incorporate a positive control during the method development, and then show a side-by-side comparison of the old and new method. Without these two important considerations, it will be difficult to judge the reliability and significance of a method or method improvement.

4.3.3.5 Looking into the future: differential display RT-PCR versus DNA microarrays

Given the great interest in DNA microarray techniques in recent years, people may start to wonder about the role of differential display RT-PCR in gene expression analysis in the near future. It should be pointed out that, like differential display RT-PCR when it was first developed, DNA microarray is an emerging technique that has many technical problems (King *et al.*, 2001). Chief among these are problems for reproducibility, probe sensitivity, non-linearity in signal detection (Ramdas *et al.*, 2001), probe cross-hybridisation due to homologous cDNA sequences (Richmond *et al.*, 1999) and data management (Gibbs, W., 2001). There certainly has been a misperception that DNA microarray data (which is essentially based on reverse Northern blot) is as accurate and sensitive as Northern blot itself. Therefore,

its proponents have developed fancy computer programs dubbed ‘cluster analysis’ trying to make sense of a large number of unconfirmed gene expression differences often from a single experiment within the given biological system being studied. Concepts, such as ‘self-organising data’ emerged, as if nature’s secret can be revealed simply by the compilation of expression patterns of a limited number of unconfirmed genes. ‘Genome-wide gene expression profiling’ became a standard banner for many DNA microarray publications, which in fact only evaluated a few thousand out of an estimated 30,000 to 50,000 genes in a genome. It might be shocking to learn that even before a DNA microarray experiment was carried out, the sequences deposited on a cDNA array might have contained as much as a 30 percent error rate (Knight, 2001a). Recently, a major oligonucleotide-based microarray company had to recall several of its mouse ‘chips’ due to as much as a 65 percent error rate on the synthesised sequences (Marshall, 2001). Other problems seem to be afflicting the microarray manufacturers, as another major player in cDNA microarrays has recently dropped its microarray business altogether, citing competitions instead of technical difficulties as the major reason (Knight, 2001b). Furthermore, related problems and challenges for microarrays have been the subject of other critics (Gibbs, 2001; Liang *et al.*, 2001; Mir, 2000; Wooster, 2000).

In contrast, differential display RT-PCR has many unique advantages over DNA microarrays (table 4.7). For example, in contrast to DNA microarrays, RT-PCR-based differential display is an open system which does not depend on any prior knowledge of the genes to be analysed. As such, differential display RT-PCR can discover novel genes and is readily applicable to any biological system where no microarray ‘chips’ are available. Furthermore, differential display RT-PCR can compare more than two RNA samples side-by-side without the need for data normalisation. Through further refinement and automation, RT-PCR-based differential

Table 4.7 Comparison of differential display RT-PCR and DNA microarrays.

	<i>Differential display RT-PCR</i>	<i>DNA microarrays</i>
Year invented	1992	1995
Number of publications*	3,862	1,196
Publications per year	429	199
Detection of novel genes	Yes	No
Systems applicable	All eukaryotes	Human, Mouse, <i>Drosophila melanogaster</i> , Yeast, <i>Caenorhabditis elegans</i>
Amount of RNA required	Five to ten micrograms per 240 PCR reactions	≥ 20 micrograms per hybridisation
Cost	Low to medium	High
Do it yourself	Yes	Mostly not
Commercial services	GenHunter	Affymetrix, Research Genetics, and others

* Medline searches conducted on January 14th, 2002, by PubMed at NCBI website (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>) using keywords depicting each methodology.

display undoubtedly will continue to play a key role in gene discovery research in the post-genome era. Perhaps, a true contest between differential display RT-PCR and DNA microarrays in accuracy, sensitivity and gene coverage awaits someone who is willing to carefully compare the same biological system systematically using both methodologies. Ultimately, it will be the functional characterisations of each isolated genes by genetic, cell biological and biochemical methods, which will provide the real proof (or disproof) of the relevance of the genes to a biological system under investigation. In a preface to a method book on protein purification, Dr. Arthur Kornberg had once quoted an admonition of Efraim Racker “Don’t waste clean thinking on dirty enzymes” to illustrate the importance of good biochemical practice which is at the core of enzymology (Kornberg, 1990). Similar doctrine, “Don’t waste clear thinking on dirty data”, will certainly continue to help to produce better quality of science in the field of gene expression analysis in the next ten years.

4.3.4

GeneCalling

4.3.4.1 Introduction

The comprehensive description of differences in gene expression among samples is a powerful method of identifying genes associated with diseases, traits and biological responses to chemicals. Existing methods for expression analysis fall into three general classes: transcript sampling by sequencing (Lee & Tomasetto, 1991; Lee *et al.*, 1995; Velculescu *et al.*, 1995), transcript amplification and imaging (Bachem *et al.*, 1996; Hubank & Schatz, 1994; Ivanova & Belyavsky, 1995; Kato, 1995; Liang & Pardee, 1992) and hybridisation-based approaches (DeRisi *et al.*, 1997; Lockhart *et al.*, 1996; Schena *et al.*, 1995; Shalon *et al.*, 1996; Wodicka *et al.*, 1997). Serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995), a cost-effective transcript-counting technique, is limited by the small amount of sequence information obtained for each gene. Transcript sequencing following subtractive hybridisation also identifies differentially expressed genes, but is limited to binary comparisons (Lee & Tomasetto, 1991). Transcript imaging approaches such as differential display RT-PCR (Liang & Pardee, 1992), partitioning by type IIS restriction enzymes (Kato, 1995), representational difference analysis (RDA) (Hubank & Schatz, 1994), and amplified fragment length polymorphism (AFLP) (Bachem *et al.*, 1996) are rapid and, in theory, are comprehensive since they utilise banding patterns that are dependent upon gene expression. However, each of these requires a time-consuming cloning and confirmation process for determination of the identity of differentially expressed gene fragments.

The development of microarrays has revolutionised the capacity of hybridisation techniques (DeRisi *et al.*, 1997; Lockhart *et al.*, 1996; Schena *et al.*, 1995; Shalon *et al.*, 1996; Wodicka *et al.*, 1997) to identify differences in gene expression. Hybridisation-based approaches are rapid and provide the identity of differentially expressed genes of known sequence immediately. In practice, however, hybridisation methods are limited by an inability to detect or discover completely novel genes

with no expressed sequence tag representation, thus making work in most organisms impossible.

The GeneCalling® method was developed to circumvent these limitations. The method permits high-throughput, reproducible detection of most expressed sequences with a sensitivity of greater than one part in 100,000 transcripts. It allows the discovery of differentially expressed genes, both known and novel, from any species and with useful sequence information to determine the potential function of novel genes captured (Shimkets *et al.*, 1999). The method has been applied to a wide variety of disease models in a wide variety of species, addressing problems as diverse as identifying novel human cancer gene targets (Herrmann *et al.*, 2001; Kahn *et al.*, 2000), understanding how drugs and diet affect animal models of disease (Basson *et al.*, 2000; Rininger *et al.*, 2000), and understanding the basis of trait differences in related strains of corn (Bruce *et al.*, 2000 and 2001).

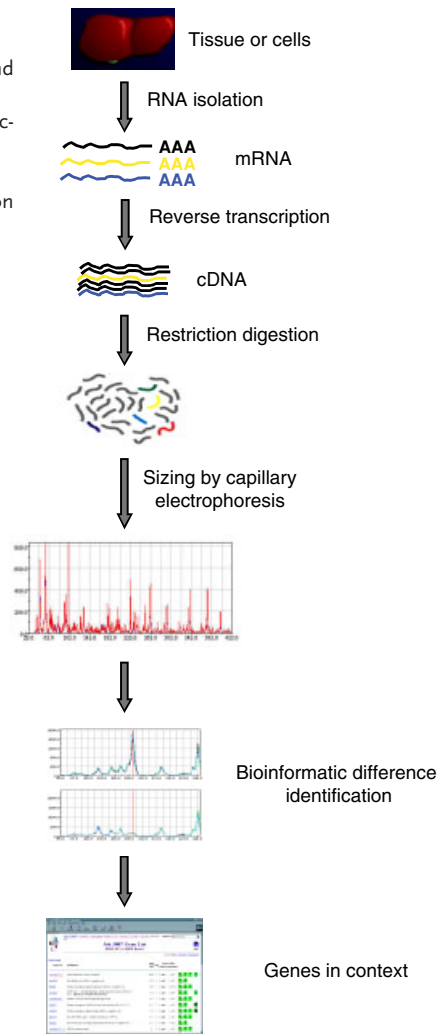
4.3.4.2 Principles and basics

The GeneCalling® chemistry has three main steps: (i) Restriction endonuclease digestion of cDNA, (ii) adapter ligation, and (iii) PCR amplification (figure 4.55). Following double-stranded cDNA synthesis from poly(A)⁺ mRNA, cDNA pools are digested with different pairs of restriction enzymes with six-base pair recognition sites. Complementary adapters are ligated to the digested cDNA, and adapter-specific primers are used to direct 20 cycles of PCR, amplifying fragments containing sites for the pairs of restriction enzymes used. One adapter-specific primer is biotin-labelled while the other is labelled with the fluorescent dye FAM which contains amine-reactive succinimidyl esters of carboxyfluorescein. Following PCR amplification, the biotin-labelled DNA is purified on immobilised streptavidin. Denatured single-stranded DNA fragments are resolved by electrophoresis on ultra-thin polyacrylamide gels or capillaries, and FAM-labelled fragments are detected upon laser excitation. Since the biotin label is necessary for purification and the FAM label is necessary for detection, all detected fragments result from restriction digestion with both enzymes. Typically, 96 GeneCalling® reactions are performed, each with a separate pair of endonucleases, on triplicate samples.

Three independent reactions from the same cDNA sample are compared for quality of electrophoretic peak resolution and reproducibility of peak patterns. Composite traces from each sample are generated and then compared between three independent samples for peak quality and reproducibility. A composite trace is calculated for each sample group based on average peak height and variance, and composite traces are compared among sample groups using software designed to identify peaks representing differential expression.

Complementary DNA fragments representing differentially expressed genes can be identified by database searching with the six-base pair restriction enzyme recognition sequences at the fragment ends and the exact length of each fragment (determined electrophoretically with subtraction of the linker length). Database searching for genes predicted to have restriction fragments of matching lengths enables the immediate identification of all of the genes whose sequences reside in

Figure 4.55 Process flow for the GeneCalling method. In the GeneCalling method, tissue or cells are processed through mRNA isolation and double-stranded cDNA synthesis by standard methods. cDNA is digested with pairs of restriction enzymes, ligated to linkers, amplified, and sized by capillary electrophoresis. Specialised software for gene expression difference detection identifies the position, size, and database sequence of differentially expressed genes.



that database, and flags fragments derived from novel genes by virtue of their absence from the database. Database lookup can provide a unique assignment of gene identity, and the detection of multiple fragments derived from the same gene which show differential expression of the same directional modulation increases the likelihood that the prediction of the gene identity is correct (figure 4.56).

A method for rapid confirmation of the identity of DNA fragments determined by database searching was developed in order to avoid time-consuming cloning, Northern blot analysis or further PCR-based quantification requiring time-consuming optimisation. The reaction containing the fragment of interest is performed a second time using the same end primers but in the absence and presence of an excess of an unlabelled oligonucleotide whose sequence is derived from the pre-

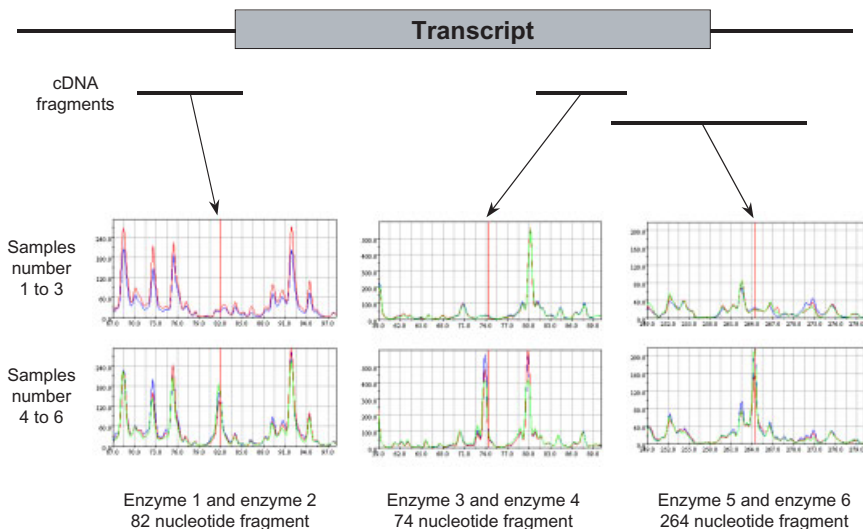


Figure 4.56 GeneCalling data example. A typical transcript whose cDNA would produce three captured products from the GeneCalling method is shown. In each of the three cases, samples number 1 to 3, consistent with the fact that each of these fragments corresponds to the same gene in the database. samples number 4 to 6 show significantly

dicted gene fragment. If the identity of the fragment was predicted correctly, the unlabelled oligonucleotide will outcompete the universal oligonucleotide for priming that fragment and appear in the chromatogram to ablate that peak specifically without affecting the amplification of the other fragments (Shimkets *et al.*, 1999). In addition to confirming the identity of a gene, this method allows more precise estimation of the magnitude of gene expression differences.

While designing oligonucleotides to database-retrieved sequences can only be performed when the gene sequence in question is already in a database, this oligonucleotide competition is also useful as part of the isolation process of novel genes. After a band of interest is excised from a gel, a custom oligonucleotide is designed against each distinct gene sequence that is recovered from that fraction, and the oligonucleotide ablating the peak of interest corresponds to the differentially expressed gene.

Restriction fragments which match in end sequence and length to known genes are used as templates for the design of unlabelled oligonucleotide primers. An unlabelled oligonucleotide designed against one end of the restriction fragment is added in excess to the original reaction, and is re-amplified for additional 15 cycles. This reaction is then electrophoresed and compared to a control reaction reamplified without the unlabelled oligonucleotide to evaluate the selective diminution of the peak of interest.

4.3.4.3 Examples and presentation of typical results

GeneCalling® has been applied to numerous biological systems in order to elucidate how gene expression changes contribute to the underlying pathophysiology of disease. This method has been applied to a surgical model of heart disease in order to identify potential markers and/or therapeutic targets for cardiac hypertrophy (Shimkets *et al.*, 1999). In this study, genes known to be associated with cardiac hypertrophy were recapitulated, and genes not previously associated with heart disease were captured. Since this model was performed in the rat and the sequence database for rat at that time was rather sparse, GeneCalling® was an excellent choice for a transcript profiling method in this system since the novel genes were easily isolated.

GeneCalling® was also applied to an *in vitro* model of angiogenesis, in which endothelial cells under various culture conditions form primitive blood vessels (Kahn *et al.*, 2000). Analysis of gene expression changes correlating with blood vessel development included potential new pathways for angiogenesis.

In addition to the study of models of disease, GeneCalling® has been applied to the study of drug response. The identification of surrogate markers for drug efficacy or toxicity in drug families such as the glitazones is enabled by this comprehensive gene expression profiling method (Rininger *et al.*, 2000). Understanding how diet also affects gene expression was studied using a variety of short versus long-chain fatty acids as treatments on human colonocytes (Basson *et al.*, 2000).

In addition to its utility in the study of the human condition, GeneCalling® has been applied to the study of quantitative traits in corn. Related strains with a difference in a key phenotype were compared using GeneCalling® in order to identify the genes responsible for the trait, with the hope of someday manipulating that pathway to get stronger plants (Bruce *et al.*, 2001). Using GeneCalling®, entire pathways under the control of transcription factors can be explored (Bruce *et al.*, 2000).

To date, GeneCalling® has been applied to a wide variety of both animal and plant systems for trait identification, marker and pathway identification, and has utility both in the drug discovery arena as well as the study of drug response.

4.3.4.4 Discussion

The principle advantages of GeneCalling include the flexibility to discover known and novel dysregulated genes, the ability to apply this technology to any organism containing tangible RNA, the capturing of the transcript's centre which provides protein-coding information, the ability to sensitively distinguish rare and abundant transcripts, the ability to independently measure transcript abundance multiple independent times in a single experiment, and the ability to comprehensively measure the majority of transcripts in a cell. These characteristics make GeneCalling® an attractive system for the drug discovery industry as well as a variety of other molecular biology applications.

4.3.5

iAFLP – Introduced amplified fragment length polymorphism**4.3.5.1 Introduction**

Amongst various methods for analysing gene expression, RT-PCR has the highest sensitivity and specificity in transcript detection (Becker-Andre & Hahlbrock, 1989). The major drawback of genome-wide application of RT-PCR is low throughput, especially when quantitative data is desired. To analyse thousands of different transcripts with concentrations that differ as much as 1,000-fold (Okubo *et al.*, 1992), non-competitive PCR is not feasible because it requires a standard curve for every primer pair (Singer-Sam *et al.*, 1990). On the other hand, quantification by means of competitor fragments for each gene is not realistic. Kikuya Kato proposed that engineered total cDNA from different source might be used to replace the competitor DNA for every transcript species (Kato, 1997). Introduced amplified fragment length polymorphism (iAFLP) is a high-throughput method for quantifying transcripts that was developed by expanding Kikuya Kato's idea and it allows relative quantification of transcript concentration in as many as six mRNA sources in a single PCR reaction (Kawamoto *et al.*, 1999). iAFLP has been applied to the analysis of common cells in the gene by tissue expression matrix of BodyMap (Kawamoto *et al.*, 1999) (see chapter 7).

4.3.5.2 Principle and basics

A set of adapters that mimics a length-polymorphic locus is prepared. These adapters have common sequences at both ends and sequences of different lengths internally. The polymorphic portion of the adapters is composed with degenerate adenine, cytosine and thymine residues to neutralise sequence-specific effects of the adapters on the amplification kinetics (figure 4.57.B). Ligation of these adapters to enzyme-digested cDNA fragments from multiple sources yields source-specific length-polymorphism at the ends of cDNA fragments. From a pool of such adapted cDNAs, any gene sequence can be amplified as length polymorphic fragments when a gene-specific primer and an adapter primer are used (figure 4.57.A). Because of the nearly identical sequence content and common primer annealing sites, the polymorphic fragments have similar amplification efficiencies. Therefore, the relative concentrations of the transcript in different sources are reflected in the amounts of products of corresponding sizes (iAFLP profile). By applying the PCR product to an automatic sequencer with fragment analysis software, the amount of each polymorphic fragment can be quantified.

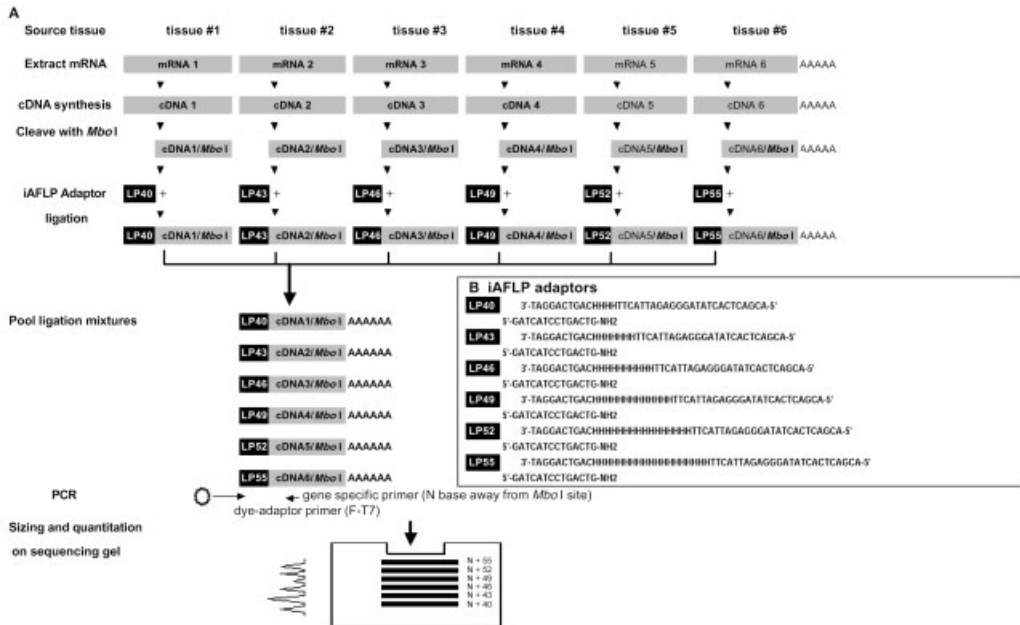


Figure 4.57 The introduced amplified fragment length polymorphism (iAFLP) method. **A)** A schematic diagram of procedures for expression profiling by iAFLP. (1) cDNA synthesis in six separated tubes. (2) Restriction enzyme digestion with *Mbo* I. (3) Ligation with one of the iAFLP adaptors (LP40 to LP55) shown in B). (4) Pool the ligation mixture in one tube. (5) PCR with antisense gene-specific primer and a fluorescent dye (F)-labelled primer for adapter sequence (F-T7(25): F-ACGACTCACTATAGGGA-GATTACTT). (6) Apply PCR products to automatic sequencer. **B)** iAFLP adapter primers. H represents degeneracy of adenine, cytosine and thymine. Non-ligating recessed 3'-end of adapter cassettes are blocked by introducing NH₂-groups. The strand that anneals to dye-primers is generated only by extension of the antisense primer.

4.3.5.3 Typical results

Starting from cDNAs from six tissues, single PCR reaction with one gene specific primer yields data comparable to multi-tissue Northern hybridisation (figure 4.58.A). Strict design of gene-specific primers is not necessary, since non-specific PCR products do not affect the results unless they are dominant (figure 4.58.B). Designing the primers in the vicinity of the adapter usually yields better results than in the distant position. The 5'-terminal nucleotide on the gene-specific primer (the 3'-end of the labelled strand of the PCR product) influences the elongase activity of the *Taq* polymerase that makes the PCR product longer than the expected size by one base (usually by the addition of adenine). Primer-dependent fluctuation in the profile of twin peaks is not desired when automatic detection of specific peaks is planned. Introducing extra guanine on the 5'-end of the gene-specific primer enhances the elongase activity and makes twin peaks into singles (figure 4.58.B). For the analysis of human genes, recommended primer sequences are given at the BodyMap site (<http://bodymap.ims.u-tokyo.ac.jp>).

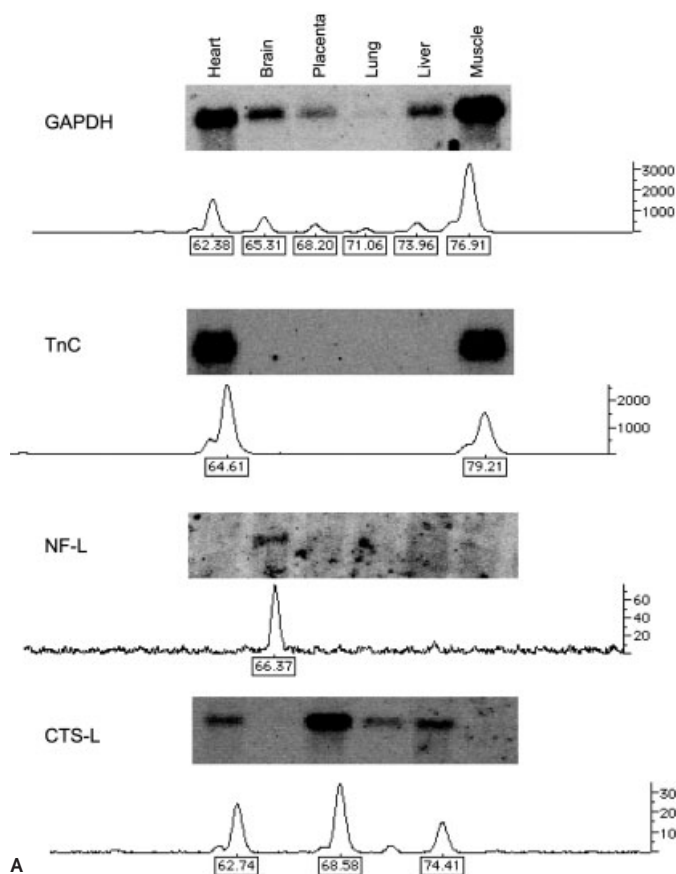
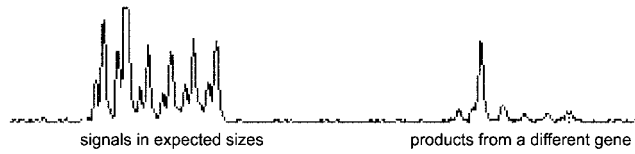


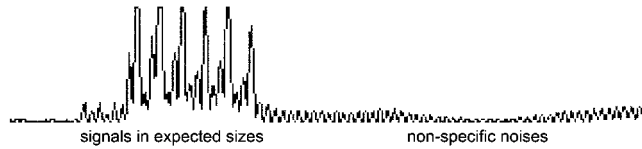
Figure 4.58 cDNAs from six different tissues analysed by introduced amplified fragment length polymorphism (iAFLP) and Northern blot. **A)** iAFLP profiles showing tissue distribution of various transcripts. A pool of cDNA templates was prepared from six human tissues

and amplified with four different gene-specific primers. Hybridisation patterns of multi-tissue blots to corresponding gene probes are inserted above electropherograms. For primer sequences, see Kawamoto *et al.* (1999).

Cross-hybridisation to different transcripts



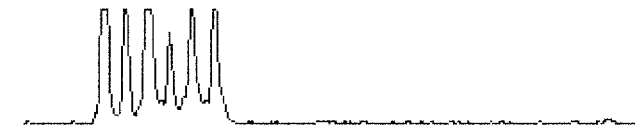
Non-specific priming



Large N/N+1 ratio



Lowering N/N+1 by adding G to 5'-end of the primer



B

Figure 4.58 B) Effects of primer sequence to the iAFLP profiles. *Top:* Cross-hybridisation to different transcripts. A set of peaks with similar intervals exists in different size range. Unless these bi-products overlaps with expected products, the quantification results are not affected.

Middle: Non-specific products having broad size distribution are seen. *Bottom:* Products resistant to elongase activity. Non extended strand (N) exceeds extended strands (N+1). By adding extra guanine to 5'-end of gene-specific primer, N+1 strands became dominant.

4.3.5.4 Examples

With the present protocol that allows direct comparison of six sources, expression in an unlimited number of sources can be compared using one common reference cDNA sample divided into six different pools. Figure 4.59.A shows the iAFLP profiles for 30 different tissues amplified from six pools (central nervous system to connective tissue) of six adapted cDNAs. As a reference cDNA, all cDNAs prior to adapter ligation (mix) were pooled and included in every pool as a sixth component. Antisense primer anneals to most 3'-*Mbo*I fragments of cDNAs for major histocompatibility complex, class I, E (HLA-E) heavy chain. By normalising all peaks relative to the signal from the reference cDNA, the ratio of concentrations in 30 human tissue mRNAs was determined (figure 4.59.B). This reference allowed comparison of nearly 100-fold differences in expression, although the dynamic range of the detection platform is on the order of ten-fold. Results for similar analyses with nearly 20,000 human transcripts are available at the BodyMap site (<http://bodymap.ims.u-tokyo.ac.jp>) (Kawamoto *et al.*, 2000).

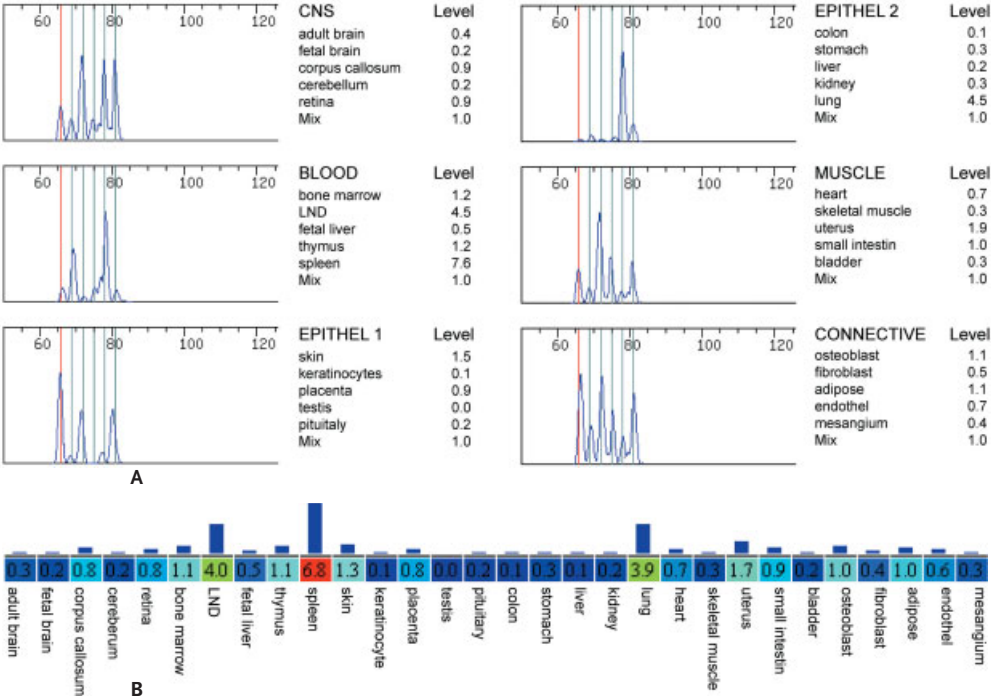


Figure 4.59 Analyses of 30 different tissues by introduced amplified fragment length polymorphism (iAFLP). A) iAFLP profiles of a gene (major histocompatibility complex, class 1, member E, HLA-E, heavy chain) showing anatomical distribution patterns across 30 tissues. Six pools of cDNA templates (central nervous system to connective tissue) were amplified with one gene-specific primer and one adapter primer. **B)** Relative concentration in 30 source RNAs were calculated by normalising peak heights with reference peaks (every sixth in all pools). Such results are provided for nearly 20,000 human genes in the BodyMap site (<http://bodymap.ims.u-tokyo.ac.jp>).

4.3.5.5 Discussion

The iAFLP method has several advantages over other methods. The signal for any gene-specific primer is validated by the size of the PCR product, which is determined to single nucleotide resolution in iAFLP. The quantitative aspect is validated by calculating the ratio of the total amount of products from individual sources to the product from the mixture of these sources. This is an advantage over, for example, DNA microarray hybridisation experiments, which is weak in gene-by-gene data validation. Gene selection for the study is highly flexible, and newly discovered genes are easily tested by researchers. In iAFLP, more than 10,000 genes can be tested with one microgram of poly(A) mRNA without any pre-amplification step and thousands of samples can be tested with the same set of primers synthesised at a nanomolar scale.

Disadvantages of the iAFLP method are as follows: For genome-wide application, the cost for primer synthesis is substantial. To manage the PCR reactions and the sequencing runs, which are as numerous as the tested genes, multi-channel liquid dispenser machines and high-throughput sequencers are virtually indispensable. Accordingly, iAFLP is not suitable for testing many genes in a small numbers of tissues. As is the case with DNA microarray hybridisation experiments, gene-to-gene comparison of expression level in one tissue is not possible by iAFLP.

4.3.6

RAGE — Rapid analysis of gene expression

4.3.6.1 Introduction

Global gene expression technologies such as cDNA or oligonucleotide microarrays and SAGE can give comprehensive snapshots of the state of gene expression in a cell type or tissue, but at high cost in materials and personnel. Rapid analysis of gene expression (RAGE) was originally conceived as a low cost alternative technology appropriate for any reasonably equipped molecular biology laboratory interested in analysing the expression of intermediate numbers of genes, (*i.e.* ten to one thousand), in a directed manner. The original design parameters included the potential for analysing any chosen gene without extensive development of new primers or alteration of reaction conditions, high sensitivity, direct assessment of the statistical significance of measurements, and the potential for gene discovery. The original description of a method satisfying these parameters was published in 1999 (Wang *et al.*, 1999c), and an U. S. patent covering the technique was issued in 2001 (MacLeod *et al.*, 2001). The technology is currently available in kit form (GeneSystem320) from Capital Genomix, Inc. (Gaithersburg, Maryland, USA).

4.3.6.2 Principles and basics

The basic strategy employed in RAGE is to construct a library of so-called RAGE-tags, short, defined fragments of each mRNA present in a population, that will be used as template in subsequent PCR reactions. Bioinformatics software supports

primer selection with primers selected combinatorially from a small, pre-designed set that specifically amplifies the RAGETag from a chosen gene based on the sequence of the RAGETag. The relative concentration of the corresponding mRNA in the original population is inferred from the level of product formation after PCR amplification. As with other quantitative PCR methods, care must be taken to ensure that product formation is linear with input mRNA. RAGE results do not routinely give an absolute concentration of a specific message, but rather relative amounts of any given mRNA can be compared between biological samples.

The steps used in a typical RAGE analysis are illustrated in figure 4.60. Complementary DNA is prepared from an mRNA sample using a biotinylated oligo(dT) primer and the cDNA is immobilised on a streptavidin magnetic bead. Two restriction enzymes, *Hsp91* II and *Dpn* II, which cleave DNA leaving unique overhanging sticky ends, designated 'A' and 'B', are used sequentially to fragment the cDNA. The positions of recognition sites for these enzymes in each cDNA define the position of the RAGETag for each gene. Two arbitrary genes are diagrammed in the figure with 'A' restriction sites indicated as dark grey bars and 'B' restriction sites indicated as light grey bars. Initially, the immobilised cDNAs are cleaved with enzyme 'A', leaving only the 3'-most 'A' fragment attached to the beads, and the cleaved 5'-fragments are washed off and discarded. The RAGETag is then cleaved from the beads with enzyme 'B', along with other 3'-fragments, and collected. Note that at this point in the preparation, only the RAGETags contain sticky ends derived from enzyme A; the other fragments of cDNA that contaminate the preparation have 'B' sites at both ends. Taking advantage of the unique sticky ends left by the 'A' and 'B' enzymes, the RAGETags are then ligated to two unique linkers that distinguish the 'A' and 'B' ends; these provide common 'A' and 'B' primer binding sites for subsequent PCR analysis; currently used linkers are 19 nucleotides in length. The 'A'-end linker is biotinylated so that the RAGETags can be purified in the next step by binding to streptavidin magnetic beads, eliminating the unwanted cDNA fragments that contain only 'B'-ends. This library of immobilised RAGETags can then be used as template in reactions designed to amplify particular gene products. The production of a RAGETag library from cDNA can be accomplished in a single day, and multiple libraries can easily be prepared in parallel.

Specificity in the PCR amplification step is provided by using primers that extend past the 'A' or 'B' restriction sites into the unique portion of the RAGETag. Good specificity is obtained with a set of 'A'-end primers that contain the 'A'-end linker sequence but extend four nucleotides into the specific portion of the RAGETag, and a set of 'B'-end primers that contain the 'B'-end linker sequence and extend three nucleotides into the RAGETag from the opposite direction. Thus, the total set of primers needed for RAGE analysis is 4^4 (256) 'A'-end primers and 4^3 (64) 'B'-end primers. For any known gene, the sequence of the corresponding RAGETag can be determined from GenBank mRNA entries, and the sequence of the specific 'A'-end and 'B'-end primers that will amplify this RAGETag can be inferred. Thus, PCR amplification of the RAGETag library with the specific 'A' and 'B'-end primers should give rise to a product of known size. This size is defined by the distance between the 3'-most 'A' restriction site in the gene's cDNA and the

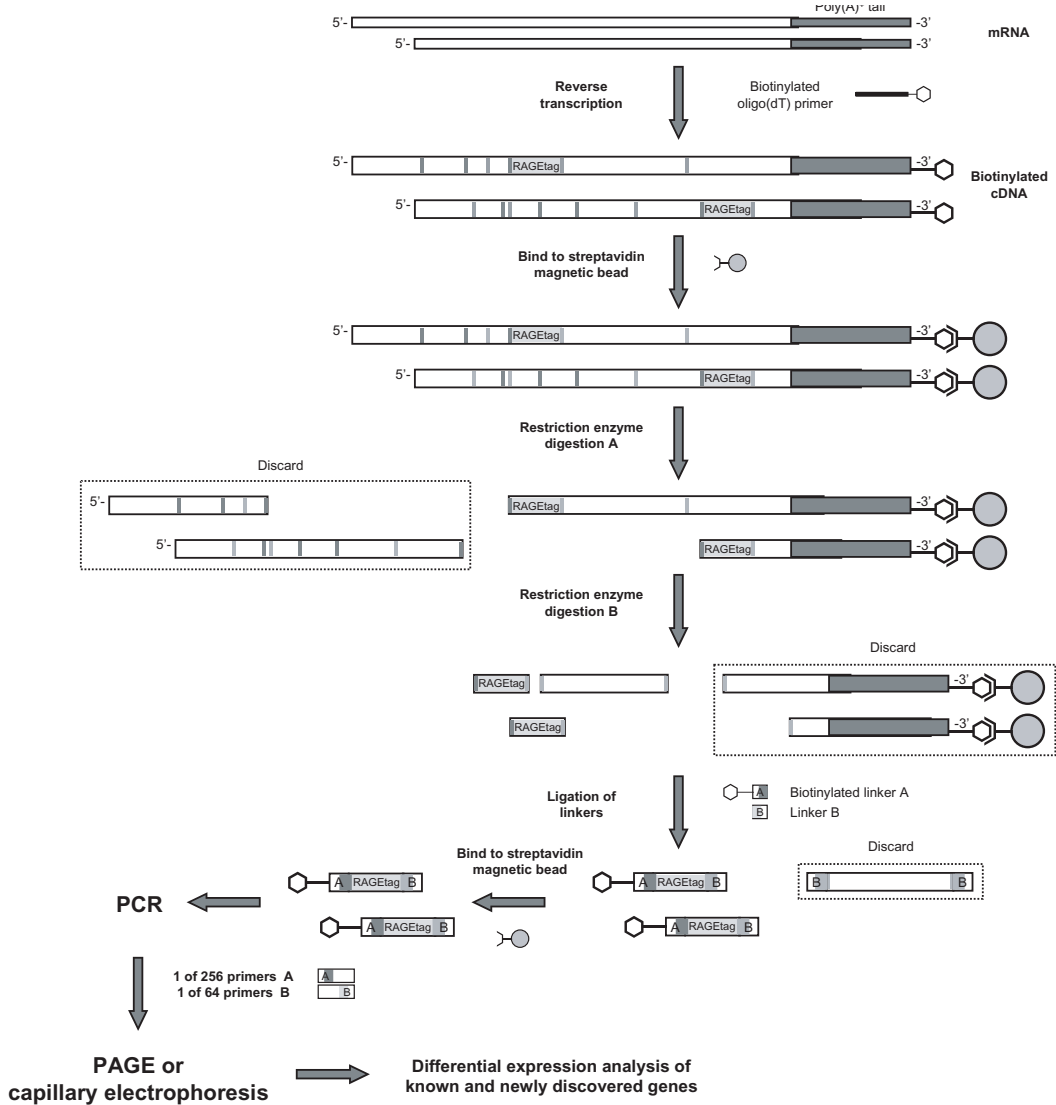


Figure 4.60 Flow-diagram of the RAGE method. See text for details.

closest 'B' restriction site in the 3'-direction, and can of course be predicted from the mRNA sequence.

Analyses of typical RAGE reactions on polyacrylamide gels with fluorescent staining are shown in figure 4.61. In this example, PCR reactions were run using RAGetags prepared from wild type (–) and E2F1-transgenic keratinocytes (+) as template, and reaction products were analysed in adjacent lanes. RAGE primers were analysed for three control genes (lanes A to C) and five genes that are

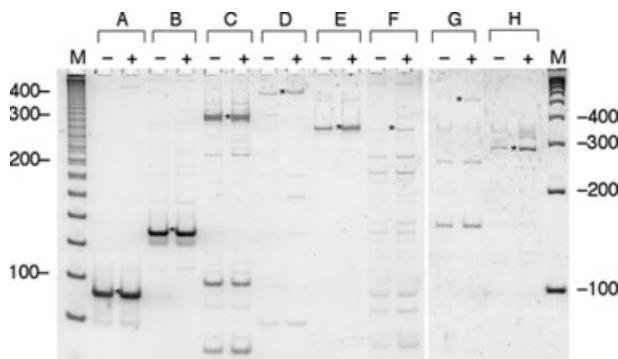


Figure 4.61 Analysis of gene expression in keratinocytes by RAGE. RAGE PCR analysis was performed, using RAGETag libraries prepared from wild-type (–) or K5-E2F1 transgenic (+) keratinocytes (K5-E2F1 mice are animals that are transgenic for a construct containing the bovine keratin 5 promoter controlling the E2F1 gene; the keratin 5 promoter targets expression to keratinocytes). Reactions contained RAGE primers chosen to amplify specific genes and the expected ampimers are indicated in the figure by black stars between the lanes. The

selected genes and the size of the expected ampimers were: lane A, *Actg* (γ -actin), 98 base pairs; lane B, *Rpl5* (ribosomal protein L5), 130 base pairs; lane C, *Lmna* (lamin A), 291 base pairs; lane D, *Cdk7* (cyclin-dependent kinase 7), 392 base pairs; lane E, *Yy1* (transcription factor yin-yang 1), 248 base pairs; lane F, *Foxd3* (transcription factor forkhead box d3), 254 base pairs; lane G, *Cdkn2a/p19^{ARF}* (cyclin-dependent kinase inhibitor 2A; p19 alternate reading frame), 508 base pairs; lane H, *Brca1* (breast cancer type 1 gene), 291 base pairs.

up-regulated in the transgenic keratinocytes (lanes D to H) were used in the reactions. In each case, the expected size of the specific PCR product is indicated by a black star between the two lanes. In the case of the control genes (γ -actin, ribosomal protein L5, and lamin A), major bands of equal intensity corresponding in size to the expected product were seen with both templates. For the experimental genes illustrated here, differential expression between the wild-type and E2F1 transgenic keratinocytes was observed. Indeed, in the case of the p19^{ARF} gene, which is known to be positively regulated by E2F1, the expected PCR product was virtually undetectable in the wild-type cells (lane G). For all genes of interest, the expected identity of the PCR product was confirmed by elution of the band from a preparative gel followed by sequencing. In many of the analyses, multiple bands with various molecular weights are seen. Some of these bands represent non-specific mis-priming events. However, in some cases the unexpected bands represent unidentified genes (further discussed below) and therefore provide a rich source for gene discovery.

Two restriction enzymes with four-base recognition sites which leave four-base overhangs, *Hsp92* II and *Dpn* II, are currently used as enzymes ‘A’ and ‘B’, respectively. The four-base recognition sequences occur at random approximately once every 256 base pairs. This results in RAGETags that average about 128 base pairs in length; over 90 percent of RAGETags are smaller than 500 base pairs. Because these enzymes are frequent cutters, most mRNAs contain at least one of each

kind of site; empirically, about five to ten percent of known genes were found that lack either one or both kinds of restriction site, or have closely spaced or overlapping sites that do not give a specific PCR product. However, about half of all mRNAs do not have a *Dpn* II site to the left of the 3'-most *Hsp*92 II site and therefore the RAGEtag library (called the 'A'/'B' RAGEtags) prepared using the protocol outlined in the figure will not represent this half of the transcriptome. To assay these genes, a second RAGEtag library ('B'/'A' RAGEtags) is made by reversing the order in which the restriction cuts are made. Use of both libraries therefore allows the researcher freedom to assay approximately 90 to 95 percent of the transcriptome.

To completely assay the entire transcriptome, each library would have to be amplified with each possible combination of RAGE primers. The total number of unique amplification reactions is therefore (two libraries) $\times$ (256 'A'-end primers) $\times$ (64 'B'-end primers) or 32,768. Assuming the total number of genes in a typical mammalian genome is in the range of 30,000 to 60,000, each unique RAGE amplification reaction will produce one to two specific PCR products, on average. However, in most cases the size of the PCR products produced from two different genes will be distinguishable using standard electrophoretic separation techniques. The corollary of this point is that in most cases a RAGE reaction designed to assay a particular gene-of-interest will simultaneously provide information on one or more other transcripts which may or may not be previously identified. This feature gives RAGE an enhanced potential for gene discovery. Since the average RAGEtag is approximately 128 base pairs in length, sequencing unknown PCR products after gel purification usually gives enough information to uniquely identify corresponding ESTs.

4.3.6.3 Examples and typical results

The RAGE technique was first used to analyse altered gene expression patterns in epidermal keratinocytes of mice constitutively over-expressing the cell-cycle-regulated transcription factor E2F1 (Wang *et al.*, 1999c). A number of genes known to be positively regulated by E2F1, including *Ccne1* (cyclin E1), *Cdc2d* (cyclin-dependent protein kinase; cell division control protein 2 homologue D), *Cdkn2a/p19^{ARF}* (cyclin-dependent kinase inhibitor 2A; p19 alternate reading frame), and *Tk1* (soluble thymidine kinase), were found to be up-regulated in the transgenic keratinocytes. In addition, because of the sensitivity of the RAGE technique, the tumor suppressor gene *Brca1* (breast cancer type 1 gene), which is expressed at extremely low levels in keratinocytes, was found to be over-expressed in the E2F1-transgenic cells. It was subsequently shown that E2F1 directly transactivates the *Brca1* promoter (Wang *et al.*, 2000), providing the first functional link between these two important tumour suppressors.

In the same study, several unexpected RAGE PCR products were noted whose expression was elevated in the transgenic cells. Sequencing of the PCR products led to the identification of one PCR product with a newly discovered apoptosis-related gene, *Pdcd6ip* (programmed cell death 6 interacting protein; previously

called Alg2-interacting protein-1; *Aip1*). The sequence of a second up-regulated PCR product was similar to that of a human cDNA for an epidermal differentiation-related gene. These RAGE results led to the identification of a family of ten murine epidermal differentiation-related genes, the *Sprrl* (small proline-rich-like) genes (Wang *et al.*, 2001d), and the subsequent identification of 17 members of the orthologous human family, *SPRL* (Wang *et al.*, personal communications).

The RAGE methodology has also been used to analyse changes in gene expression in normal human mammary epithelial cells and bronchial epithelial cells in response to carcinogen treatment, in cultured human tumour cells responding to cytokine treatment, in studies of breast tumor progression, and in defining molecular signatures of early human breast cancer.

4.3.6.4 Discussion

Some of the advantages and limitations of the RAGE method, particularly in comparison to microarray technologies, are as follows.

4.3.6.4.1 Gene discovery

As noted above, each RAGE reaction has the potential to assay one or more genes other than the target gene, and typically in any medium scale project (assay of a few hundred genes) several differentially expressed, unidentified genes will be found. Sequencing of the PCR product usually gives enough sequence information to search for matches in the EST databases. In the absence of an EST match, enough sequencing information is often available to design primers for 3' and 5'-rapid amplification of cDNA ends (RACE) (Frohman, 1993), and subsequently to clone the cDNA of interest. DNA microarrays can only assay those genes or ESTs that are present on the array, and there is no possibility of identifying genes that are not represented in an EST database.

4.3.6.4.2 Flexibility

Using RAGE, the set of genes being tested can easily be tailored to the problem of interest, and can be changed at any time. In particular, newly discovered genes or genes whose relevance is newly established can immediately be added to the set being assayed. Because the primers used for different genes only differ at three or four positions (about 80 percent identity), new assays can be performed without altering the reaction conditions. RAGEtag libraries can be stored indefinitely and re-assayed for different sets of genes in the future. A public database is available (<http://sciencepark.mdanderson.org/ggeg>) for predicting the pair of RAGE primers to use for any human, mouse or rat mRNA sequence in GenBank. In contrast, much less flexibility is typically possible with microarray techniques, and the speed with which new genes can be added to an array is limited. Most microarrays assay a mixture of genes designed to provide coverage of multiple different scientific interests. Thus, many of the genes assayed will not be of interest for a particular project, and not all genes relevant to a particular project are likely to be included in any given microarray.

4.3.6.4.3 Maintenance

A small set of reagents, namely the 320 RAGE primers and several linker oligonucleotides, is all that is needed for the full capabilities of RAGE; these can be synthesised in-house or commercially and replenished as needed. A particularly important advantage of the RAGE approach is that the same set of primers can be used to analyse all eukaryotic transcriptomes. In contrast, for cDNA microarray techniques thousands of cDNA clones must be obtained, arrayed and maintained for each species to be tested. For oligonucleotide arrays, thousands of oligonucleotides must be synthesised, and different sets must be synthesised for each different eukaryotic species to be analysed. In general, the researcher is limited to those genes and species that are deemed desirable by the commercial firm marketing the technology.

4.3.6.4.4 Statistical significance

Over a wide range of expression levels, replicate RAGE determinations have a coefficient of variation of about 25 percent. Statistically this means that quadruplicate reactions should routinely be run in order to be able to detect changes in expression two-fold or greater. Although direct estimates of variability for microarray determinations are not routinely published, a general consensus seems to be that coefficients of variation for replicate trials are at least as high as 25 percent.

4.3.6.4.5 Expense

The RAGE technology is most advantageous in projects where analysis of medium numbers of genes, on the order of 50 to 1,000, is desired, especially if multiple samples must be analysed. No special equipment to make or analyse arrays is needed, and the reagents required number in the hundreds rather than in the thousands. As the number of genes assayed increases, the expenses associated with performing multiple PCR reactions (both time and materials) also increase. For high-throughput applications, specialised equipment (*e.g.*, robotic liquid handling equipment, or capillary electrophoresis with multi-channel fluorescent detection) is required.

The above discussion is framed in terms of microarrays *versus* RAGE as alternative technologies. Similar comparisons could be made pitting RAGE against SAGE or real-time RT-PCR. In fact, each technology has its own unique advantages and disadvantages, and a more fruitful way of approaching global gene expression analysis is to utilise the strengths of one technology to complement the limitations of another technique. For example, in many kinds of projects it is useful to first use a more comprehensive technology such as SAGE or cDNA microarrays to analyse a large number of genes in a small number of samples. From this analysis, a more focussed set of genes of interest can be obtained which can then be analysed more economically in multiple samples using RAGE.

Developmental projects currently underway to improve the RAGE technology focus on increases in throughput and sensitivity by utilising capillary electrophoresis as the analytical modality. Trials with a 16-capillary machine equipped for five colour detection suggest that good reproducibility can be obtained by running in-

dividual RAGE reactions with fluorescent RAGE primers, multiplexing four reactions, each containing a spectrally-distinct dye, and using a fifth dye as an internal loading standard. Detection by this method appears to be at least 250-times more sensitive than scanning of fluorescently-stained polyacrylamide gels (MacLeod, unpublished data), and a throughput approaching 3,000 assays per day should be attainable. In addition to supporting large scale projects involving species with partially defined transcriptomes (*i.e.* human, mouse, and rat), these high-throughput methods may allow the use of RAGE as a rapid way to define the transcriptome in eukaryotic species where little information is currently available. Assay of an entire transcriptome by RAGE is possible, and methods to increase the efficiency of information retrieval by RAGE are currently under development.

4.3.7

READS™ — Restriction enzyme analysis of differential sequences

4.3.7.1 Introduction

Restriction enzyme analysis of differential sequences (READS™) is a gel-based method for the analysis of differential mRNA expression that was described by Yatindra Prashar and Sherman M. Weissman in 1996. Similar to the differential display RT-PCR approach of Peng Liang and Arthur B. Pardee (Liang & Pardee, 1992), this method uses the polymerase chain reaction (PCR) to amplify cDNA fragments in order to allow detection of gene expression profiles. In contrast to differential display RT-PCR, READS™ PCRs are performed under more stringent conditions so that only the extreme 3'-end fragments of a restriction endonuclease-digested cDNA pool are amplified. This is done without physical separation of the cDNA fragments prior to PCR amplification. For different cDNAs, restriction digestion produces 3'-cDNA fragments of different sizes. Using different kinds of restriction endonucleases, it is possible to systematically analyse the complete mRNA pool of a eukaryotic cell as a pattern of bands derived from gel electrophoresis. In contrast to differential display RT-PCR, the resulting 3'-cDNA fragments are selectively amplified using a 5'-primer directed against a special adapter that is ligated to the 5'-ends of the 3'-cDNA fragments and a special set of PCR primers that are also used for cDNA synthesis. A comprehensive analysis of differentially expressed mRNAs between different physiological states of cells or tissues is performed by comparing the gel patterns of amplified 3'-end cDNA fragments derived from different samples. As with for the differential display RT-PCR approach, no prior knowledge of the transcript sequences is needed (Prashar & Weissman, 1996 and 1999). The READS™ technology is currently available as a service provided by Gene Logic, Inc. (Gaithersburg, MD, USA; information about the technology and the gene expression profiling service is available at <http://www.genelogic.com>).

4.3.7.2 Principle and basics

Figure 4.62 provides a schematic illustration of the READS™ procedure. In the first step, total RNA is isolated from the cells or tissues of interest. At this point, it is important to obtain RNA that is free of any genomic DNA contaminations because any residual DNA may be amplified during the subsequent PCR amplification leading to wrong results. Therefore, it is recommended to treat total RNA with RNase-free deoxyribonuclease I prior to further processing. Different subsets of

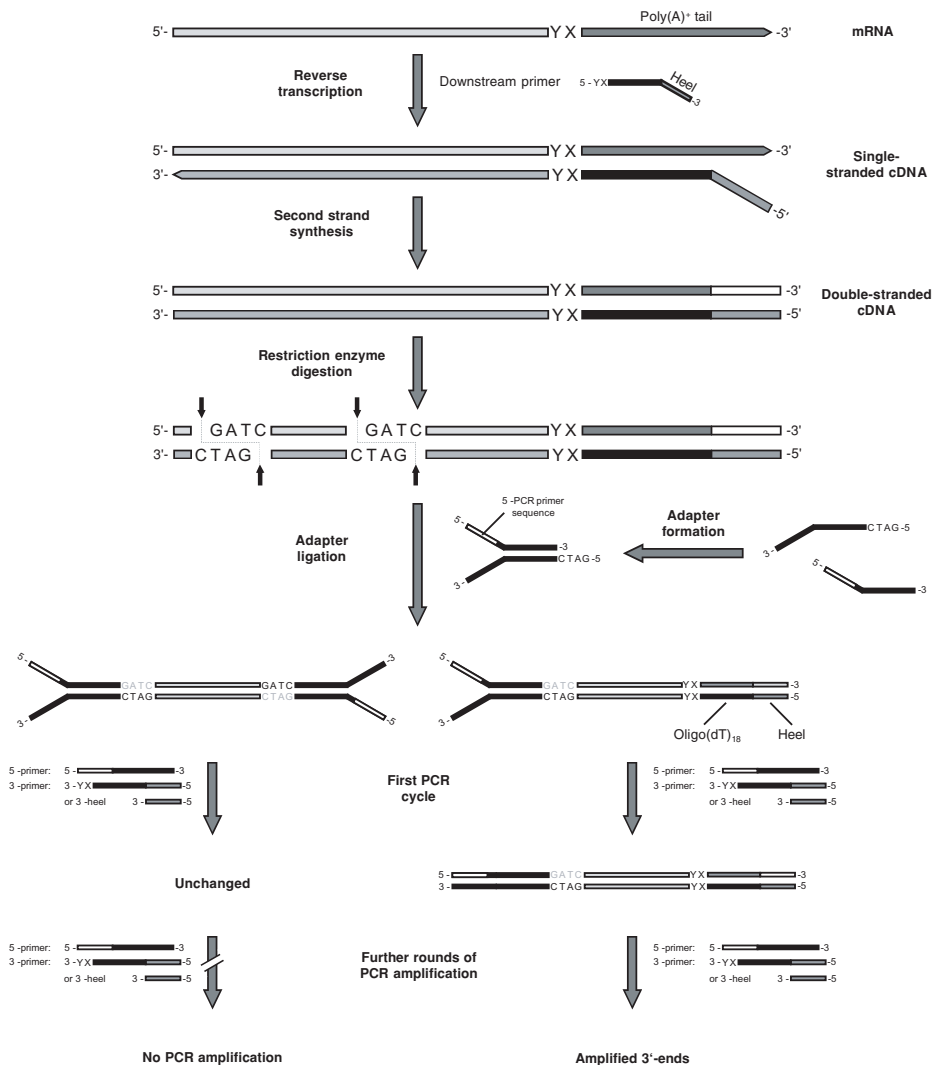


Figure 4.62 Principle of the restriction enzyme analysis of differential sequences (READS™) procedure. See text for details. The figure is adapted and modified from Prashar & Weissman (1996 and 1999) with permission from *Proceedings of the National Academy of Sciences of the USA*.

For the selective subsequent PCR amplification of the 3'-cDNA fragments, a Y-shaped double-stranded adapter, called fly adapter, with a single-stranded GATC 5'-overhang at one site is formed (see figure 4.63 and table 4.8). For this purpose, the two oligonucleotides FA-1 and FA-2 (see table 4.8) are annealed to form the Y-shaped adapter primer. This adapter primer has three distinct regions. On the 5'-end, the adapter has a stretch of non-complementary sequences on the opposite strands that give rise to the Y shape of the adapter because this part is not able to form a double strand (see figures 4.62 and 4.63). In the middle portion, the sequences of both oligonucleotides are complementary on the opposite strands and can form a double strand. On its 3'-end, the adapter has a four-nucleotide 5'-overhang that is complementary to that generated by the six-nucleotide-recognising restriction enzyme, which makes a staggered cut at the site of the cDNA. By changing the four nucleotides on the oligonucleotide FA-2, overhangs can be generated for any restriction enzyme of interest. It is noteworthy that there are some restriction endonucleases that are not denatured by heat subsequent to digestion. Therefore, it is desirable to design the overhang of the adapter oligonucleotide FA-2 such that once the fly adapter is ligated to the cDNA it does not regenerate the enzyme site.

by restriction endonuclease digestion using *Bgl* II, *Bcl* I, or *Bam* HI enzyme). The figure is reproduced with permission of *Methods in Enzymology* from Subrahmanyam *et al.* (1999).

Table 4.8 Primers used for cDNA synthesis, adapter formation and PCR amplification in the READS™ procedure used in the study described by Prashar & Weissman (1999). Please note that the oligonucleotides used in different studies based on READS™ may be different in their sequences.

Oligonucleotide name	Sequence	
	Heel	XY
RP 8.0 AA	5'-TAA	TTT
RP 8.0 AC	5'-TAA	TTT
RP 8.0 AG	5'-TAA	TTT
RP 8.0 AT	5'-TAA	TTT
RP 5.0 CA	5'-CTC	TTT
RP 5.0 CC	5'-CTC	TTT
RP 5.0 CG	5'-CTC	TTT
RP 5.0 CT	5'-CTC	TTT
RP 6.0 GA	5'-TGA	TTT
RP 6.0 GC	5'-TGA	TTT
RP 6.0 GG	5'-TGA	TTT
RP 6.0 GT	5'-TGA	TTT
RP 8.0 heel	5'-TAA	TTT
RP 5.0 heel	5'-CTC	TTT
RP 6.0 heel	5'-TGA	TTT
FA-1 (fly adapter 1)	5'-TAG	TTT
FA-1 (fly adapter 1)	5'-GAT	TTT

In the next step, a selective PCR amplification of the 3'-cDNA fragments is performed. In this PCR, primers are used that are either identical to the primers RP 5.0 XY, RP 6.0 XY or RP 8.0 XY used for cDNA synthesis (see table 4.8) or that are identical to the heel of these primers (RP 5.0, RP 6.0 or RP 8.0; see table 4.8) that serve as the 3'-primer depending on the cDNA subset used while the FA-1 oligonucleotide 5'-TAG CGT CCG GCG CAG CGA C-3' that is a part of the non-complementary stretch of the Y-shaped fly adapter and serves as the 5'-primer. During this PCR, only the 3'-end cDNA fragments are amplified, because the 5'-end cDNA fragments lack a sequence complementary to the reverse primers RP 5.0 (XY), RP 6.0 (XY) or RP 8.0 (XY).

During this PCR step, only the 3'-end restriction fragments of the cDNA are amplified that have the Y-shaped adapter ligated on their 5'-ends and have a heel sequence on the 3'-end. In contrast, there is no amplification of internal cDNA fragments because they lack the heel sequence while the 5'-primer for which the complementary sequence does not exist on the antisense strand cannot anneal to either one of the Y-shaped adapters ligated on both ends of the internal cDNA fragment. During the first PCR amplification cycle, the 3'-primer anneals to its complementary heel sequence of the 3'-end cDNA fragment on the sense strand and initiates DNA synthesis, which extends the new strand into the Y region of the adapter (figure 4.63), thus synthesising a new sequence that is complementary to the non-complementary sense strand region of the Y adapter that is the same sequence as that of the 5'-PCR primer. Consequently, the 5'-PCR primer can now anneal to this newly synthesised sequence of the strand. In the subsequent PCR cycles, both of the primers continue to amplify only the 3'-end of the cDNA fragments.

Adaptor primers used for the PCR can be radioactively labelled using, for example, ³²phosphorus, thus allowing detection of PCR products by auto-radiography. The PCR fragments derived from different samples are separated electrophoretically on polyacrylamide gels and cDNA bands showing different intensities in two samples are extracted from the gel, re-amplified, directly sequenced or sub-cloned and sequenced (Liang & Pardee, 1995; Prasher & Weissman, 1999; Werle *et al.*, 1994).

4.3.7.3 Typical results

In a first approach, the READS™ method was established with respect to optimisation of reaction conditions using untreated Jurkat cells or Jurkat cells activated by incubation with phorbol 12-myristate 13-acetate (PMA) plus phytohemagglutinin for four hours (Prashar & Weissman, 1996). In this experiment, 700 bands representing differentially expressed genes were identified (figure 4.64). About three to four percent of all fragments were up-regulated whereas about two percent were down-regulated. The authors sub-cloned a total of 16 bands that were sequenced and examined by RT-PCR. Of these, 15 were verified to be differentially expression in untreated and activated Jurkat cells using semi-quantitative RT-PCR (figure 4.65). On average, 100 to 150 discrete bands per lane can be evaluated visually. The size of the fragments analysed is in the range of 75 to 600 base pairs.

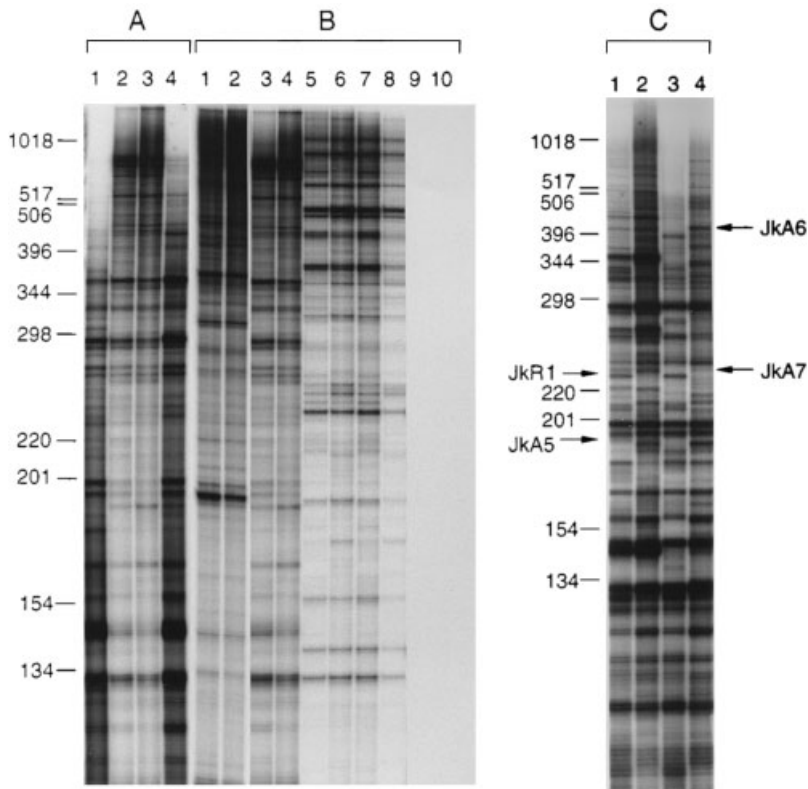


Figure 4.64 Reproducibility of display patterns and their diversity generated by different restriction enzymes and anchors in oligo(dT)₁₈ primers with a heel as shown in table 4.8.

A) Reproducible patterns were observed on restriction enzyme analysis of differential sequences (READS™) display gels of *Bgl* II-digested cDNAs prepared by using oligo(dT) primer RP 6.0 XY from RNA samples derived from untreated (lanes 1 and 2) and activated (lanes 3 and 4) Jurkat cells isolated in two separate experiments. **B)** Lanes 1, 3, 5, 7, and 9: cDNA samples derived from untreated cells; lanes 2, 4, 6, 8, and 10: cDNA samples derived from activated Jurkat cells. Different READS™ display patterns were produced on *Bgl* II digestion of cDNAs prepared by using different anchor bases (5'-CAG GGT AGA CGA CGC TAC GCT₁₈ GA-3' in lanes 1 and 2; 5'-CTC TCA AGG ATC TTA CCG CTT₁₈ AT-3' in lanes 3 and 4). However, digestion of the cDNAs using the latter primer with different restriction enzymes produced different display patterns (*Bgl* II in lanes 3 and 4; *Bam* HI in lanes 5 and 6).

Aliquots of cDNAs produced using the latter primer digested with *Bam* HI in separate experiments showed consistent patterns between samples derived from untreated (lanes 5 and 7) and activated (lanes 6 and 8) Jurkat cells when ligated to adapter and PCR amplified. However, in the absence of adapter ligations, no pattern was generated (lanes 9 and 10). **C)** Complementary DNAs derived from samples from untreated (lanes 1 and 3) and activated (lanes 2 and 4) Jurkat cells using primer 5'-CTC TCA AGG ATC TTA CCG CTT₁₈ AT-3' are displayed upon digestion with *Bgl* II (lanes 1 and 2), while in lanes 3 and 4 *Bgl* II-digested and adapter ligated cDNA was re-digested with restriction enzyme *Hin* fi before PCR amplification. Arrows point to bands revealed as true differences on re-cutting. Numbers shown on the left represent the size of the cDNA fragments in base pairs. The figure is reprinted from Prashar & Weissman (1996) with permission from *Proceedings of the National Academy of Sciences of the USA*.

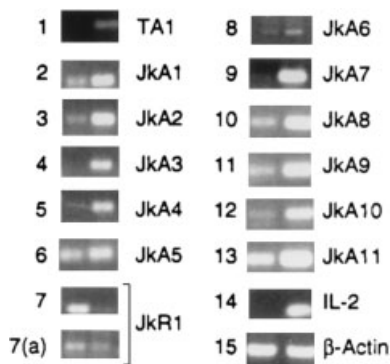


Figure 4.65 Analysis of differences from restriction enzyme analysis of differential sequences (READS™) display gel by semi-quantitative RT-PCR of total RNA samples to verify the results of the READS™ display gel. In each of the panels 1 to 15 the left lane represents cDNAs derived from untreated cells while the right lane represents cDNA derived from activated Jurkat cells. In panel 7, but not in panel 7a, cDNA derived from peripheral blood T-cells was used for RT-PCR. Bands representing differentially expressed mRNAs were extracted from the READS™ display gel, sub-cloned, sequenced, and mRNA-specific RT-PCR primers were made. RT-PCR samples were analysed on an agarose gel and stained with ethidium bromide. The figure was reproduced with permission from *Proceedings of the National Academy of Sciences of the USA* (Prashar & Weissman, 1996).

In a second approach, Newburger *et al.* (2000) analysed the expression profiles of human neutrophils in comparison to neutrophils exposed to opsonised non-pathogenic *Escherichia coli* K12 using READS™. In contrast to the classic view of the neutrophil that portrays it as a short-lived, terminally differentiated cell with a highly condensed and therefore relatively inactive nucleus, the authors found an unexpected complexity of the gene expression pattern in the neutrophils. Several hundred mRNA species (about eight percent of all displayed cDNAs) showed significant changes in their expression two hours after introduction of the bacteria. Interestingly, this level of change was thought to be comparable to the modulation of gene expression that occurs, for example, on activation of human T-cells as demonstrated by Prashar & Weissman (1996).

To complete these studies, the patterns of gene expression in neutrophils obtained after exposure to pathogenic and control gram-negative bacteria such as wild-type *Yersinia pestis* strain KIM5 (Kurdistan Iran man), the gram-negative bacterium that causes bubonic plague, and non-pathogenic *Yersinia pestis* strain KIM6 lacking a 70-kilobases virulence-encoding plasmid were examined. In each case, the displayed pattern of mRNA expression was different in part for each type of bacteria tested.

The results of the latter comprehensive studies have been described in more detail by Subrahmanyam *et al.* (2001). In all, 17 different restriction enzymes were used to produce the gene expression patterns using the READS™ technique, and the fragments from each enzyme digest were amplified with each of the twelve possible 3'-terminal dinucleotides on the oligo(dT) primer (see table 4.8). Using these samples, a total of 1,887 bands were sequenced of which about 80 percent revealed sequences with sufficient quality to give high probability matches to known sequences (for a detailed view of the distribution of categories of these transcripts, see Subrahmanyam *et al.*, 2001). In total, 350 known genes and 292 expressed sequence tags or unknown sequences were found to be differentially expressed by two hours after incubation of neutrophils with bacteria.

Within the two hours of treatment with bacteria, changes in the level of expression of mRNAs were observed including those for a variety of cytokines, receptors, apoptosis-regulating proteins, and membrane trafficking regulators. In addition, a large number of mRNAs was up-regulated that appear to represent a common core of activation response genes that have been identified as early response products to a variety of stimuli in a number of other cell types. The activation response of neutrophils to non-pathogenic bacteria was greatly altered by exposure to *Yersinia pestis*, which may be a major factor contributing to the virulence and rapid progression of plague. The findings indicated that active regulation of gene expression plays a major role in the neutrophil contribution to the cellular inflammatory response and that neutrophils are capable of extensive, rapid, and complex changes in gene expression. Supplementary information on these studies is available in Subrahmanyam *et al.* (2001) and on the website of the Department of Molecular Biophysics and Biochemistry and the Department of Genetics (Yale University, School of Medicine, New Haven, Connecticut, USA) at <http://bioinfo.mbb.yale.edu/expression/neutrophil/>.

In a third approach, the changes in mRNA expression during the process of mouse promyelocytic cell maturation to neutrophils were analysed in a cell line derived from murine bone marrow cells that is arrested developmentally by a dominant-negative retinoic acid receptor (Lian *et al.*, 2001). An extensive catalogue was prepared of the changes in mRNA expression profiles that occur during morphologic maturation of the bone marrow cells to mature neutrophils in the presence of all-*trans* retinoic acid using the READS™ technique complemented by analyses performed with oligonucleotide arrays. Cellular mRNA derived from the mouse promyelocytic cell line was analysed at 0, 24, 48, and 72 hours after treatment with all-*trans* retinoic acid. Nine restriction enzymes were used in this READS™ approach. During differentiation, 1,109 fragments corresponding to 837 transcripts were found to be differentially expressed. These represented approximately 279 known genes, 112 expressed sequence tags, and 59 putative new genes, each with a perfect or fair polyadenylation signal at an appropriate distance from the oligo(dT) priming site.

In a few cases, discrepancies were found in the pattern of expression of an mRNA, as estimated by READS™ and by oligonucleotide array analysis. The authors examined the six most extreme cases and analysed the levels of change in mRNA expression for these genes by Northern blot analysis. In five cases, the Northern blot results agreed with the results of the READS™ analysis, whereas the results of one mRNA agreed with the oligonucleotide array but duplicate bands derived from READS™ showed a relatively high level of expression in the zero hours sample that did not correlate with the Northern blot. Figure 4.66 shows Northern blot analyses of eleven mRNAs identified by the approach described by Lian *et al.* (2001).

Among the differentially expressed genes, eight cytokines and chemokines, 52 mRNAs encoding receptors, 29 mRNAs encoding several types of granules, and 13 transcription factors were identified by one or the other method. A number of the receptors of mature neutrophils showed a prominent induction of their

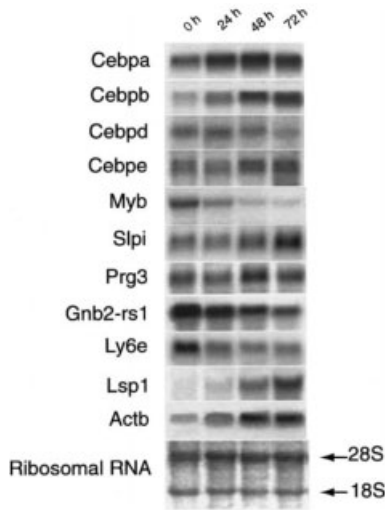


Figure 4.66 Northern blot analysis of selected differentially expressed mRNAs identified by restriction enzyme analysis of differential sequences (READS™).

Equivalent amounts of total RNA derived from a mouse promyelocytic cell line induced by all-*trans* retinoic acid at different time points (0, 24, 48, and 72 hours) were resolved by formaldehyde/agarose gel electrophoresis, stained with ethidium bromide to verify the amount of loading. Eleven mRNAs were separately probed on the filters. The gene symbol of each probe is listed at the left of a related Northern blot result. One of the RNA-blotted membranes is shown with methylene blue-stained 28S and 18S RNA subunits demonstrating the quality and quantity of RNA loaded in individual lanes. The following gene symbols are used in this figure: *Actb*, cytoplasmic β -actin; *Cebpa*, CCAAT/enhancer binding protein (C/EBP) α ; *Cebpb*, CCAAT/enhancer-binding protein (C/EBP) β ; *Cebpδ*, CCAAT/enhancer-binding protein (C/EBP) δ ; *Cebpe*, CCAAT/enhancer-binding protein (C/EBP) ϵ ;

Gnb2-rs1, guanine nucleotide-binding protein β -2-related sequence 1; *Lsp1*, lymphocyte-specific 1 (S37 or pp52); *Ly6e*, lymphocyte antigen 6 complex, locus E; *Myb*, myeloblastosis oncogene; *Prg3*, proteoglycan 3 (eosinophil major basic protein 2); *Slpi*, secretory leukocyte protease inhibitor. The figure is reproduced from Lian *et al.* (2001) with permission from *Blood*.

mRNA late in maturation. Rarely, mRNA encoding receptors was down-regulated. This is consistent with previous findings that myeloid maturation is accompanied by an increased responsiveness of the cell to external stimuli (Lian *et al.*, 2001). Neutrophils also express several types of granules that develop at different stages of myeloid maturation (Gullberg *et al.*, 1999; Lawson *et al.*, 1998; Sigurdsson *et al.*, 1997). Levels of all mRNAs encoding primary, secondary and tertiary granule proteins increased as the cells matured. The changes in certain transcription factors, such as the moderate down-regulation of *myb* and *myc* and the up-regulation of the *Max* dimerisation protein, or *MAD*, were consistent with the shift of the cells from a proliferative to a differentiated state (Amati *et al.*, 1994). In addition, the increase in mRNA expression of members of the CCAAT/enhancer-binding protein (C/EBP) family during late maturation is consistent with previous findings (Morosetti *et al.*, 1997; Tenen *et al.*, 1997; Yamanaka *et al.*, 1998). Some other changes in the mRNA expression of transcription factors were more difficult to explain and the differential expression of several transcription factors showed no conclusive characteristics (Lian *et al.*, 2001).

A database (dbMC) of genes or expressed sequence tags that were differentially expressed during myeloid differentiation was also constructed that included two sub-databases, dbMCd (information detected by READS™) and dbMCa (information derived from the oligonucleotide arrays). This database includes GenBank matches, Locus Link or UniGene clusters, expression patterns, tissue distribution,

synonymous protein or gene name(s), notations of possible functions, poly(A) signal sites, information on sequence quality, and hyperlinks to the database searches, sequence trace files, and related references. This supplementary information is available at <http://bioinfo.mbb.yale.edu/expression/neutrophil>.

In addition, mRNA expressions data were compared to protein expression data derived from two-dimensional high-resolution gels of bone marrow control cells and cells differentiated for 72 hours in the presence of all-*trans* retinoic acid. Fifty protein spots whose relative intensity changed prominently during differentiation were examined by mass spectrometry. The results suggested a poor correlation between mRNA expression and protein abundance, indicating that it may be difficult to extrapolate directly from individual mRNA changes to corresponding ones in protein levels as estimated from two-dimensional gel electrophoresis (see also chapter 1). Of these, 19 were found in the dataset of mRNA expression; the mRNA for five genes was reported as absent, and 13 genes were present during mouse promyelocytic cell differentiation. Comparison of the expression patterns revealed only four genes of 18 present on the oligonucleotide arrays whose expression was consistent at the mRNA and protein level. None of these was on the list of the genes that were differentially expressed significantly (five-fold or greater change by oligonucleotide array analysis or two-fold or more by READS™).

4.3.7.4 Refinements of the READS™ procedure

A complete transcriptome of a cell or tissue is thought to consist of at least 12,000 to 15,000 different transcripts. Using the READS™ approach described here, it is not possible to cover the complete transcriptome because each six-base cutting restriction enzyme cuts only about eight percent of the cDNAs at positions between 50 and 400 bases from the poly(A) tract so that about ten to twelve six-base cutting restriction enzymes will be needed to approach complete representation of each cDNA once, each being used with several different anchored oligo(dT) primers (Prashar & Weissman, 1996). Therefore, for a comprehensive analysis of differential mRNA expression of a cell or tissue twelve to twenty-four different six-base-cutting restriction endonucleases should be used (Prashar & Weissman, 1999).

However, Yatindra Prashar and Sherman M. Weissman (1996) estimated that, at best, 100 to 150 discrete bands can be detected in a single gel lane. Therefore, at least 100 lanes need to be run under different conditions to study the overall pattern of gene expression in any single cell type.

A modification of the READS™ method described above helps to identify false-negatives (Prashar & Weissman, 1996). Overlapping bands that mask true differences can be resolved by re-cutting the adapter ligated 3'-cDNA fragments before PCR amplification. If a site for a restriction enzyme used for re-cutting is present in one of two co-migrating bands, it will be cut into one fragment with an adapter and the other with the heel, which cannot be amplified by PCR and will therefore be eliminated.

Using this modified READS™ approach, additional differences between untreated and activated Jurkat cells were identified when cDNA digested with *Bgl* II and ligated to the adapter was further digested with *Hin* fl. The choice of enzymes for primary or secondary cutting is arbitrary although a combination of more than one enzyme can be used for re-cutting. Another advantage of re-cutting is that recovery of low-abundance cDNAs is enhanced because removal of high-abundance bands by re-cutting allows to amplify these fragments for a further round without influencing their relative abundance. Re-cutting can also be used to minimise redundancy between fragments in different lanes. cDNA fragments digested with *Bgl* II can be re-cut with *Bam* HI and *vice versa*, so that the two samples share no amplified products. Confirmation that a band corresponds to a known cDNA sequence can be obtained by re-cutting the cDNA with appropriate restriction enzymes prior to amplification. A large number of variations of the display patterns can therefore be produced using this method in order to look for differentially expressed genes by (i) a combination of a number of different two-base anchored oligo(dT) primers with a heel for making cDNAs, (ii) a number of different restriction enzymes that can be used for primary cutting of these cDNAs, and, (iii) the number of restriction enzymes used for secondary cutting for each primary cut.

In the original procedure described by Prashar & Weissman (1996 and 1999) twelve individual cDNA preparations are used due to the twelve different combinations of the dinucleotides X and Y of the primers that are used for generating different subpopulations during cDNA synthesis (see table 4.8). To avoid the preparation of twelve individual cDNA populations of every RNA sample of interest by priming with each of the twelve primers, RNA can be primed with a modified oligo(dT)₁₈V primer (where V is adenine, cytosine or guanine, but not thymidine). Using this modification described by Subrahmanyam *et al.* (1999), it is necessary to make only one cDNA preparation of every RNA sample. At the PCR stage, each of the twelve XY-anchored primers shown in table 4.8 is used to set up twelve individual PCRs. This modification reduces the amount of total RNA that is required for analysing an mRNA expression profile using READS™.

4.3.7.5 Summary and discussion

In principle, READS™ is a modification of the differential display RT-PCR developed by Liang & Pardee (1992). The refinements of the READS™ technique make it a sensitive tool leading to reproducible results. It commonly allows detection of relative changes in the level of mRNA expression of the order of 1.5-fold to two-fold at the lower limit. At the upper end, relative differences in abundance of three to four orders of magnitude can be measured. The most important advantage of the READS™ procedure is that it is an open system, thus allowing the identification of known and unknown genes.

The most abundant mRNAs observed in the studies of Subrahmanyam *et al.* (1999) are unlikely to represent more than one to two percent of the total cellular mRNA. The authors therefore suggested that the READS™ approach makes it pos-

sible to detect bands representing mRNAs present at less than one copy per 100,000 copies of total mRNA, corresponding to approximately one mRNA molecule per cell for a typical cell, thus allowing the identification of low abundant mRNAs that are differentially expressed.

As is true for all differential display variants, READS does not allow measurement of absolute amounts of individual mRNAs. The signals of the most abundant cDNAs, in which self-annealing competes with primer annealing to cDNA to a much greater extent than with rare mRNAs, are relatively reduced by the resulting decrease in the amplification probability per cycle of PCR (Mathieu-Daude *et al.*, 1996). In addition, longer cDNAs are amplified by PCR less efficiently than shorter cDNA and amplification of cDNAs of similar length may vary because of base composition.

All methods based on oligo(dT) priming for cDNA generation bear the risk that some of the resulting cDNA will be generated from intranuclear RNA species, including incompletely spliced product and even intergenic transcripts. Priming may also occur at poly(A) tracts often present in interspersed repetitive sequences such as Alu elements or in ribosomal RNAs (Schmid, 1996; Weiner *et al.*, 1986).

In summary, READS™ is an elegant strategy to identify differentially expressed mRNA. It bears the same risks that have been proposed for differential display RT-PCR and related methods. READS™ technology is provided by Gene Logic, Inc. (Gaithersburg, Maryland, USA) as a commercial service to analyse gene expression.

4.3.8

TOGA – Total gene expression analysis

4.3.8.1 Introduction

In the era of genomics, with most of the genome sequence known, we require methods that can detect not only the known genes but also those that have not yet been added to the complete list. We require accurate quantification of all messenger RNAs, but particularly those that are expressed in the one to twenty copies per cell range, because these represent the bulk of the expressed genes. We require high-throughput methods that can allow us to address several related questions simultaneously, so as to direct our attention to those molecules that are most correlated with the processes which we wish to understand. Given the expansion of biological research in the last two decades, we require ways to link any discovery that our robot might make to the sequences that are in databases and in all of the published literature. The TOGA method described below was first developed as a manual technology. Digital Gene Technologies subsequently automated the process and built the informatics, which accomplish the goals mentioned above. The details of this method have been published (Lo *et al.*, 2001; Sutcliffe *et al.*, 2000 and 2001).

4.3.8.2 The TOGA method

TOGA, an acronym for total gene expression analysis, investigates the complete set of messenger RNAs in the cell (figure 4.67). It begins with the isolation of total RNA – 20 nanograms of total RNA is sufficient. Synthesis of cDNA is applied with a pool of primers that anneal to the very beginning of the poly(A) tail on the mRNAs in each total RNA sample. After cDNA synthesis, which begins on each RNA from a fixed point, the cDNAs are cleaved with a restriction endonuclease that uses a four-nucleotide recognition sequence. In this example, *Msp* I is shown, which recognises the sequence CCGG. Since CCGG is a frequently observed sequence within mRNAs, the enzyme cleaves most cDNAs at several positions along their length. Next, the 3'-fragments (all those cDNA pieces with the poly(A) tail primer attached) are isolated and ligated to a short primer-binding sequence at the 5'-end which allows to use that product in subsequent polymerase

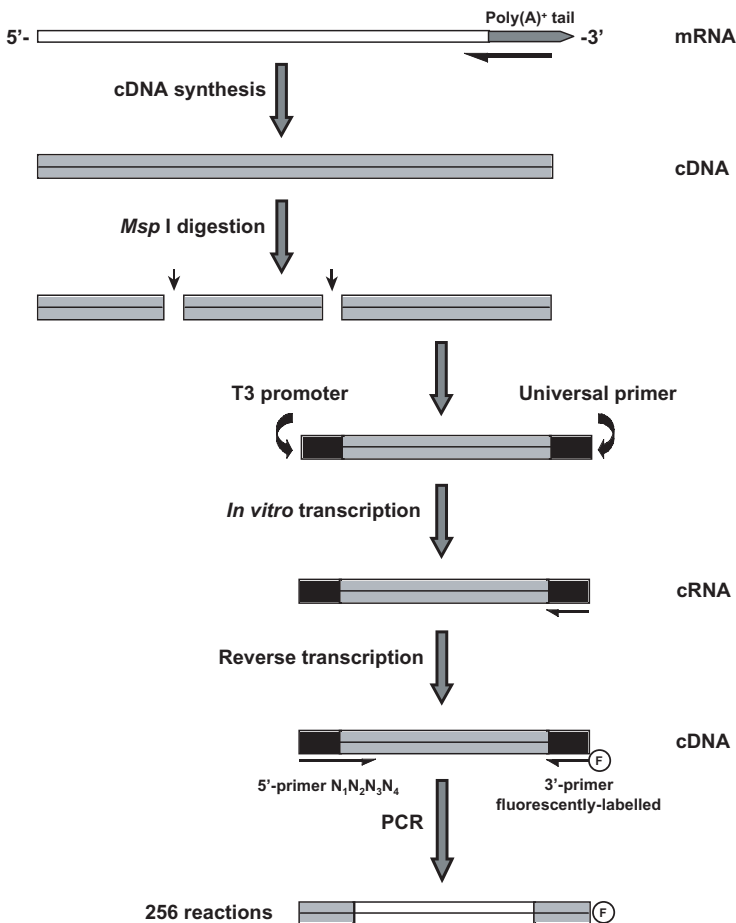


Figure 4.67 The total gene expression analysis (TOGA) method. A diagrammatic outline of the steps in the TOGA method, adapted from Sutcliffe *et al.* (2000).

chain reaction steps. The important step is the very last one, in which the binding site primer is extended by four nucleotides into the cDNA inserts. At each position, there could be either adenine, cytosine, guanine or thymine, and so therefore there are $4 \times 4 \times 4 \times 4$, or 256, primers which are used individually. These 256 reactions are done using primers with a fluorescent primer on the 3'-end to generate fluorescently-labelled PCR products from the starting cDNAs.

A thought experiment, if one imagines that in the starting population of RNAs there were 10,000 different mRNA species, by following these steps those 10,000 species are sub-divided into 256 pools, each pool containing approximately 40 members. These 40 distinct products are a small enough number that their lengths can be individually measured and recorded into a database.

To summarise all the important principles of the method: TOGA addressing begins with the four nucleotides that are recognised by a restriction endonuclease, which cleaves the RNAs near the 3'-end. They are sorted into 256 pools by virtue of the four nucleotides adjacent to the cleavage site. That provides eight nucleotides of sequence. Because the cDNA was originally cleaved at a discrete site upstream of the poly(A) tail, when we include the length, we get an address for each RNA. The identity, or address, of the eight nucleotides plus the length of the fragment (to the poly(A) tail), is sufficient information to recognise an RNA if its sequence has already been deposited in a database. Because of the digital addressing of each sequence tag, these are referred to as DSTs, for digital sequence tags.

An important aspect of the TOGA process is the use of automation; robots conduct nearly every step in the process. With these robots, it has been possible to optimise all 256 PCRs, so that each one is absolutely specific for the four parsing nucleotides. If one were to use only a single enzyme, as in the *Msp* I example, then only approximately 65 % of the messenger RNA would be detected, which would leave 35 % undetected. TOGA repeats the process with a total of four iterations, so that with the second enzyme we detect an independent 65 %. A third enzyme detects a third independent 65 %, and so on, so that after four iterations, one has sampled greater than 98 % of the RNAs with at least one of the enzymes, most of them with three or four. The products are separated by capillary electrophoresis, each product fragment is detected by fluorescence and its length and intensity are recorded. The laboratory information management system (LIMS) that manages all of the robotic reactions knows what is in each lane in the capillary electrophoresis analysis, that means it knows what the RNA sample was, it knows what the primer was, it knows what the eight nucleotides at the 5'-end are, so that when the fragment is detected, its length and fluorescence intensity are recorded into a file. Each fragment is addressed by the TOGA address, the digital address. That automatically generated database is cumulative; each new sample that goes in provides a new data point. The data are easily queried because they are all electronically filed. Because of the simplicity of the TOGA address, it is possible to use that information to link instantaneously any newly-detected RNA with the GenBank database, as will be explained below.

The robot is built around an arm which collects 96-well bar-coded reaction trays from a rack. There is a barcode reader which logs each tray into the LIMS, and then

transfers it to a pipettor. The substrates, the PCR primers, and the reagents, are brought to the pipettor in bar-coded trays by the arm and validated by the LIMS. After the reactions are assembled, they are brought to a plate-sealing device that puts a gasket on top of each plate so that there is no evaporation or cross-contamination during the PCR. The trays are next brought to the thermocycling centre, where they are loaded into the thermocyclers. The LIMS knows the thermocycling program for each reaction and therefore knows how to program each of the thermocyclers. When the reactions are completed, the arm collects the trays and they are moved to a capillary electrophoresis system. The products in each reaction are measured by length and the database of lengths is created automatically.

The data appear as in figure 4.68.A: for each RNA sample, there are 256 panels of this sort, and in figure 4.68.B the same panel for each of 8 different samples. This is a study in which mouse brain has been dissected into eight regions: the hypothalamus, the hippocampus, the cerebral cortex, the amygdala, the brainstem, the striatum, the cerebellum, and the olfactory bulb. TOGA was run on each of those eight dissected regions, generating 256 PCR product sets times eight samples is shown. The data were automatically put into the database. After inspecting the data, all the analyst needs to do is ask a question: "Show me all the RNAs whose concentration in the hypothalamus is at least four-fold the concentration in the brainstem." From the TOGA Portal (described below), we learn that there are 53 members that meet that criteria, out of over 15,600 detected expressed genes. The first one shown has the TOGA address AGTC168. A candidate identity, shown by a mark at the top of the panel at that address, indicates the mRNA encoding preprohypocretin, a known hypothalamic RNA (de Lecea *et al.*, 1998).

The icons atop each TOGA panel (figure 4.68.A) indicate candidate gene identities that have been compiled by running a 'virtual TOGA' analysis of all known mRNA transcript sequences in the database. For each sequence, the *Msp* I site nearest to the 3'-end was identified, along with the 3'-adjacent four bases. The distance to the start of the poly(A) tail was calculated, and a TOGA address is then assigned. In the case of preprohypocretin, the TOGA address in the *Msp* I data set, AGTC168, reflects the four base parsing primer sequence that follows the CCGG *Msp* I recognition sequence, and the 168 base distance from this site to the poly(A) tail.

In situ hybridisation studies have shown that preprohypocretin is expressed in a few thousand neurons of the dorsal lateral hypothalamus and nowhere else in the brain (de Lecea *et al.*, 1998). That is a tiny percentage of the neurons in the mouse hypothalamus. The RNA from those few thousand neurons is sufficient to give a fairly substantial TOGA signal in the hypothalamus panel (figure 4.68.A), indicative of the sensitivity of the TOGA method. This study also found many other known hypothalamic-specific RNAs, including those encoding vasopressin and oxytocin in the next two panels (figure 4.68.B), as well as more than twenty hypothalamic RNAs that were not represented in the database, and, thus, are potentially novel. Clearly, TOGA represents a powerful anatomical discovery tool.

4.3.8.3 The TOGA Portal

The analysis of TOGA experimental results is accomplished through the TOGA Portal, a web-based application that integrates gene expression data with information from genomics databases and provides a series of bioinformatics tools within a single graphical user interface. Figure 4.69 shows a screen capture of the TOGA Portal graphical user interface. Tools and utilities are organised into subsections for data analysis and review (analysis tools and review tools), experiment-related information (experiment info), keyword search, printing and data import/export functions (utilities), sequence analysis programs and LIMS access (DGTools) and paths to other bioinformatics software and databases (links). In the example in figure 4.69, the results are displayed in a list format, which illustrates the association of a TOGA address identifying an mRNA alongside the gene expression data measured in each of the experimental samples together with predicted DST candidates

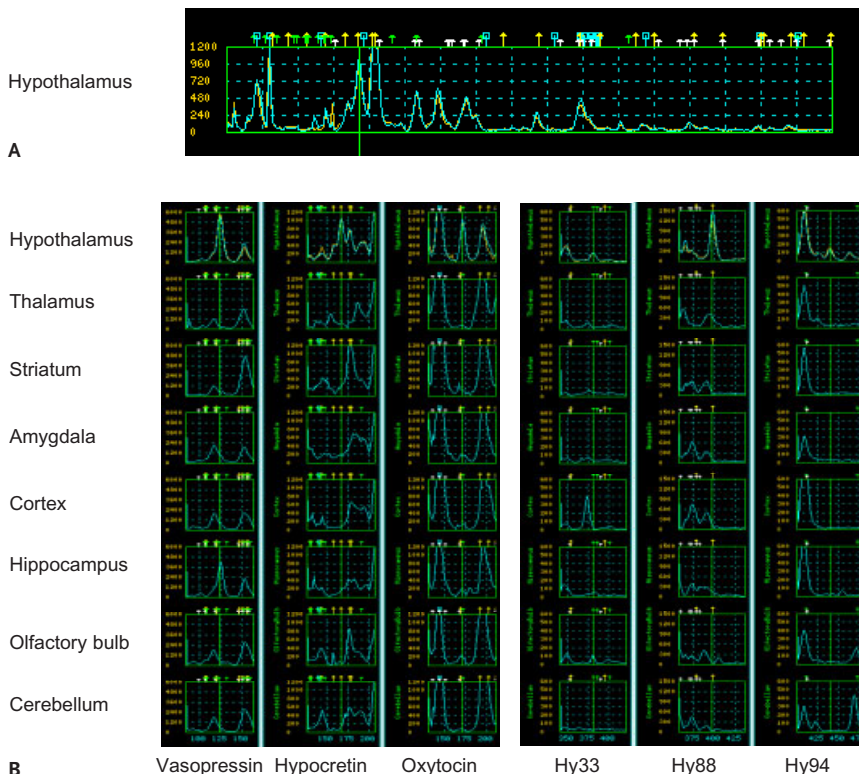


Figure 4.68 Total gene expression analysis (TOGA) mRNA expression profiles. A graphical display of TOGA data taken from one of the 256 PCR reactions (AGTC parsing bases) from a mouse hypothalamus RNA sample (figure 4.68.A). The vertical line drawn through the peak indicates the position of the preprophy-

pocretin digital sequence tags (DST). A comparison of expression profiles for three known (vasopressin, hypocretin, oxytocin) and three novel (Hy33, Hy88, Hy94) hypothalamic-enriched mRNAs across eight brain regions is shown in figure 4.68.B (Lo *et al.*, 2001; Sutcliffe *et al.*, 2001).

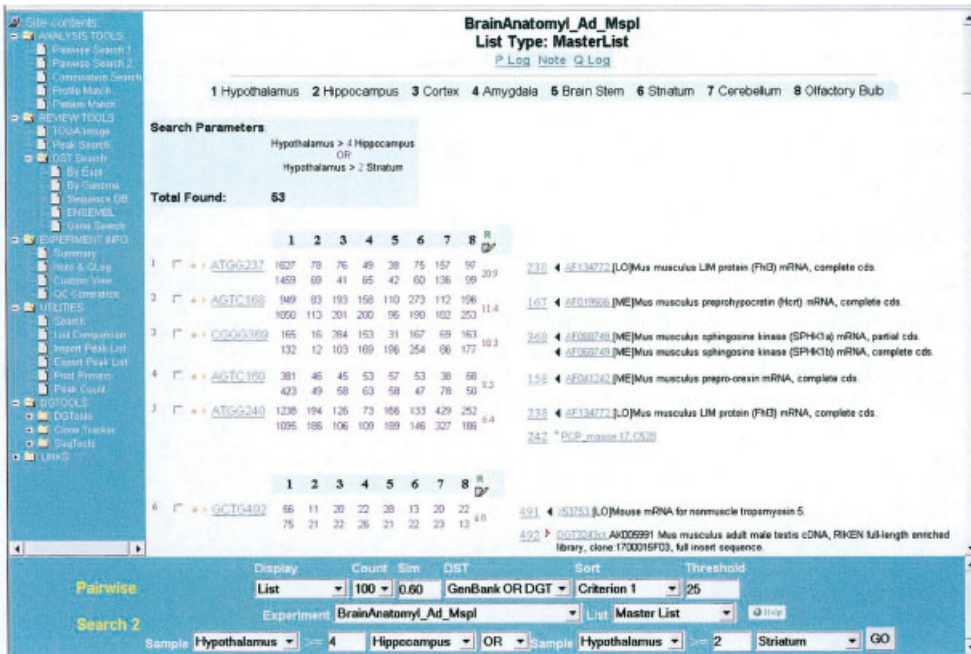


Figure 4.69 The Total gene expression analysis (TOGA) Portal. The tools directory is displayed on the left of the graphical user interface, showing folders for major sections. The tool window on the bottom portion of the screen layout contains pull-down menus and parameter setting boxes used to execute the indicated function (Pairwise Search 2) on selected data. The results of functions performed on the data are returned in the central data window and the experiment name and list type are displayed along with sample identities corre-

sponding to the data columns. Hyperlinks in the central window provide instant access to primary data for each peak (TOGA address hyperlink) and to GenBank database files, assembled EST clusters, or annotated DGT (Digital Gene Technologies) sequence (accession number hyperlink) for each listed candidate digital sequence tag (DST) sequence. Checkboxes alongside each list entry allow the user to flag and define new sublists for subsequent searches and analysis.

derived from sequence databases. The intuitive browser-style interface and bioinformatics tools create a framework within a single application for sophisticated, in-depth gene expression analysis.

In the TOGA Portal, the data sets are organised and accessed from a central database as individual experiments that contain defined sets of samples. The process for data analysis typically proceeds in an experiment-by-experiment fashion within the context of larger projects that are constructed from an interrelated set of experiments that might constitute, for example, several animal models of human cardiovascular disease. The results from profile and pattern searches on these single experiments, which occasionally identify as many as hundreds of differentially expressed genes, are further interrogated by employing comparison tools to allow examination of a gene's expression profile across multiple experiments within the

project. In addition, the data management tools of the TOGA Portal allow scientists to regroup, reorder, and combine samples to permit a more focused analysis on subsets of samples and genes. The TOGA Portal also enables virtual experiment creation, and these can be designed from any selected group of samples in the larger database that share the same species of origin and were processed using a common TOGA 5'-anchor enzyme (e.g., *Msp* I).

As with most large-scale gene expression studies, the starting point for conducting an analysis is the initial assessment of data quality. In TOGA, the automated data processing routines create a set of files that store statistical quality measures, including standard correlation coefficients, for duplicates or replicates of the same sample and for all cross-sample correlations that can be determined within an experiment. The calculated parameters also include a measure of duplicate variance, since the data are compiled from a series of 256 independent reactions for each sample processed in TOGA. Entire data sets from duplicates or comparable samples can also be viewed in traditional scatter plots to evaluate variance or highlight trends in the data. Duplicate data are displayed as overlaid traces for each of the 256 TOGA reactions in a graphical view, which serves as an excellent visual means to confirm data quality in the context of individual PCR reactions.

Visualisation of genome-scale experiments and their accompanying primary data represents a major challenge for gene expression informatics applications (Basset *et al.*, 1999). The analysis tools found in the TOGA Portal aim to reduce these data sets by detecting and bringing into focus the biologically relevant expression patterns in the experimental data. The group of review tools is designed principally for inspection of the primary TOGA data, which often occurs in parallel with analysis of differential expression profiles. In the TOGA Portal, the retrieval and display of data is driven from peak lists, which contain either entire experiment or sample data sets (master lists) or specific subsets that are created off of the master list during review sessions. Users can also import lists of addresses for known genes for review and can utilise search tools for retrieving data from single genes by name or TOGA address. The gene expression data for individual peaks or mRNAs are viewed in a variety of formats, including bar charts, pseudocolour representations, tabular charts, and the TOGA panel electropherogram view (see figure 4.68.A). In each case, analysts can select from data sets compiled pre or post-normalisation. In the graphical view, features such as point-and-click navigation, zoom, and *xy*-axis scaling are used to enhance details in the plotted data. The collection of visualisation tools in the TOGA Portal allows the review process to be customised to fit a scientist's particular needs or preferences.

The TOGA Portal's expression profile search and analysis tools provide investigators with a series of powerful techniques to explore experimental data and capture key results from the review process. A pairwise search strategy underlies several of these tools, and an example is provided in figure 4.69. In this particular experiment, samples from microdissected mouse brain regions were queried to identify mRNAs with high levels of expression in the hypothalamus relative to other structures. In the tool window, search parameters were set to enforce duplicate similarity (Sim) and a minimum expression value (threshold). Pull-down menus

were used to select from among eight samples in the experiment to run pairwise searches for finding gene expression differences between hypothalamus and hippocampus, with ratios set to detect mRNAs with four-fold or greater expression in the hypothalamus; a similar search at a two-fold level was entered for hypothalamus and striatum samples. The results from these searches were returned in the data window and indicated that the expression patterns of 53 mRNAs (out of approximately 15,600 with detectable expression in the hypothalamus) satisfied at least one of the two search criteria. Logical operators (AND, OR, NOT) can be employed on subsequent searches to find subsets of expression profiles matching both criteria or that are mutually exclusive of one another. For more complex queries, the combination search tool will generate an exhaustive series of pairwise searches that can be configured and launched, as well as recursively searched.

Another component of the TOGA data review process takes advantage of search capabilities that can detect user-defined patterns across a series of samples. These types of searches require input to derive a mathematical vector for execution. The pattern search tool provides a form for setting relative expression values in each of the samples, a typical example being a linear increase in expression over an experimental time course. Alternatively, an expression pattern of interest discovered by pairwise search methods can be used as input in the profile match tool. The profile searches run by all of these tools generate lists that can be saved, annotated and further compared and merged with existing results or directly exported into applications such as GeneSpringTM (Silicon Genetics, Redwood City, CA, USA) for further data mining purposes.

4.3.8.4 Comparison to other profiling methods

How does TOGA compare with the other expression profiling methodologies that are commercially available? First, there are the array, or 'chip', methods (Lockhart *et al.*, 1996; Schena *et al.*, 1995). TOGA is an open system; it looks at the RNAs, and has the RNAs report whether they are present or absent. By contrast, a DNA array is a closed experiment. It can report only on the sequences that have been arrayed on the DNA microarray. Therefore, arrays are applicable for known genes, and for species for which substantial genomic data are already available, but it is not a gene discovery tool. Because arrays depend on hybridisation, they therefore must follow the rules of hybridisation, and any given hybridisation is dependent on the sequence that is to be hybridised. There is a floor for sensitivity, and that floor is different for each RNA, with the range of sensitivity on any array being between one part in 100,000 and one part in 10,000, depending which RNA one is looking at. Arrays require about 1,000-fold more RNA than does the TOGA method, thus TOGA allows investigations using microdissected or cell-sorted material to get RNA profiles that are not possible with DNA microarrays.

The SAGE (serial analysis of gene expression) method is used extensively in the cancer field (Velculescu *et al.*, 1995). This method relies on sequences tags to indicate the frequency of any given expressed gene. From the 3'-ends of RNAs, fragments usually fourteen nucleotides in length are isolated, and then polymerised.

The polymerised tags allow high-throughput sequencing machines to count the number of copies of each of these tags. From a statistical point of view, it is possible to account for nearly every mRNA as long as enough sequence information is accumulated. Because approximately 300,000 copies of messenger RNA are expressed per cell, in order to get a complete overview of these, and to get adequate quantification on their expression levels, one must accumulate perhaps 1,000,000 or more tags to get statistically valid information. This is because SAGE is a digital method, not an analogue method. The generation of sufficient sequence tag data is extremely expensive, especially for studies involving multiple RNA samples.

There are several methods that fall under the grouping of 'differential display' (Liang & Pardee, 1992; Prashar & Weissman, 1996; Welsh *et al.*, 1992). Those, other than TOGA, each incorporate some random step, either in the priming for cDNA synthesis, or in the enzymes that are used to generate products. In some cases, those random events lead to false-positive differences. In all cases, they create a bias towards abundantly-expressed genes and away from rarely expressed genes. Because these methods incorporate random steps, it is not possible to build databases easily out of the information. By contrast, with TOGA and its simple addressing mechanism, identification of the candidate identity for any RNA is facile and database generation and queries are straightforward.

4.3.8.5 Summary

The data generated by TOGA are highly reproducible, because the PCR-based method is carried out by a robot. One consequence of that is that the databases become a permanent record, and they can be called upon in later experiments electronically. Using the database, one can compare RNA concentrations in different samples that might have been generated in completely unrelated experiments. The detection level of present capillary electrophoresis systems is below one part in a million, which is between a tenth and three-tenths of a copy of RNA per cell. TOGA has a very low RNA requirement – 20 nanograms of total RNA. It is very high-throughput – 6,000 samples can be processed per robot per year with very short turnaround time. Therefore, TOGA is a descriptive tool reporting on relative concentrations of RNA, and its short turnaround time allows its data to be used heuristically to help design further experiments to refine the questions that one addresses. The method recognises both known and novel RNAs, and can be used to examine RNAs in any species (plant or animal) – it does not require that an RNA to be already in a database to be known. Because the data are automatically accumulated, they can be queried instantaneously, and they can be related to the literature. The addressing method is simple, and that simplicity allowed the method to be automated. That simplicity also has enabled very powerful informatics.

4.4

Sequencing-based techniques

4.4.1

MPSS™ — Massively parallel signature sequencing

4.4.1.1 Introduction

New developments in genome science are creating an increasing interest in systems biology research. Systems biology is best defined simply as the analysis of all the components of a complex biological system. It involves a new paradigm for conducting biological research where the organism or biological system is analysed in its entirety rather than by just studying a limited number of components one at a time. An important need for conducting systems biology experiments is to be able to accurately analyse the transcriptome of the system under analysis. Information on all genes that are expressed within the system need to be collected, and an accurate assessment of the number of copies of mRNA from each gene should be part of the analysis.

Several DNA microarray platforms (Duggan *et al.*, 1999; Evertsz *et al.*, 2000; Hughes *et al.* 2001; Lipshutz *et al.*, 1999; Lockhart & Winzeler, 2000; Zhou *et al.*, 2000), serial analysis of gene expression (SAGE) (Velculescu *et al.*, 1995; Zhang *et al.*, 1997), cDNA sequencing and a variety of other technologies reviewed in other sections of this book are available for analysing the expression of thousands of genes simultaneously. In this section, one of the newest tools for evaluating in-depth gene expression, called massively parallel signature sequencing (MPSS™) (Brenner *et al.*, 2000a; Tyagi, 2000), is described. MPSS™ is an open-ended platform that analyses the level of expression of virtually all genes expressed in a sample by counting the number of individual mRNA molecules produced from each gene. There is no requirement that genes be identified and characterised prior to conducting an experiment. MPSS™ has a routine sensitivity of a few molecules of mRNA per cell, and the data sets are in a digital format that simplifies the management and analysis of the data. MPSS™ results are particularly useful for generating the type of complete data sets that will help to facilitate the development of relational databases for systems biology research.

4.4.1.2 Principles and basics

4.4.1.2.1 MPSS™ produces data by counting individual mRNA molecules in a sample

The principle of MPSS™ as a gene expression tool is very simple. It works by simultaneously counting a large number of molecules of mRNA in a sample. Individual mRNAs are identified through the generation of a 17-base 'signature sequence' at a unique site on the molecule, and, in a typical MPSS™ data set, over a million molecules are counted simultaneously. Bioinformatics tools are used to sort out how many molecules of mRNA from each gene are present in the sample. In the end, the level of expression for each gene in an MPSS™ data

set is represented by the number of transcripts present in a million molecules counted (transcripts per million).

4.4.1.2.2 Process for generating signature sequences

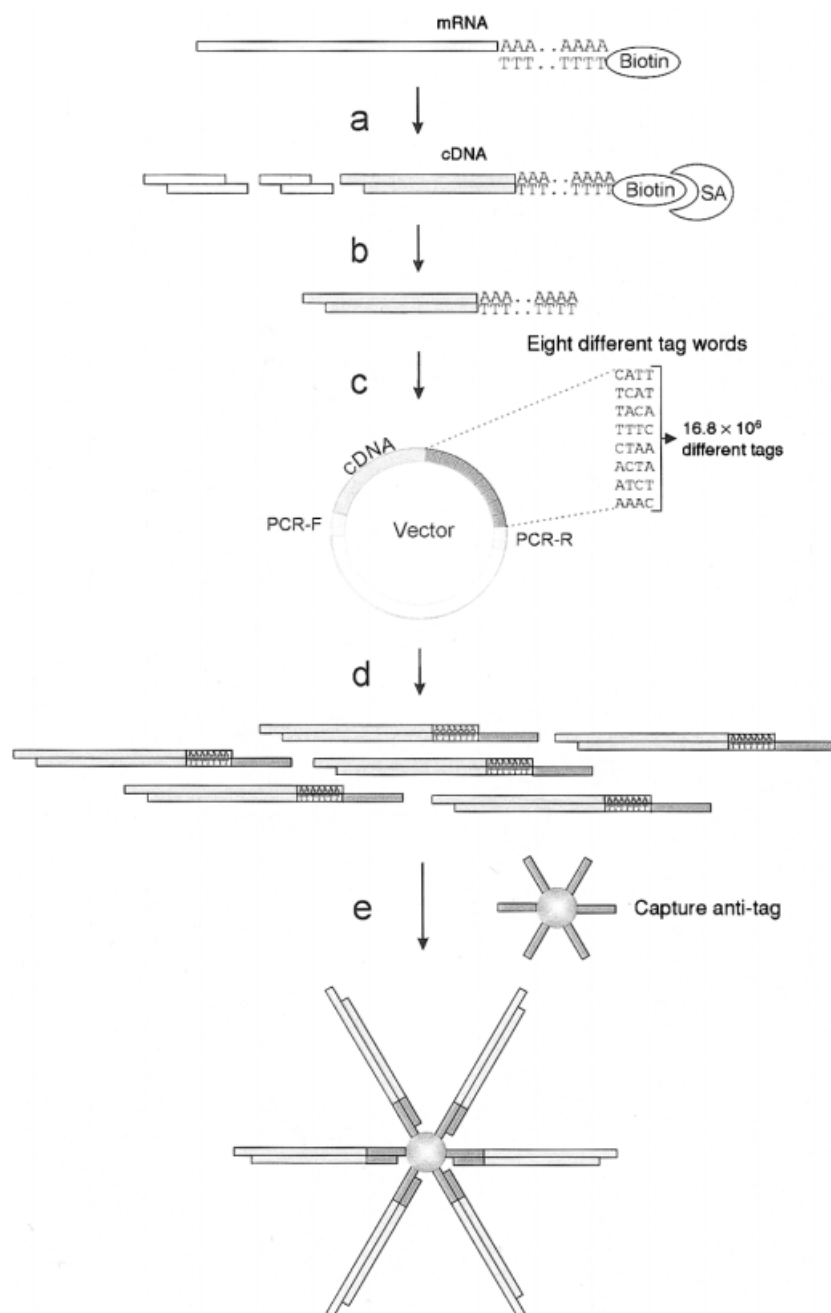
The key to MPSS™ is the ability to identify a large number of mRNA molecules by efficiently generating 17-base signature sequences on over a million molecules of mRNA at once. This is accomplished using two different technologies that have been developed at Lynx Therapeutics, Inc. (Hayward, California, USA). The first is called Megaclone™, which is the technology that allows millions of individual molecules of cDNA to be efficiently cloned onto microbeads (Brenner *et al.*, 2000b). The second technology, MPSS™, uses the cloned cDNA on the microbeads as a substrate to perform a million different 17-base sequencing reactions simultaneously (Brenner *et al.*, 2000a). Each technology will be described in detail below.

4.4.1.2.3 Megaclone™ technology to clone cDNA molecules on beads

To generate the substrate for the MPSS™ sequencing reactions, the Megaclone™ technology is used to clone double-stranded cDNA onto five micron microbeads (Brenner *et al.*, 2000b) (figure 4.70). For this purpose, cDNA is prepared from poly(A)⁺ mRNA using a biotin-labelled oligo(dT) primer. The oligo(dT) is designed to prime each mRNA molecule exactly at the poly(A) junction. The cDNA fragments are then digested with *Dpn* II (recognition sequence GATC), and the 3'-most *Dpn* II-poly(A) fragments are purified utilising the biotin label at the end of each molecule. Each cDNA fragment is then tagged with one of the 16.8 million different 32-base synthetic oligonucleotide tags developed at Lynx Therapeutics (see below for a description of the tags). After amplification of the tagged cDNA molecules, the 32-base tag at the 3'-end of each molecule is made single-stranded with T4 DNA polymerase, and then the tags on the cDNA fragments are hybridised to a set of 16.8 million different microbeads, each containing a covalently attached 32-base oligonucleotide (called an anti-tag) that is complementary to one of the 32-base oligonucleotides used to tag the molecules. Each microbead contains a vast excess of the 32-base anti-tag, so the amplified cDNA fragments with the same tag will hybridise and 'load' a single microbead in the 16.8 million targeted bead set. On average, 100,000 amplified copies of each cDNA will hybridise to a single bead. The nick between the hybridised tag/complementary tag on the microbead is then repaired to create a covalent linkage between the cDNA molecules and the microbead.

In the end, the Megaclone™ process yields a library of microbeads where each molecule of mRNA in the starting sample is represented by one microbead. Each microbead in the library contains approximately 100,000 identical cDNA fragments from one mRNA. The cDNA fragment on each microbead contains the sequence between the 3'-most *Dpn* II site and the poly(A) junction on each mRNA. The attachment of the cDNA fragment to the microbead occurs at the poly(A) end of the fragment.

The 32-base tag/complementary tag system has a number of unique properties that are critical for making the Megaclone™ microbead loading process work. Each



◀ **Figure 4.70 Summary of the Megacalone™ technology.** Poly(A)⁺ mRNA molecules are converted into double-stranded cDNA molecules and are then digested with *Dpn II* (step a). The 3'-most cDNA fragments are selected utilising a biotin tag (step b) and then cloned into a specially designed plasmid vector containing a 32 base pair oligonucleotide tag (step c). There are 16.8×10^6 different 32-base sequences available in the reference tag library, and each cDNA clone contains a different sequence (Brenner *et al.*, 2000a). A library of cDNA inserts, along with their adjacent 32-base oligonucleotide tags, are PCR amplified, and the resulting linear molecules are partially treated with an exonuclease to make the 32-base tag single-stranded (step d). The 32-base tags at the end of each cDNA are hybridised to 32-base complementary tags that are covalently linked to five micron microbeads (step e).

There are 16.8 million different complementary tags, each of which corresponds to one of the 16.8 million different 32-base oligonucleotides that is used to tag the molecules (Brenner *et al.*, 2000a); therefore, for every tag on a cDNA molecule, there will be one bead with a complementary 32-base tag available for hybridisation. Each bead contains a vast excess of one particular 32-base complementary tag sequence. Once the cDNA molecules are hybridised to the beads, the nicks are sealed enzymatically. The end product is a library of microbeads, each with approximately 100,000 identical cDNA molecules covalently attached to its surface. The following abbreviations are used in this figure: SA, streptavidine; PCR-F, complementary site of forward PCR primer; PCR-R, complementary site of reverse PCR primer.

32-base tag is composed of eight carefully designed 4-base 'words' (figure 4.71.A). The tag set is comprised of all combinations of the eight 4-base words ($8^8 = 16.8 \times 10^6$), which yields a set of 16.8 million different tags. Each word is designed to contain one G/C base pair and three A/T base pairs to ensure that all 16.8 million possible tag/anti-tag duplexes have about the same melting temperature (figure 4.71.B). The tags contain only cytosines and the anti-tags contain only guanines, which is a feature to prevent any of the tags or anti-tags from hybridising with themselves. Most importantly, any given tag is different from its 56 nearest neighbours by a total of three nucleotides (figure 4.71.A), which means that the hybridisation of any tag with one of its nearest neighbours will produce a duplex that has a melting temperature that is several degrees centigrade less than that produced by hybridisation with its perfect anti-tag (figure 4.71.B). Therefore, careful control of

Perfect match

CATT CATT TTTC ATCT ATCT ATCT AAAC CATT
GTAA GTAA AAAG TAGA TAGA TAGA TTTG GTAA

One-word mismatch

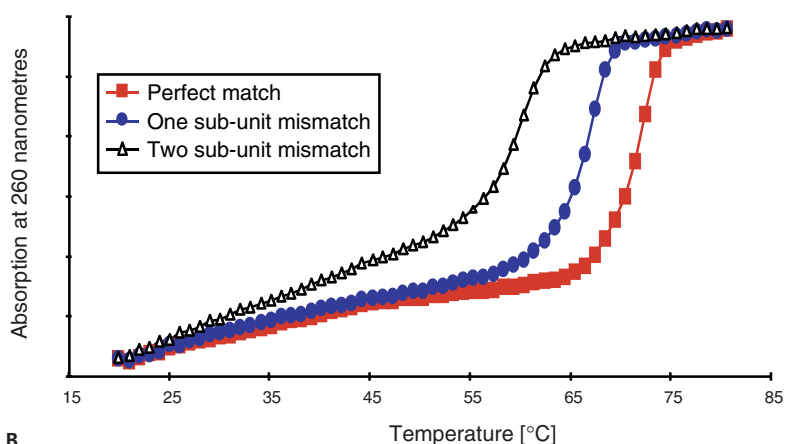
CATT CATT TTTC ATCT ATCT ATCT AAAC CATT
GTAA GTAA AAAG TTTG TAGA TAGA TTTG GTAA

Two-word mismatch

CATT CATT TTTC ATCT ATCT ATCT AAAC CATT
GTAA GTAA AAAG TTTG TAGA GTAA TTTG GTAA

A

Figure 4.71.A (s. next page)



B

Figure 4.71 Nature of the 32-base oligonucleotide tag/anti-tag set used for Megaclone™. A) The 16.8 million different tag/anti-tag set is produced from all combinations of the eight 4-base words shown in bold on the top line. An example of one 32-base tag with its perfect complementary anti-tag is shown at the top. An example of a 32-base tag with one of its 56 closest neighbours with one subunit mismatch

is shown in the middle. An example of a tag/anti-tag duplex with a two subunit mismatch is shown at the bottom. Underlined nucleotides represent mismatched bases in the duplex. Tags are in bold text; anti-tags are in regular text. B) Melting curves using tag/anti-tag duplexes comprised of a perfect match, a one subunit mismatch or a two subunit mismatch.

the temperature and buffer conditions during the bead loading hybridisation reaction will drive the formation of perfect tag/anti-tag duplexes and minimise the cross-hybridisation of any tag with its nearest neighbours.

4.4.1.2.4 Generating signature sequences with MPSS™

The microbeads loaded with cDNA fragments using Megaclone™ are used directly for the MPSS™ sequencing reactions (Brenner *et al.*, 2000a). Since each cDNA molecule is attached at its poly(A) end, the 17-base signature sequencing reactions occur from the *Dpn* II end of each molecule. Approximately one million microbeads are packed into a specially designed flow-cell in a way that allows them to form a tight monolayer by stacking along channels in the flow-cell (figure 4.72). The flow-cell is connected to a computer-controlled microfluidics network that delivers the reagents for the sequencing reactions. Each microbead is stationary in the flow-cell and is monitored with a charge-coupled device (CCD) camera that is capable of detecting a fluorescent signal on each of the microbeads. Fluorescent signals arising from each microbead are monitored during the sequencing reactions (see below) in order to determine the sequence from the end of each cDNA molecule attached to the microbeads.

To initiate the sequencing reaction, a 4-base single-stranded overhang is produced at the end of each cDNA molecule. This is done by first digesting the DNA on the microbeads with *Dpn* II, and then ligating an initiating adapter to

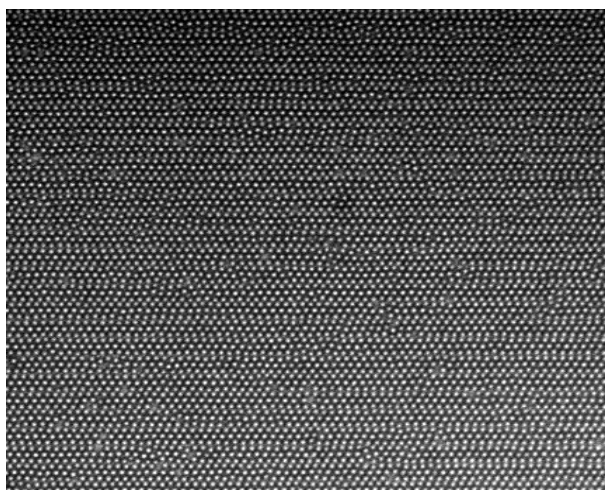


Figure 4.72 Picture of the microbeads loaded into the flow-cell through the charge-coupled device (CCD) camera. The camera on the MPSS™ device has the capacity to generate images of the beads with visible light and after fluorescent activation.

the resulting GATC *Dpn* II sticky-end. The initiating adapter contains a *Bbv* I type IIs restriction recognition site. Digestion of the sample with *Bbv* I produces a 4-base single-stranded overhang at a position nine to thirteen nucleotides away from the recognition sequence (figure 4.73). This generates initiated molecules with a 4-base single-stranded overhang immediately adjacent to the GATC *Dpn* II recognition site. In practice, two different initiating adapters, called two-stepper and four-stepper, are used for initiating the sequencing reaction in order to stagger the first four bases in the single-stranded overhang. This is done to address the issue of palindromes that arise in the 4-base overhangs during the sequencing reactions. The issue of palindromes was addressed in Brenner *et al.* (2000a) and will not be described in detail here.

Once the molecules are initiated, the sequencing reactions are performed directly in the flow-cell. In order to know how the sequencing reactions work, it is first important to understand the concept of the process. The system was designed to generate sequence in four-nucleotide increments through the hybridisation of a synthetic adapter to the 4-base single-stranded overhang on each initiated cDNA. Each adapter is part of a set of molecules that each contain a defined 4-base single-stranded overhang with one of all 256 different possible combinations of four nucleotides ($4^4 = 256$). Each molecule is also encoded in a manner that allows the sequence of its 4-base overhang to be decoded through a series of hybridisation reactions in the flow cell. Therefore, the sequence of the first four nucleotides on the cDNA is generated by simply deciphering which member of the adapter set has hybridised to the cDNA on each microbead. Each synthetic adapter also contains a type IIs restriction recognition site, so additional sequence beyond the first four

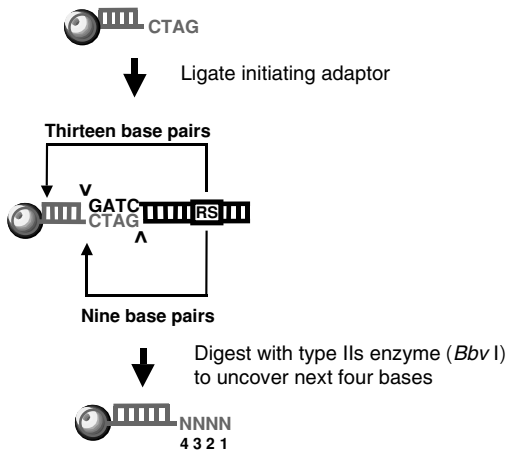


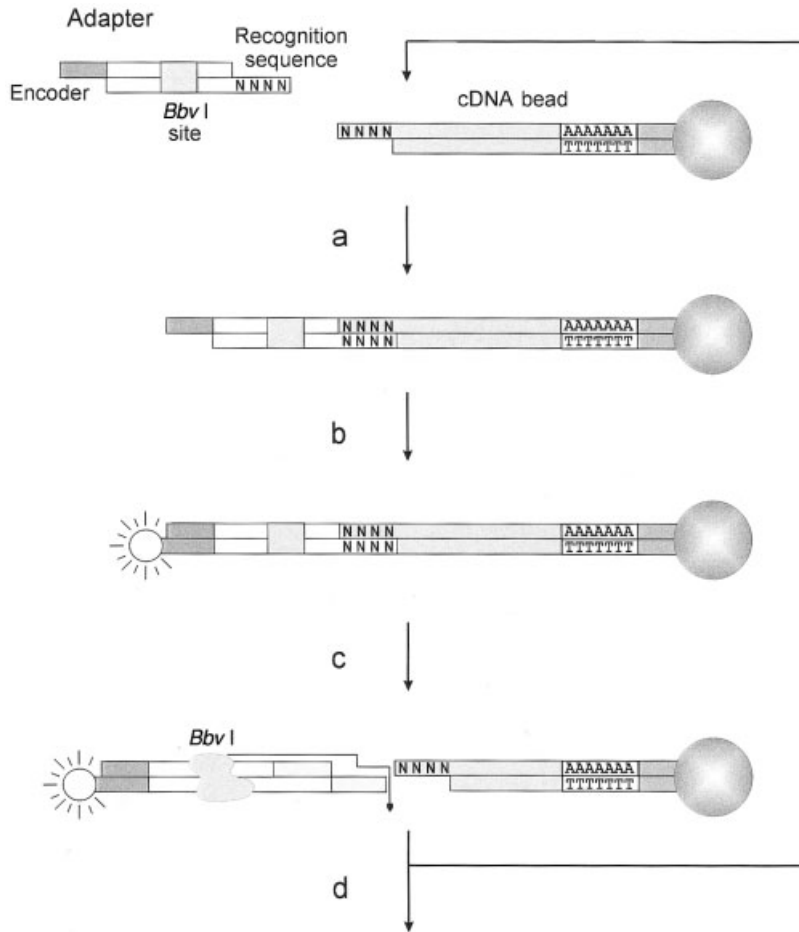
Figure 4.73 Process for generating initiated cDNA molecules for the sequencing reactions.

One microbead (of the million in the flow-cell) with one double-stranded molecule of DNA with a GATC single-stranded overhang (of the approximately 100,000 cloned molecules on each bead) is shown at the top. The length of the molecule is not drawn to scale. Ligation of a synthetic initiation adaptor occurs through the GATC overhang on the cDNA molecule. The RS site in the adaptor is a recognition sequence for

the type IIs restriction enzyme, *Bbv* I. Digestion with *Bbv* I reaches nine to thirteen nucleotides away from the recognition site to a position outside of the adaptor to cut the DNA; this produces an initiated cDNA with a 4-base single-stranded overhang. The nucleotides numbered as 1, 2, 3 and 4 correspond to the first four nucleotides adjacent to the *Dpn* II GATC restriction site that will participate in the first round of sequencing.

nucleotides is produced by repeating the process after removing the first synthetic adaptor with a type IIs restriction enzyme. The recognition site for the enzyme is positioned in the adaptor in a way that produces a new single-stranded overhang that is immediately adjacent to that from the previous round of analysis.

A summary of the actual details of the sequencing reactions is provided in figure 4.74. The first step of the process involves the addition of a set of encoded adapters to the flow-cell. The members of the encoded adapter set each contain two important features: a recognition sequence, which is a defined 4-base single-stranded sequence at one end of each molecule, and the encoder sequence, which is a single-stranded sequence at the other end of the molecule (see below for details). The 4-base recognition sequence on an encoded adapter will hybridise to a complementary sequence on the single-stranded end of the initiated cDNA on the microbeads, and ligase is then added to covalently link the encoded adapter to the ends of the molecules. The next step involves determining which encoded adapter, with a defined recognition sequence, has become attached to each cDNA molecule. This is accomplished through a series of hybridisation reactions with 16 different fluorescent-labelled decoder probes, where each probe hybridises only to the encoder sequence on specific members of the encoded adapter set (see below). The recognition sequence for the encoded adapters on each microbead is then deciphered



Determine the 17mer-signature of each cDNA bead

Figure 4.74 Summary of the sequencing reactions for MPSS™. Encoder adapters are ligated to the ends of the cDNA molecules attached to the microbeads (step a). Sixteen different fluorescently labelled decoder probes are then sequentially hybridised to the encoded adapter ends in order to deduce the first four nucleotides

at the end of each molecule (step b). The encoded adapter from the first round is then removed by digestion with *Bbv* I, which exposes the next four nucleotides as a 4-base single-stranded overhang (step c). The process is repeated several times in order to generate a total of 17 bases of sequence (step d).

using the encoded sequence as a guide (see below). To collect additional sequence, the encoded adapter from the first round is removed by digestion with *Bbv* I to produce a 4-base single-stranded overhang that is immediately adjacent to the first four nucleotides from the first round of sequencing. This is possible because the encoded adapters contain a synthetic *Bbv* I site at a defined position within the molecule. The process is then repeated several times in order to generate a 17-base signature sequence for the cDNA on microbead in the flow-cell.

The encoded adapter set of 1,024 different molecules is comprised of four groups, each with 256 different members (Brenner *et al.*, 2000a) (figure 4.75). Each group contains all possible combinations of nucleotides in the recognition sequence ($4^4 = 256$), so there will be an encoded adapter for any possible 4-base overhang on the initiated cDNA molecules on the microbeads. There are 16 different encoder sequences that are used to build the encoded adapter set, and each is used to identify a specific nucleotide at one of the positions on the 4-base recognition sequence. For example, in the first group, all encoded adapters with a cytosine at the first position (and all combinations of nucleotides at the second through fourth position) are labelled with the first encoder sequence, and those with thymine, adenine or guanine at the first position are labelled with the second, third and fourth encoder sequence, respectively. Groups 2, 3 and 4 are labelled in a si-

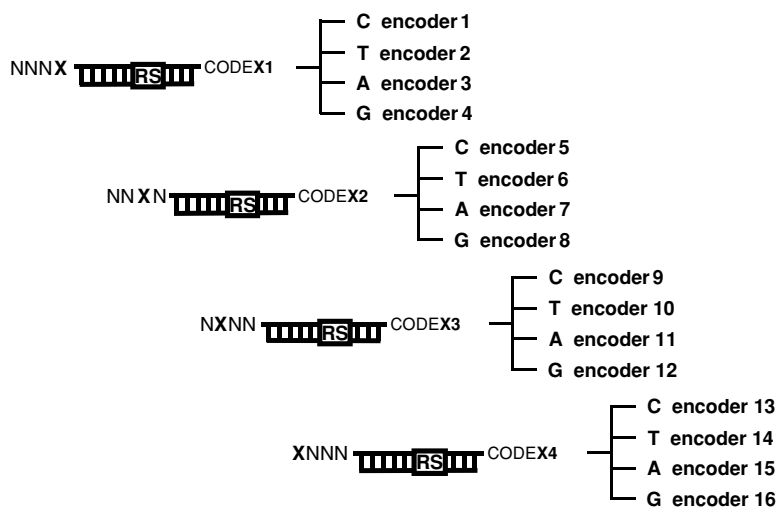


Figure 4.75 Composition of the encoded adapter set of molecules. There are four different groups of adapters that contain 256 individual members, each with one of all possible combinations of nucleotides on the single-stranded recognition sequence (shown with 3 N's and one X). There are 16 different encoder sequences, four of which are used for each group of adapters. In the first group, shown at the top, encoder sequence number 1 is used on

molecules with a cytosine at the X position. The second through fourth encoder sequences are used when the X position is a thymine, adenine, or guanine, respectively. Similarly, the remaining twelve encoder sequences are used on the second, third and fourth groups in order to label the nucleotide at the X position on each molecule. The following abbreviation is used in this figure: RS, recognition site of the restriction enzyme.

milar manner with the remaining twelve encoder sequences in order to label the specific nucleotide at positions 2 through 4. This approach makes it possible to decipher 256 different sequence combinations on the 4-base overhangs on the cDNA molecules using only 16 different fluorescent-labelled decoder probe hybridisation reactions. It is important to note that four different encoded adapters hybridise with the cDNA on each microbead. These different members of the encoded adapter set contain the same recognition sequence, but each has a different decoder sequence that allows the nucleotide at each position to be deciphered.

4.4.1.2.5 Data handling and calculation of RNA abundance

A typical flow-cell loaded with about a million microbeads will yield high quality 17-base signature sequences from about 250,000 to 400,000 microbeads. Since MPSS™ data sets are additive in nature, data sets from multiple flow-cells with microbeads loaded from the same starting mRNA sample can be combined. This routinely produces experimental data sets that contain signature sequences from more than one million counted molecules per sample (see below). Assuming that an average cell in a biological sample contains approximately 300,000 individual mRNA molecules (Velculescu *et al.*, 1995), a sampling of one million mRNA molecules provides greater than a three-fold over-sampling per cell, which provides potentially the capability to detect a single mRNA per cell. Additional sensitivity can be achieved, if necessary, by sequencing more beads loaded from the same mRNA sample. This level of sensitivity becomes especially important when biologically heterogeneous samples are being analysed where only a small percentage of cells in a sample may be expressing a gene of interest.

Each signature sequence in an MPSS™ data set is identified, compared electronically to all other signature sequences, and all identical signatures are counted. In most cases, signatures from about one million counted mRNAs, collected from several flow-cells, are identified and counted, so the data for each gene in the data set is routinely normalised and expressed as transcripts per million. Analysis of a complete MPSS™ data set makes it possible to calculate the numbers of genes that are expressed at various levels within the sample. For example, it is possible to readily calculate the genes that are expressed at greater than 1,000 transcripts per million, between 100 and 1,000 transcripts per million, between ten and one hundred transcripts per million, and less than ten transcripts per million. This type of analysis for many samples has revealed that in the extensive collection of human tissues and cells studied to date at Lynx Therapeutics, a large percentage of genes are expressed at a level of between ten to one hundred transcripts per million (data not shown).

4.4.1.2.6 Associating signature sequences with genes

The signature sequence generated with MPSS™ needs to be associated with a specific gene in order to understand the biological significance of the data. Signature sequences can be connected to known genes by comparison with data in the available genomic and expressed sequence tags (ESTs) databases for the organism under study. This is usually an efficient process, although occasionally it is not pos-

sible to find a signature for a gene known to be expressed in a particular sample. This happens when either a gene does not contain a *Dpn* II site, or when there is a sequence polymorphism in the *Dpn* II site. These problems can be easily overcome by digesting the cDNA with a second enzyme with a different recognition sequence. Additionally, incomplete sequence representation of a particular gene in the current EST and cDNA clone databases can complicate the process of assigning a signature sequence to a gene. Not all cDNA clones that have been sequenced are complete to the poly(A) addition site, so the sequence that corresponds to an MPSS™ signature for a specific gene may not be represented in an EST database. In this event, and in the case where there is little genomic information available for the organism under analysis, other procedures would need to be used to identify the gene associated with a specific signature sequence.

4.4.1.3 MPSS™ data sets are digital in nature and cover virtually the entire transcriptome

It is highly desirable to have the complete transcriptome of a sample available for experiments in systems biology. The inherent sensitivity to detect low-level expressed transcripts coupled with the fact that even currently unknown genes are analysed with MPSS™ ensures that virtually all genes in a sample are represented in an MPSS™ data set. Additionally, and most important, MPSS™ data is digital in nature, where the level of expression for each gene is expressed in transcripts per million. There is no need to first convert data into a digital format, which is often problematic, particularly with some technologies where the data is captured in an analogue format (Aach *et al.*, 2000). MPSS™ data sets can be seamlessly integrated into relational databases for complex data mining operations. Therefore, MPSS™ has many of the performance characteristics that are compatible with the requirements for generating gene expression data content for experiments in systems biology.

The digital nature of MPSS™ data also facilitates the sharing of data between research groups. The fact that each gene with MPSS™ is expressed in terms of transcripts per million means that a data set generated in one laboratory can be readily shared with other investigators also using MPSS™. With other technologies, there are many experimental variables that can significantly impact the nature of the data set produced in different laboratories. For example, with microarrays, variability in the manufacture of the microarray along with differences in the procedures for probe preparation, hybridisation and washing of the microarray can significantly impact the detection sensitivity and background for any given experiment. These factors can substantially complicate the ability to compare one data set with another. On the other hand, MPSS™ data sets can be shared and compared directly without having to address sensitivity and background issues. This allows one data set to be compared with all others.

Additionally, an important attribute of MPSS™ data over those produced by other technologies, like microarrays, relates to the fact that MPSS™ data can be treated as 'categorical' from a statistical point of view (Reinartz *et al.*, 2002; Kal *et al.*, 1999). This makes it possible to capitalise on the large number of measurements of a

given signature in the data set (typically ten to one thousand or more depending upon the gene) as well as the size of the entire data set (typically over a million) to calculate whether the gene giving rise to this signature is differentially expressed in several different samples. Therefore, for example, for a statistic significance < 0.001 , it is possible to detect a two-fold change for a gene that is expressed at a level of 30 to 40 transcripts per million. However, for genes that are expressed at a higher abundance, it is possible to detect a substantially smaller difference. A 40 percent difference can be ascertained for genes that are expressed at about 200 transcripts per million. These characteristics are in contrast to the analyses for gene expression data generated by hybridisation based methods, such as microarrays, where a significance test is possible only if the experiment is replicated several times (Kerr *et al.*, 2001), and where differential expression can usually be detected only for genes with relatively high levels of expression (Duggan *et al.*, 1999) and with a large difference between samples (Everts *et al.*, 2000).

4.4.1.4 Comparison of MPSS™ to cDNA sequencing, SAGE and to microarray technologies

Several other technologies are available for conducting in-depth gene expression experiments. Here, MPSS™ will be compared with cDNA sequencing, serial analysis of gene expression (SAGE) and the DNA microarray technologies. These comparisons capture most of the advantages/disadvantages of MPSS™ over the other large-scale gene expression profiling technologies that are currently in use.

Direct sequencing of cDNAs was the first digital technology for measuring gene expression. Both MPSS™ and direct cDNA sequencing involve the generation of a cDNA library as the first step of analysis. Once the cDNA library is made, however, sequencing of a sample of cDNA clones from each library involves the purification and sequencing of DNA using standard procedures that are both costly and time consuming. With Megacclone™, at least one million cDNA molecules are cloned onto beads, and with MPSS™ over a million clones are sequenced simultaneously. The time, effort and cost to generate data from a million mRNAs in a sample with MPSS™ is a small fraction of that required to sequence the same number of clones using conventional technologies.

SAGE is also a transcript counting technique that generates a tag sequence for each mRNA (Velculescu *et al.*, 1995; Zhang *et al.*, 1997). However, MPSS™ has two noteworthy advantages over SAGE. The first relates to the length of the signature produced. With current methods, the SAGE tag for each mRNA is 14 nucleotides, while MPSS™ produces a 17-nucleotide signature. The longer signature with MPSS™ makes it much less ambiguous to assign the signature to a specific gene on the genome. Additionally, and most important, a typical SAGE tag set is comprised of only 20,000 to 60,000 sequenced mRNAs, which is unlikely to have the depth to adequately represent all genes in the sample, especially those expressed at low copy levels. The automated nature of MPSS™ makes it possible to efficiently produce a very large data set of signature sequences with substantially greater depth than SAGE.

With respect to the various DNA microarray platforms that are currently in use, MPSS™ is most notable in that it is a technology that has the potential to capture virtually all genes present within the sample, not just those that have been placed on the microarray. No prior knowledge of a gene's sequence is required for MPSS™. While this is most relevant to non-human species whose genomes have not been sequenced, it also applies to genes on the human genome that have not yet been identified and annotated. Also, with some array platforms, the data is expressed as a ratio of expression of a gene in one sample compared to another. Converting these ratios into a number that accurately reflects the expression of a gene in a particular sample is potentially problematic (Aach *et al.*, 2000; Zhou *et al.*, 2000). Microarrays also have the limitation that homologous genes can cross-hybridise, which makes it impossible to detect the expression of individual members of highly homologous gene family members (Evertsz *et al.*, 2001). With MPSS™, the signature sequence, which is often in the 3'-untranslated region, is usually different for individual genes within a family. Therefore, it is possible, in many cases, to differentiate highly homologous genes from each other.

While microarrays have limitations for in-depth gene expression analyses, they have the advantage of being very useful for the high-throughput analysis of multiple samples. Therefore, it may be useful to think of the microarray and MPSS™ technologies as being complementary in nature – different tools for different types of experiments. For example, to generate in-depth and quantitative gene expression data for building complex relational databases for systems biology applications, MPSS™ may be the technology of choice. After these databases are mined for interesting biological information, it may be necessary to test whether sets of genes are differentially expressed in a large number of samples (*e.g.*, tumours of a specific type). Here the microarray platform would clearly be the technology of choice. Having access to both MPSS™ and at least one of the microarray technologies would seem to be most ideal for many investigators.

4.4.1.5 Summary

Compared to several existing technologies, MPSS™ has the advantage that it provides in-depth quantification of virtually all genes that are expressed in a sample. Since there is no requirement for prior knowledge of any gene or genome, it is possible to generate quantitative gene expression data sets from any organism. Additionally, since an MPSS™ data set typically involves over a million signature sequences, it has the sensitivity to accurately quantify genes that are expressed at very low levels within a cell. The fact that the data is digital in nature allows MPSS™ data sets to be seamlessly shared and integrated into large relational databases. These features make MPSS™ an ideal technology for generating data content for experiments in systems biology.

4.4.2

SAGE – Serial analysis of gene expression**4.4.2.1 Introduction**

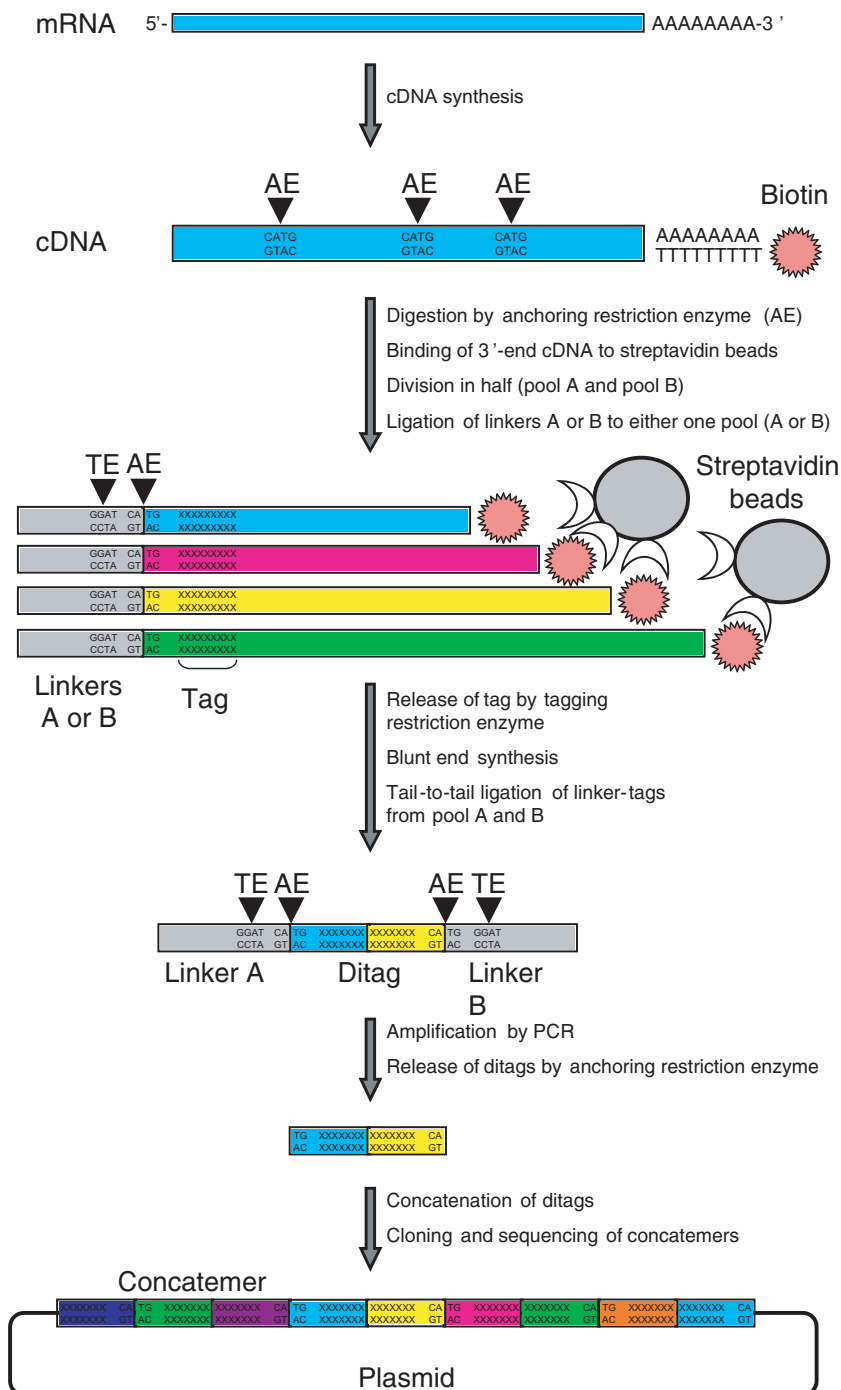
Serial analysis of gene expression (SAGE) is a high-throughput experimental technique, designed by Victor E. Velculescu and co-workers, to gain a detailed qualitative and quantitative measure of gene expression without having prior sequence information (Velculescu *et al.*, 1995). It can be used not only to obtain a global gene expression spectrum of a particular type of cell or tissue, but also to identify genes that are differentially expressed in cells or tissues that are exposed to different conditions, derived from various diseased states or from wild-type and transgenic animals. SAGE can be applied on all eukaryotic cells. Up to now, it has been used in men and mice, in the malaria parasite *Plasmodium falciparum* (Munasinghe *et al.*, 2001), in yeast (Kal *et al.*, 1999; Velculescu *et al.*, 1995) and even in rice (Matsumura *et al.*, 1999).

The original SAGE technique is based on the isolation of unique DNA sequence tags, derived from a defined position in individual mRNAs, which are concatenated serially into long DNA molecules for sequencing (Velculescu *et al.*, 1995). SAGE may be used to study general housekeeping mRNA profiles, signalling pathways, and cell or tissue-type specific expression profiles under physiological and pathological conditions. SAGE is useful for the study of rare and abundant transcripts and for the discovery of novel genes. For all these studies, SAGE relies on DNA sequencing capacity, the knowledge of the genomic sequence of the species of interest and availability of SAGE databases to identify genes and gene expression profiles. The number of fully sequenced genomes is steadily increasing, and this increase is being accompanied by an increase in the number of scientific papers using SAGE technology. As a result, millions of SAGE tags have now been deposited in public SAGE databases (Velculescu *et al.*, 2000). This wide availability of SAGE libraries has enabled comparisons of universal gene expression profiles among different tissues, leading to transcriptome mapping (Caron *et al.*, 2001). The analyses reveal both gene responses that are shared among tissues and others that are tissue-specific.

The field where SAGE has been most comprehensively applied is undoubtedly cancer research. The data from these experiments have been collated in several specialised cancer databases such as the Cancer Genome Anatomy Project (CGAP) at the National Cancer Institute (<http://cgap.nci.nih.gov/>) (Strausberg, 2001). Many other SAGE databases exist, a commonly used starting-point is the SAGE database at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/SAGE>).

4.4.2.2 Principles and basics

A stepwise explanation of the original SAGE technique (Velculescu *et al.*, 1995) is illustrated in figure 4.76. Following isolation of mRNA from the cells or tissue of interest, double-stranded cDNA is synthesised using biotinylated oligo(dT) primers. Digestion of the cDNA by the so-called ‘anchoring restriction enzyme’ *Nla*



- ◀ **Figure 4.76 Schematic representation of the SAGE technique.** RNA is obtained from the cells or tissue of interest and is transcribed to cDNA. Upon cleavage of the cDNA with an anchoring restriction enzyme (AE) the 3'-ends are collected and linkers are attached. The tagging restriction enzyme (TE) recognises its cleavage signal in the linker and cleaves downstream, resulting in the release of a small cDNA tag. These linker tags are ligated into linker ditags, which are amplified by PCR. Then the ditags are released by incubation with the anchoring restriction enzyme and ligated into concatemers. The concatemers are cloned into a plasmid and sequenced to obtain information on gene expression.

III results in cDNA fragments, the 3'-biotinylated cDNA ends of which will bind to streptavidin beads. The remaining cDNA fragments are discarded. Since *Nla* III has a 4-base recognition site, it will cleave on average at every 256 (4^4) base pairs. Since nearly all mRNA transcripts are longer than 256 nucleotides, the enzyme will digest virtually all transcripts. The 3'-end cDNA fragments are resulting from digestion divided into two portions which are ligated to two different linkers (A and B) containing the recognition site for the so-called 'tagging restriction enzyme' *Bsm* FI. Short fragments – the so-called tags – are enzymatically released by digestion with *Bsm* FI, which cleaves on average 13 to 14 base pairs downstream of its recognition site. Blunt ends are created and the released linker tags from pools A and B are ligated tail-to-tail to form a ditag still containing the linkers. The linker ditags are amplified by PCR and again cleaved by *Nla* III to release the actual ditag. These ditags are separated by polyacrylamide gel electrophoresis (PAGE) and isolated from the gels. They have overhanging CATG sequences at both sides and are allowed to ligate into large cDNA concatemers. The obtained concatemers vary in size. Therefore, to enhance DNA sequencing efficiency, only the larger cDNA molecules are isolated and cloned into a plasmid vector. By comparing the concatemer sequences with SAGE databases, qualitative and quantitative information about the obtained SAGE tags can be obtained. Careful statistical analysis is required to establish significant differences between expression levels of a certain gene, present in different libraries. Commonly, a χ^2 -test is used (Man *et al.*, 2000). To obtain more sequence information about unidentified SAGE tags, the 13 to 14 base pair tag sequence can be used as a probe to isolate the original gene from a cDNA library (Velculescu *et al.*, 1995).

As explained above, the SAGE technique is based on the isolation of unique sequence tags, derived from a defined position in individual mRNAs. The position is defined by the anchoring restriction enzyme used to digest the cDNA. In the case of *Nla* III, the tag is derived from the nine to ten base pairs following the final CATG sequence preceding the poly(A) tail. Since the position of the tag is defined, the length of the tag is determined by the cleavage specificity of tagging restriction enzyme, a type II restriction enzyme. In the case of *Bsm* FI, cleavage occurs 13 to 14 base pairs (CATG plus additional nine to ten base pairs) downstream of its recognition site. A nine base pairs tag is able to discriminate between 262,144 transcripts (4^9). We now know that the human genome very probably contains between 30,000 and 40,000 genes (Lander *et al.*, 2001; Venter *et al.*, 2001). Therefore, theoretically each SAGE tag is unique to a single gene transcript.

4.4.2.3 Typical results

To illustrate SAGE in practice, an experiment to identify genes that are expressed as a result of cell activation will be discussed. In this experiment, mRNA was isolated from activated and resting cultured human vascular smooth muscle cells. In this study, approximately 47,000 SAGE tags were sequenced for each condition. Comparison of the two SAGE libraries revealed 105 SAGE tags that are at least five-fold induced upon activation of the smooth muscle cells. Table 4.9 contains the top 20 list of these up-regulated tags, including secreted cell activation markers, such as cytokines (tags number 1, 9 and 13), chemokines (tags number 11 and 19) and growth factors (tags number 4, 7 and 18). Often more than one tag from the same gene are found, as is the case in the present list for cardiac ankyrin repeat protein, activin β A and plasminogen activator inhibitor 2. A second tag is often der-

Table 4.9 The 20 most induced tags in the library of stimulated smooth muscle cells compared to resting smooth muscle cells.

No.	Tag	R	A	P-value	Corresponding gene	UniGene cluster
1	5'-TGGAAGCACT-3'	0	127	< 0.001	Interleukin 8	Hs.624
2	5'-ATGGTACTGA-3'	0	44	< 0.001	Cardiac ankyrin repeat protein	Hs.74019
3	5'-CAGAATTGAA-3'	0	37	< 0.001	Cardiac ankyrin repeat protein	Hs.74019
4	5'-AGTAGTATGA-3'	0	30	< 0.001	Basic fibroblast growth factor	Hs.56066
5	5'-CCCTTCTATA-3'	0	26	< 0.001	Plasminogen activator inhibitor-2	Hs.75716
6	5'-GCCTTGGGTG-3'	1	23	< 0.001	Leukemia inhibitory factor	Hs.2250
7	5'-AGTTTGAAAG-3'	1	22	< 0.001	Activin β A	Hs.727
8	5'-CCAAAATTCT-3'	1	17	< 0.001	Expressed sequence tag	Hs.193700
9	5'-ACACATTTTA-3'	0	17	< 0.001	Interleukin 11	Hs.1721
10	5'-CCTATATTCC-3'	1	15	< 0.001	Expressed sequence tag	Hs.301228
11	5'-GTACTAGTGT-3'	3	45	< 0.001	Monocyte chemotactic protein 1	Hs.340
12	5'-CCCTTCTGTA-3'	0	13	< 0.001	Plasminogen activator inhibitor 2	Hs.75716
13	5'-GGCACCTCAG-3'	3	36	< 0.001	Interleukin 6	Hs.93913
14	5'-TTATTATGA-3'	1	12	0.001	Tissue inhibitor of metalloproteinase 3	Hs.245188
15	5'-CTTGACATAC-3'	0	12	< 0.001	Dual specificity phosphatase 1	Hs.171695
16	5'-TATGTGCTTT-3'	0	11	< 0.001	No Match	—
17	5'-TAATATTTTT-3'	1	11	0.003	Actinin α 4	Hs.182485
18	5'-TCTTCCCTGT-3'	1	10	0.005	Activin β A	Hs.727
19	5'-TTGAAACTTT-3'	1	10	0.005	GRO- α	Hs.789
20	5'-CAGTCAATAT-3'	2	19	< 0.001	Expressed sequence tag	Hs.28792

R is the number of tags from resting smooth muscle cells;

A is the number of tags from activated smooth muscle cells.

ived from cDNA primed on internal poly(A) stretches within the mRNA or from alternative polyadenylation, as is the case for cardiac ankyrin repeat protein and activin β A. The second tag for plasminogen activator inhibitor 2 is derived from a gene polymorphism. Moreover, even among the highly expressed tags, an occasional tag does not match with any gene as is observed for tag number 16. Some tags are derived from expressed sequence tags (ESTs) in the database (tags number 8, 10 and 20) and require further analysis.

4.4.2.4 Examples

The SAGE literature is extremely diverse since SAGE is a technique that can be applied in numerous fields of research. Therefore, it is impossible to acknowledge all scientists who performed innovative work using SAGE. However, a few key manuscripts should be mentioned. Brad St. Croix and co-workers from the laboratories of Kenneth W. Kinzler and Bert Vogelstein, who developed the SAGE technique, performed a comprehensive SAGE analysis of the mRNA profile of normal human endothelial cells and tumour endothelial cells, derived from a colon carcinoma (St. Croix *et al.*, 2000). Briefly, the conclusions from this study were that the profiles of normal and tumorigenic endothelial cells are very similar, with only a small number of differential transcripts are detected. These differentially expressed genes are not specific for endothelial cells from colon carcinomas, but are expressed by endothelial cells in a variety of tumours. Significantly, this restricted group of genes is also expressed by other 'activated' (angiogenic) endothelial cells, such as those involved in wound healing and *corpus luteum* formation. These observations indicate that differential gene expression in tumorigenesis is not unique but resembles that of other cellular repair processes. Furthermore, Victor E. Velculescu and co-workers analysed for the first time the entire yeast transcriptome, showing that the majority of regulatory genes are present at a low copy number (Velculescu *et al.*, 1997). Huib Caron and co-workers introduced human transcriptome mapping (Caron *et al.*, 2001). Chromosomal mapping of all SAGE tags derived from all publicly available SAGE libraries revealed clustering of highly expressed genes in chromosomal domains, so-called regions of increased gene expression (RIDGES). Remarkably, on each chromosome, mRNA expression levels are mostly related to gene clustering. These important discoveries have paved the scientific path for a new view of gene expression.

4.4.2.5 Discussion

Although SAGE is a very effective approach for determining the expression of mRNA populations, there are significant biases in the observed results that are inherent in the experimental process and the computer sampling. Over the last few years, many of these biases have been addressed. Large amounts of mRNA and high-quality tag and ditag fragments are required to construct a good SAGE library. Four different modifications lead to sufficient SAGE tags with significantly reduced input of RNA, which is useful if only small (surgical) samples of tissue

are available (Datson *et al.*, 1999; Neilson *et al.*, 2000; Peters *et al.*, 1999; Virlon *et al.*, 1999). Three of these methods, called microSAGE, PCR-SAGE and SAGE-Lite will be discussed elsewhere in this volume. In addition, to obtain better quality and quantity linker tags and ditags, extra purification steps can be performed (Angelastro *et al.*, 2000; Lee *et al.*, 2001).

The efficiency of the original SAGE protocol is also limited by the small average size of cloned concatemers. Contaminating unbound linkers disturb efficient concatenation. For this reason, biotinylated PCR primers were introduced. Unbound biotinylated linkers generated in this fashion can be removed by binding to streptavidin-coated magnetic beads (Powell, 1998). To overcome the problem of contaminating aggregates of small concatemers, a heating step was added before polyacrylamide gel electrophoresis (Kenzelmann *et al.*, 1999). Both modifications yield substantially enhanced average length of cloned concatemers and, thus, more efficient sequencing, reducing time and costs.

A major problem of the SAGE technique is how to further analyse unidentified tags that represent uncharacterised transcripts. A variant of SAGE based on RT-PCR that uses a primer containing the SAGE tag as the sense primer, and a single base anchored oligo(dT) primer as an antisense primer, can provide new sequence information (Chen *et al.*, 2000; Matsumura *et al.*, 1999; van den Berg & van der Leij, 1999). The RT-PCR product can be used as a template either for Northern blotting or for *in situ* hybridisation experiments.

An algorithm has been developed to assist in deciding which unidentified SAGE tags are most worth pursuing in terms of tissue-specificity (Moreno *et al.*, 2001).

There have also been improvements in analysing and managing of SAGE data with new software tools, such as eSAGE, USAGE and POWER_SAGE (Man *et al.*, 2000; Margulies *et al.*, 2000; van Kempen *et al.*, 2000).

Even though quite a number of problems have been solved over the recent years, there are still certain drawbacks when using SAGE. For instance sequencing errors may occur leading to an invalid tag-to-gene assignment. In addition, ten-base pairs tags are not always unique. Multiple tags can be, for example, obtained from one transcript due to the occurrence of alternative polyadenylation sites. Furthermore, polymorphisms in the population may result in a variety of tags for the same gene. Also, some tags are shared by a family of genes, and are therefore unable to discriminate between individual family members. In addition, not all SAGE databases are of equal quality in providing tag-to-gene mapping, the 3'-untranslated regions of some genes or ESTs are unknown and the orientation of an EST may also be unclear. In such cases, the identity of the gene of interest must be verified by conventional techniques.

A bias in GC content has been identified in publicly available SAGE libraries (Margulies *et al.*, 2001). This was found to be due to spontaneous denaturation of AT-rich ditags at higher temperatures. It is thought that this bias affects about one fifth of current SAGE libraries.

Technical problems in the SAGE protocol include the fact that the tagging restriction enzyme *Bsm* FI can cleave on average 13 to 14 base pairs downstream of its recognition site, but depending on incubation conditions may cleave at be-

tween twelve to sixteen base pairs. As a result, upon ligation of two tags of different length, one can not be certain of the length of either tag, resulting in invalid tag identification. Moreover, *Bsm* FI cleaves on average once every 500 base pairs, thus destroying the 3'-end cDNA fragments that contain an additional *Bsm* FI recognition site, resulting in a lack of tags for these mRNAs. Also, the use of two different linkers may create difficulties, since half of the ligated linker-ditags contain the same linkers A or B, instead of one linker of each kind. These linker-ditags disturb efficient amplification by PCR. To tackle these problems other anchoring restriction enzymes, tagging restriction enzymes and linkers can be used, as in a large study which generated 80,000 tags from mRNA of the cervix adenocarcinoma cell line HeLa (Yamamoto *et al.*, 2001). Recently, more groups have started to employ different anchoring and tagging restriction enzymes. However, since most groups have used the original enzymes *Nla* III and *Bsm* FI, SAGE databases and related profiling algorithms generally depend on the original SAGE tags, and can therefore be readily compared.

4.4.3

MicroSAGE

4.4.3.1 Introduction

Serial analysis of gene expression (SAGE) is a powerful expression profiling tool, allowing qualitative and quantitative analysis of thousands of transcripts simultaneously (Velculescu *et al.*, 1995; Velculescu *et al.*, 1997; Zhang *et al.*, 1997). In SAGE, short (about 14 base pairs) sequence tags are isolated from each transcript at a defined position, ligated to long multimers, cloned and sequenced. In this way, large numbers of transcripts can be analysed simultaneously, thus generating detailed expression profiles. The tag frequency is directly proportional to the relative transcript abundance in the original mRNA pool, while the tag sequence allows identification of the corresponding transcript in database searches (Velculescu *et al.*, 1995).

SAGE analysis of expression profiles has provided a wealth of data on transcriptional responses in a variety of tissues, cell types and cell lines of diverse origin including human, mouse, rat, yeast and plant, allowing the identification of differentially expressed genes in response to disease or to numerous stimuli (for an overview, see the SAGE homepage at <http://www.SAGEnet.org>). In particular, many studies have focussed on elucidating the transcriptional changes that occur in human cancer, thus identifying markers for tumour invasion, progression and metastasis (Argani *et al.*, 2001; Parle-Mcdermott *et al.*, 2000; Porter *et al.*, 2001; Zhang *et al.*, 1997). Mechanistically, SAGE can provide clues to the metabolic or signalling pathways involved if the expression of several genes in a particular pathway appears to be altered. For example, SAGE analysis of *n-myc*-transfected and control neuroblastoma cells revealed up-regulation of a large number of genes involved in ribosome assembly and activity, suggesting that *myc* genes function as major regulators of the protein synthesis machinery (Boon *et al.*, 2001). In addition, SAGE can provide a detailed characterisation of the transcriptional make-up and

the genetic complexity of a tissue. For example, extensive descriptive SAGE expression profiles have been generated of skeletal muscle (Welle *et al.*, 1999), hippocampus (Datson *et al.*, 2001) and liver (Yamashita *et al.*, 2000). A full description of the SAGE method and discussion of its general advantages and disadvantages may be found elsewhere in this chapter.

Due to the relatively large amount of starting RNA that is required for SAGE (2.5 to five micrograms of poly(A)⁺ mRNA), most SAGE studies have been performed using samples derived from cell lines, yeast cultures, large solid tumours or whole tissue, which consist of millions of cells and are plentiful sources of RNA. However, it is often difficult if not impossible to obtain sufficient starting material for SAGE analysis, for example, when dealing with tissue biopsies, scarce post-mortem material, tumour metastases and small micro-anatomical structures.

A problem related to the use of whole tissue samples is the fact that many tissues are highly heterogeneous, being composed of multiple cell types that display unique expression profiles. Therefore, when whole tissue, individual expression profiles are combined to form a global picture of gene expression. This has consequences not only for the detection of changes in expression that are limited to a small subpopulation of cells of the tissue, but may also mask more generalised transcriptional effects. For example, a five-fold rise in the expression of a particular gene in a cell type that constitutes only five percent of the total tissue will be diluted 20-fold and may no longer be detectable in the whole tissue profile. In addition, if a specific gene is up-regulated in one of the major cell types of a tissue and down-regulated to an equal extent in another equally abundant cell type, the net result will be no change in expression (figure 4.77). By using specific cell popula-

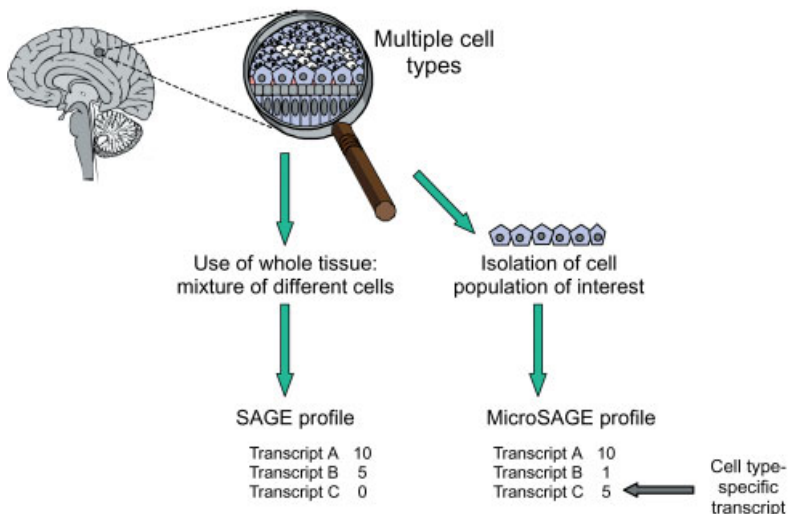


Figure 4.77 Expression profiling in complex heterogeneous tissues, for example brain, generates a global picture of gene expression due to mixing of the expression profiles of

individual cell types. This may mask differential expression of genes restricted to a specific cell population within the whole tissue.

tions for expression profiling rather than whole tissue, detection of novel classes of differentially expressed genes with a highly localised expression or specific regulation in a particular cell type is feasible. However, the rate-limiting step in SAGE analysis of specific cell types or substructures is obtaining sufficient high quality RNA.

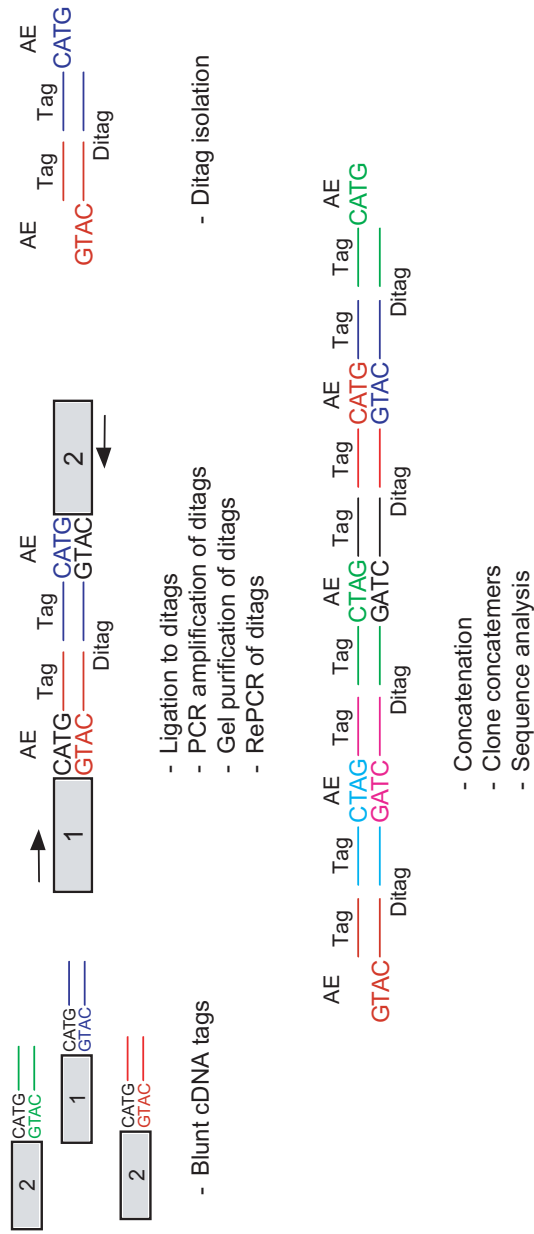
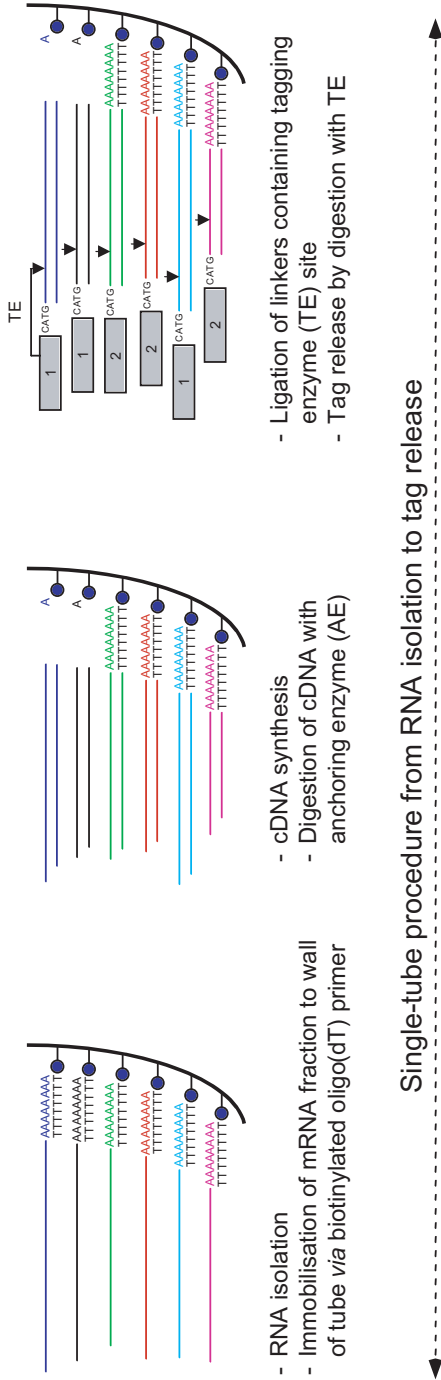
To allow expression profiling in limited sources of RNA or specific cell types of complex heterogeneous tissues, the SAGE procedure requires technical modifications. To address this issue, a modification of the SAGE procedure was developed, called microSAGE, allowing the use of 500 to 5,000-fold less starting material (Datson *et al.*, 1999).

4.4.3.2 Principles and basics

To enable the application of SAGE to small quantities of tissue, several modifications were made to the original procedure (table 4.10). The modifications mostly involve the first steps of the procedure, from RNA isolation to PCR, but leave the basic principles of SAGE unaltered. SAGE is characterised by a large number of sequential reactions and purifications, which can give rise to a significant loss of material. Compared to the original SAGE method, microSAGE is simplified due to the incorporation of a single-tube procedure replacing several of the many steps in SAGE. All steps from RNA isolation to tag release are performed in a single tube in which the RNA, and later the cDNA, remains attached to the wall of the tube by means of streptavidin/biotin binding (figure 4.78). This obviates the need to perform a phenol/chloroform extraction and ethanol precipitation between each step. The enzymes of the previous reactions are simply removed by heat inactivation and disposal of the solution. After washing and a change of buffer, the next reaction can be performed in the same tube. The most important advantage of this single-tube procedure is therefore a reduction of the number of manipulations and, thus, in the accompanying loss of material. Furthermore, total RNA is used rather than poly(A)⁺ mRNA, obviating the need for an additional mRNA extraction

Table 4.10 Comparison of SAGE (serial analysis of gene expression) and microSAGE.

SAGE	MicroSAGE
2.5 to five micrograms poly(A) ⁺ mRNA (<i>i.e.</i> large solid tumours, yeast cultures and cell lines)	One to five nanograms poly(A) ⁺ mRNA (<i>i.e.</i> single hippocampal punch, biopsies and microdissected material)
Multiple tube reaction	Single tube reaction from RNA isolation to tag release: a) fewer manipulations, b) reduces material loss, and c) simplified procedure.
25 to 28 cycles of PCR	28 cycles of PCR and eight to fifteen cycles of rePCR on gel-purified ditag



◀ **Figure 4.78 Schematic representation of the microSAGE procedure.** After isolation of total RNA, a biotinylated oligo(dT) primer is annealed to the RNA. The primer-RNA mixture is subsequently added to a streptavidin-coated tube. After a wash step, the mRNA fraction remains bound to the wall of the tube *via* the biotin/streptavidin bond. First strand and second strand cDNA are synthesised, and the double-stranded cDNA is digested with *Nla* III, all in the same tube. In subsequent enzymatic steps, the reaction mixture is removed from the tube which contains the immobilised cDNA, the tube is washed and the next reaction mix is added. Removal of the digestion mixture after *Nla* III digestion leaves only the most 3'-restriction fragment bound to the wall of the tube. Linkers containing a recognition site for the tagging enzyme are ligated to this 3'-*Nla* III restriction fragment. Subsequent digestion with the tagging enzyme, which recognises the site in the linker but cuts at a given distance in the unknown cDNA sequence, results in release of

the linker with attached cDNA tag of approximately ten base pairs. All steps so far from mRNA isolation up to the tag release are performed in the same single tube, reducing the loss of material significantly and simplifying the procedure. After tag release, the tags are blunted and ligated to ditags (chemical modification of the linkers prevents formation of multimers). The ditag population is amplified by PCR using primers directed against the linker sequences. After 28 cycles of PCR the ditags are gel-purified and subjected to another twelve to eighteen cycles of PCR (rePCR). The linkers are subsequently removed from the ditags by digestion with *Nla* III. Addition of ligase to the ditags results in the formation of ditag multimers reamplification, which are cloned and sequenced. Analysis of sequenced concatemers using SAGE software allows tag quantification and tags to be linked to their corresponding mRNA sequences in GenBank or UniGene clusters.

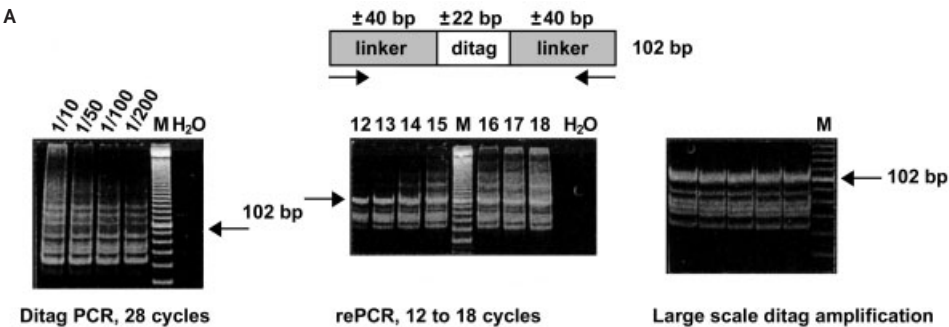
step. Instead, the poly(A)⁺ mRNA fraction is directly bound to the streptavidin-coated wall of the tube *via* a biotinylated oligo(dT) primer, which also serves as a primer in the subsequent cDNA synthesis. Another change from the original protocol is that a limited number of cycles of re-amplification PCR (rePCR) are performed on the gel-purified ditag band to generate sufficient ditags.

Using microSAGE, an expression profile can be obtained from as little as on to five nanograms of mRNA, allowing expression profiling in small tissue specimens, microdissected parts of complex heterogeneous tissues, biopsies, post mortem material, scarce biological material and so forth, thus broadening the applications of SAGE enormously.

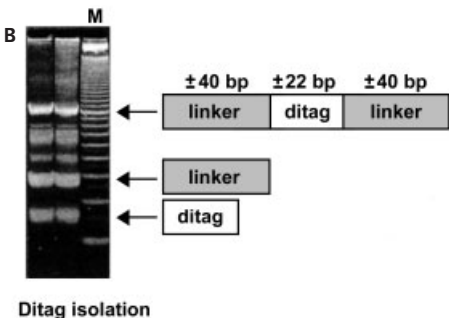
4.4.3.3 Presentation of typical results

A typical example of the different steps of the microSAGE procedure is shown in figure 4.79 (see figure legend for more details).

A

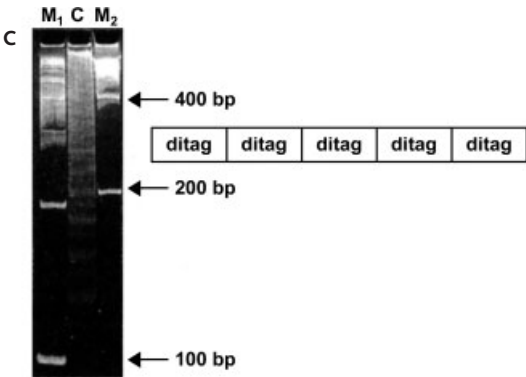


B



Ditag isolation

C



Concatemerisation of ditags

◀ **Figure 4.79 Typical example of the different steps of the microSAGE procedure.** Ethidium bromide-stained polyacrylamide gels (Twelve percent in panels **A** and **B**, eight percent in panel **C**) showing examples of several steps in the microSAGE procedure. **A)** PCR amplification of the ditag. The left panel shows 28 cycles of PCR of various dilutions (1/10, 1/50, 1/100 and 1/200) of one microlitre of the ligated ditag derived from punch material and a negative control performed on water. The 102 base pairs band corresponding to the amplified ditag is faintly visible among several other background bands. After excision of the ditag band and extraction of the DNA, a series of PCRs with varying number of cycles (in this case, twelve to eighteen cycles) is performed to determine the optimal number of cycles of rePCR (middle panel). The negative control performed on

water is amplified for 30 cycles. After large-scale rePCR (in this particular case, twelve cycles of re-PCR were considered optimal), the PCR products are concentrated and run on a preparative gel from which the 100 base pairs ditag band is excised (right panel). **B)** After digestion with *Nla* III to cleave off the linkers, the small ditag of 22 to 26 base pairs is excised and purified. **C)** The isolated ditags are ligated to concatemers that are size-separated on a polyacrylamide gel. The region of the gel containing concatemers ranging from 400 to 800 base pairs and larger than 800 base pairs are excised, after which the purified concatemers are cloned in plasmids. M: ten base pairs ladder; M₁: 100 base pairs ladder; M₂: 200 base pairs ladder; C: concatemers. The figure was reproduced with permission from *Nucleic Acids Research* (Datson *et al.*, 1999).

4.4.3.4 Examples of application of microSAGE

The central nervous system (CNS) in particular is a prime target for application of microSAGE due to its extremely complex architecture of different highly specialised cell types. Because each neuron has its own specific synaptic input, even two adjacent neurons in the same brain substructure may have different expression profiles. In addition, changes in gene expression in the central nervous system are often small (less than two-fold), unlike the huge changes in gene expression seen in other tissues such as, for example, tumour tissue (Zhang *et al.*, 1997). Another complicating factor in expression profiling in the central nervous system is the high number of different genes expressed in neuronal tissue, giving rise to an extremely complex transcriptome (Datson *et al.*, 1999). It is this combination of extreme cellular heterogeneity, relatively small changes in gene expression and complex transcriptome that hampers expression profiling in the central nervous system.

This problem is illustrated nicely by efforts to identify corticosteroid-responsive genes in the hippocampus, a brain structure with an important function in learning and memory which is extremely sensitive to aberrant levels of corticosteroids. The hippocampus is comprised of several different sub-fields (*e.g.*, CA1 to CA4, dentate gyrus) and is a heterogeneous tissue consisting of different pyramidal neurons, granule neurons, at least twelve types of interneurons, as well as different glial cell species (Vizi *et al.*, 1998). If the hippocampus is completely deprived of corticosteroids for a number of days, specific granule neurons in the inner blade of the dentate gyrus die by apoptosis, while other sub-regions appear to be unaffected (Sloviter *et al.*, 1989). In contrast, chronic excess of corticosteroids causes atrophy of CA3 pyramidal neurons (figure 4.80) (Watanabe *et al.*, 1992). CA1 neurons, by contrast, seem to be resistant to the harmful effects of aberrant corticosteroid levels. It is this combination of cellular heterogeneity and pronounced regio-

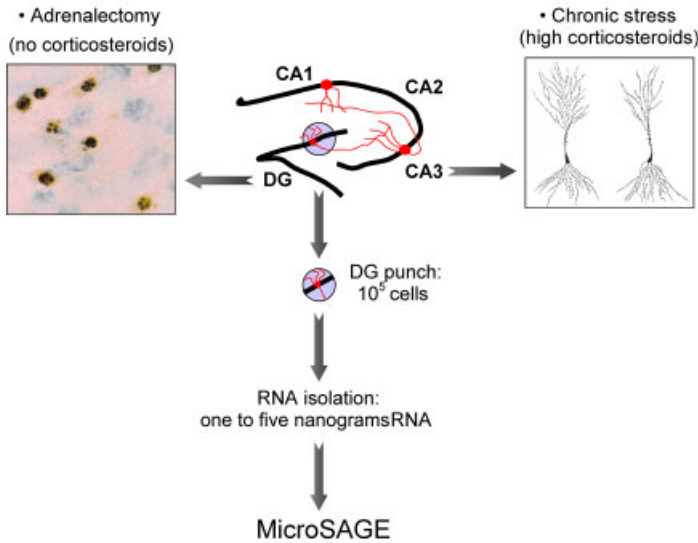


Figure 4.80 A schematic representation of the hippocampus, indicating the different sub-fields and their difference in corticosteroid-dependent neuronal vulnerability. Chronically high concentrations of corticosteroids during periods of chronic stress may lead to atrophy of CA3 neurons (right panel). However, physiological levels of corticosteroids are necessary for hippocampal viability as lack of corticosteroids after adrenalectomy may lead to apoptosis of granule cells in the dentate gyrus (left panel).

Therefore, to facilitate identification of corticosteroid-responsive genes in the hippocampus, it is preferable to specifically isolate sub-fields and use them for expression profiling rather than the whole hippocampus. A single punch containing approximately 10,000 cells was taken from the inner blade of the dentate gyrus out of a rat brain slice. RNA isolated from the punch was used as input in the microSAGE procedure.

nal differences in corticosteroid-dependent neuronal viability that complicates detection of corticosteroid-responsive genes in the hippocampus. Practically the only way to obtain sufficient amounts of RNA for expression profiling is by using whole hippocampi. However, this results in dilution of region-specific or even neuron-specific transcriptional effects of corticosteroids. It is therefore likely that transcriptional changes that are restricted to certain hippocampal sub-regions or even to certain neurons will be lost by pooling RNA from the entire hippocampus.

Using microSAGE, a SAGE profile was obtained from a single hippocampal punch containing a small part of the dentate gyrus (Datson *et al.*, 1999), showing that it is possible to zoom in on a highly specialised brain region and to obtain a region-specific expression profile (figure 4.80). This was the first demonstration that profiles could be generated from 1,000 to 5,000 times less starting material than normally required for SAGE.

A recent application of microSAGE was the investigation of the transcriptional make-up of undifferentiated human mesenchymal stem cells isolated from adult

bone marrow (Tremain *et al.*, 2001). Mesenchymal stem cells are capable of differentiating into several different cell types including adipocytes, chondrocytes and myoblasts. A single cell-derived colony of human mesenchymal stem cells, consisting of approximately 10,000 cells, was used for microSAGE analysis. A total of 17,767 tags were analysed, consisting of 2,353 unique genes. Detailed characterisation of the expression profile revealed that the mesenchymal stem cells colony simultaneously expressed transcripts that are characteristic of various mesenchymal cell lineages, thus reflecting the enormous developmental potential of these stem cells.

Since the initial publication of microSAGE, several other microSAGE-like methods have been published; these will be discussed in more detail below.

4.4.3.5 Discussion

The initial microSAGE publication demonstrated the feasibility of miniaturisation of SAGE by obtaining an expression profile of a single dentate gyrus punch, containing between 500 to 5,000 times less RNA than normally required for SAGE. An initial disadvantage in the protocol of microSAGE was the relatively frequent occurrence of a single ditag species derived from linker sequences, which reduced the average number of tags obtained per sequenced clone. Formation of this linker ditag may be reduced by lowering the amount of linkers used in the ligation to the cDNA, and by using larger washing volumes and multiple washing steps after linker ligation to remove the non-ligated linker remnants which give rise to this product. Indeed, in a microSAGE study on human mesenchymal stem cells, a number of these improvements to the protocol have been incorporated, largely solving this problem (Tremain *et al.*, 2001).

Since the first description of microSAGE, several other SAGE modifications have been described with the aim of reducing the amount of input RNA, such as SAGE adaption for downsized extracts (SADE) (Virlon *et al.*, 1999), SAGE-Lite (Peters *et al.*, 1999), PCR-SAGE (Neilson *et al.*, 2000) and miniSAGE (table 4.11) (Ye *et al.*, 2000) (described elsewhere in this chapter).

The use of minute amounts of starting material demands additional amplification of the input material compared to the original SAGE procedure to enable experimental manipulation. Almost all modified SAGE methods described so far (with exception of miniSAGE) include such an amplification step, either at the cDNA level before SAGE ditags have been established, or after formation of the ditags.

In both PCR-SAGE and SAGE-Lite, the cDNA is PCR-amplified before ditag formation using a template-switching oligonucleotide. In each first strand cDNA synthesis, the terminal transferase activity of reverse transcriptase adds a few non-template nucleotides (mostly dCTPs) to the 3'-end of the newly synthesised cDNA strand upon reaching the 5'-end of the mRNA template. When a special oligonucleotide, the so called template-switching or TS oligonucleotide which contains a few dGTPs at its 3'-end, is present in the reaction mix, it anneals with the dC stretch to form an extended template. The reverse transcriptase then

Table 4.11 Comparison of different SAGE (serial analysis of gene expression) modifications aimed at use of limited amounts of starting material.

Modification	Amount of starting material	Amplification of	Number of PCR cycles	Other characteristics
MicroSAGE (Datson <i>et al.</i> , 1999)	10,000 cells or one to five nanograms mRNA	Ditags	28 and additional twelve	Single-tube procedure from RNA analysis to cDNA tag release using streptavidin-coated tubes → Reduces loss of material
SADE (Virlon <i>et al.</i> , 1999)	0.5 milligrams tissue or 50,000 cells	Ditags	29 to 31 and additional twelve	Single-tube assay from tissue lysis to cDNA tag release using oligo(dT) magnetic beads → Reduces loss of material
PCR-SAGE (Neilson <i>et al.</i> , 2000)	900 picograms mRNA	cDNA and ditags	29 cycles of cDNA 28 to 30 cycles of ditags	cDNA amplification based on incorporation of template switching-oligo nucleotide during cDNA synthesis; cDNA is amplified using single biotinylated oligonucleotides → Fragments biotinylated at both ends; digestion with <i>Sap</i> I removes biotin at 5'-end of all transcripts cDNA amplification based on incorporation of template switching-oligo nucleotide during cDNA synthesis; cDNA is amplified using a biotinylated oligo(dT) primer and template switching-oligo nucleotide → Biotinylation at 3'-end of all transcripts
SAGE-Lite (Peters <i>et al.</i> , 1999)	100 nanograms total RNA	cDNA and ditags	22 cycles of cDNA 28 to 30 cycles of ditags	Single-tube procedure from RNA analysis to cDNA tag release using streptavidin-coated tubes → Reduces loss of material
MiniSAGE (Ye <i>et al.</i> , 2000)	One microgram total RNA	Ditags	28	Use of phase lock gel for phenol extractions → Reduces loss of material

switches templates and incorporates the template-switching oligonucleotide at the 3'-end of the cDNA. The resulting single-stranded cDNA contains the complete 5'-end of the mRNA template, as well as the sequence complementary to the template-switching oligonucleotide, allowing PCR amplification of the entire cDNA population between the 5'-template-switching oligonucleotide and the poly(A) tail (Diatchenko *et al.*, 1996; Herrler, 2000; Lukyanov *et al.*, 1997). A disadvantage of this strategy is a risk that not all cDNA molecules are amplified equally well, altering the stoichiometry of the transcripts. A particular problem is amplification bias in favour of smaller cDNA molecules, causing larger transcripts to be underrepresented in the ditag population and thus also in the subsequent SAGE profile.

In microSAGE and SADE (SAGE adaptation for downsized extracts), the additional amplification which is needed in order to allow use of small amounts of starting material is performed after formation of the ditags. The ditag population consists of an almost infinite number of possible species resulting from different combinations of tags within a ditag. Tags derived from abundant transcripts are present in numerous different ditag combinations, while tags derived from rare transcripts will be much less frequent. The formation of ditags is a crucial step in the SAGE procedure, since the direct relationship between transcript abundance and tag frequency is consolidated in the resulting ditag population. Additional PCR amplification of the ditags, as is performed in microSAGE and SADE, should not jeopardise the quantitative aspect of the SAGE data. First, since all ditag species are of approximately equal length, this PCR step should be relatively free of bias. However, because of the nature of the PCR reactions preferential amplification of some ditag species cannot be prevented entirely. These ditags will then be in the sequenced concatemers. Because the special SAGE software only includes each exclusive ditag combination once in the analysis, PCR artefacts are excluded.

Besides incorporating additional PCR cycles for ditag amplification, another essential modification of microSAGE is its use of a more efficient single-tube assay. This reduces the loss of material during the procedure and allows significant reduction of the amount of input RNA. SADE and miniSAGE both also make use of a single-tube assay for all steps of the SAGE procedure until the tags are released by digestion with the tagging enzyme. In addition, in miniSAGE, the loss of material after tag release is further reduced by using phase lock gels to increase the recovery and purity of SAGE material from the aqueous upper phase after phenol extraction following tag release. The authors of miniSAGE claim that the reduction of loss of material in miniSAGE is so low that additional PCR of cDNA or ditags is not necessary (Ye *et al.*, 2000).

It can be expected that as the technology advances, future expression profiling studies will focus more and more on transcriptome analysis of specific cell populations or even single cells. In this respect, amplification of the starting material will become increasingly important. Linear amplification of the mRNA pool might be a suitable alternative for the amplification methods described above. In linear amplification, first strand cDNA is synthesised using an oligo(dT) primer that has a tail containing a T7 RNA polymerase promoter sequence. After conversion of the single-stranded cDNA to double-stranded cDNA, all cDNA molecules contain a func-

tional T7 RNA polymerase promoter at their 3'-end. By addition of T7 RNA polymerase the cDNA molecules are transcribed *in vitro* into antisense RNA, yielding up to 1,000-fold amplification of the cDNA population in a single round (Eberwine *et al.*, 1992). Application of this method has been shown not to affect the mRNA stoichiometry and has been used to detect expression differences in single cells (Eberwine *et al.*, 1992; Nair *et al.*, 1998).

4.4.4

MiniSAGE

4.4.4.1 Introduction

Since its first description in 1995 (Velculescu *et al.*, 1995), serial analysis of gene expression (SAGE) has become an established and widely-used method to determine genome-wide gene expression profiles in eukaryotic cells and tissues. The principle of the SAGE method is described in a previous section of this book. The original SAGE procedure requires about two to five micrograms of poly(A)⁺ mRNA. For this reason, several technical modifications of the SAGE procedure have been developed such as microSAGE, PCR-SAGE, PCR-Lite and SADE to allow expression profiling in samples with starting material amounting to less than one microgram of total RNA (*e.g.*, small samples derived from laser microdissection). These modifications are described in detail in other sections of this chapter (see also Datson *et al.*, 1999; Neilson *et al.*, 2000; Peters *et al.*, 1999; Virlon *et al.*, 1999).

The use of minute amounts of starting material requires an additional amplification step. Almost all modified SAGE methods described so far include such an amplification step, either at the cDNA level before SAGE ditags have been established, or after formation of the ditags. By contrast, the miniSAGE method dispenses with an additional PCR amplification step, but nevertheless requires only one microgram of total RNA, several hundred times less than that required by the original SAGE protocol (Ye *et al.*, 2000). This is advantageous because PCR amplification potentially introduces bias and may compromise the quantitative aspects of the SAGE method.

4.4.4.2 Principles and basics

The miniSAGE technique is based on three key modifications of the SAGE procedure. (i) Use of a phase lock gel system to increase the recovery and the purity of DNA material after each phenol extraction step. (ii) Reduction of the amount of linker oligonucleotides in the ligation, thereby minimising their interference with SAGE ditag amplification and increasing the yield of SAGE ditags. (iii) Integration of an mRNA capture kit to allow the first five steps of the SAGE procedure (mRNA isolation, reverse transcription of mRNA into cDNA using a biotin-labelled oligo(dT) primer, enzymatic digestion of cDNA, binding of digested biotin-labelled 3'-terminal cDNA fragments to streptavidin-coupled magnetic beads, ligation of linker oligonucleotides containing recognition sites for a tagging enzyme to the

bound cDNA fragments, and release of cDNA tags) to be performed within one tube. This one-tube procedure prevents loss of material between successive steps.

In the miniSAGE procedure, the basic protocol is as described in the original SAGE procedure (Velculescu *et al.*, 1999) with the integration of the three modifications mentioned above. This means that in a first step, poly(A)⁺ mRNA is isolated from total RNA by use of an oligo(dT) oligonucleotide linker-based procedure. Complementary DNA is synthesised from poly(A)⁺ mRNA using a biotin-labelled oligo(dT) primer. The cDNA is then digested with an anchoring restriction enzyme such as *Nla* III. The resulting 3'-terminal cDNA fragments are bound to streptavidin-coated magnetic beads. An oligonucleotide linker containing recognition sites for a tagging restriction enzyme such as *Bsm* FI is linked to the bound cDNA. A short tag including the linker sequence is released from the beads by enzymatic digestion with the tagging enzyme *Bsm* FI. The 3'-ends of the released short tags plus the linker sequences are blunted and ligated to one another to form ditags. After PCR amplification, the ditags are digested with the anchoring enzyme, purified, concatemerised, and cloned into plasmids. Sequencing of the concatemers enables quantification of tags, thus allowing the identification and quantification of abundance of transcripts for a given sample.

4.4.4.3 Typical results

The miniSAGE method was used to generate two fibroblast SAGE libraries representing RNA derived from a healthy donor and from a patient with premature coronary artery disease and hyperapobetalipoproteinaemia (*i.e.* raised circulating apolipoprotein B concentrations) (Ye *et al.*, 2000). In total, 916 and 3,838 tags, respectively, were sequenced from these two libraries. Unfortunately, the number of sequenced tags was too small to identify significant differences in gene expression between these two patients. For this reason, most SAGE tags identified (82 percent) were found only once in the libraries.

However, the 3,838 tags derived from the second library represented 2,308 unique genes of which 1,407 (61 percent) matched to characterised transcripts, 711 (31 percent) matched to expressed sequence tags and 190 (eight percent) did not match to any characterised genes. As expected, most abundantly expressed genes were housekeeping genes such as ribosomal proteins, elongation factor 2, ferritin, glyceraldehyde-3-phosphate dehydrogenase, and thymidylate kinase (Ye *et al.*, 2000). These results are consistent with the findings of others (Ye *et al.*, 2000; Zhang *et al.*, 1997). However, the data allowed the authors to conclude that the miniSAGE procedure is a promising tool to analyse gene expression profiles in eukaryotic cells and tissues with limited starting material.

4.4.4.4 Discussion

The modified SAGE protocol called miniSAGE established by Ye *et al.* (2000) allows profiling of gene expression patterns with only one microgram of total RNA as starting material without an additional PCR amplification as it is performed in other SAGE modifications such as microSAGE, PCR-SAGE and PCR-Lite (Datson *et al.*, 1999; Neilson *et al.*, 2000; Peters *et al.*, 1999). This is an incontestable advantage compared to other SAGE modifications based on an additional PCR amplification. As yet, however, the only experience gathered with this procedure is that generated by Ye *et al.* (2000), the researchers who described the method.

4.4.5

ORESTES – Open reading frame expressed sequence tags

4.4.5.1 Introduction

Estimations of the number of genes within the human genome vary widely. Before completing the sequencing of the human genome, estimations based on different methods ranged between 30,000 and 120,000 genes encoded in approximately 3,200,000,000 nucleotides (see, for example, Antequera & Bird, 1993; Crow & Kimura, 1970; Dunham, 2000; Ewing & Green, 2000; Fields *et al.*, 1994; Lewin, 1980 and 1990; Liang *et al.*, 2000; Müller, 1967; Müller & Kern, 1967; Roest *et al.*, 2000; Smaglik, 2000; Zhuo *et al.*, 2001). To the surprise of many, the completion of the working draft sequence of the genome by Venter *et al.* (2001) and Lander *et al.* (2001) predicted the number of genes to be at the lower end of these previous estimations, *i.e.* between 30,000 and 40,000.

Estimates of gene number are based on several different improved gene-finding algorithms (for details see literature cited above). Irrespective of the approach used, each predicted gene and its genomic structure must be verified by experiment. For single genes, the experimental task is not too difficult; a cloned cDNA fragment is sequenced and the complete sequence of the cDNA is identified by, for example, using rapid amplification of cDNA ends (RACE) (Frohman, 1994; Schaefer, 1995) to characterise its 5' and 3'-ends. Large-scale identification of transcripts with the aim of detecting of open reading frames (ORFs) and the genomic structure of genes is a much more daunting matter.

Large-scale analysis of transcriptomes has its origin in the concept of sequencing expressed sequence tags (ESTs) (Adams *et al.*, 1992). In this approach, cDNA libraries are created and subjected to single-pass sequencing. The cloned cDNA fragments represent information on the abundance of transcripts and help to identify genes experimentally. Usually, the cloned cDNAs are incomplete. In the past, tags were about 300 nucleotides in length, but ESTs of more than 700 nucleotides are now common (Strausberg & Riggins, 2001). The EST approach can either be used to identify the transcriptome of a tissue or cell population or to identify differences in gene expression between different cell type or tissues (see chapter 7 for a description of the comparative EST analysis approach).

Traditional EST approaches are now complemented by newer strategies such as serial analysis of gene expression (SAGE). In this approach, short sequence tags

are produced that are located adjacent to defined restriction sites near the 3'-end of the cDNA. Theoretically, each transcript has a unique tag, thereby facilitating transcript quantification. The SAGE approach is described in a previous section of this chapter.

Several problems arise using these two methods to define a complete transcriptome: (i) Using SAGE, only short oligonucleotide tags are generated that do not allow a prediction of full-length cDNA sequences or gene structure. (ii) Approximately 65 percent of known expressed sequence tags represent the 3'-extremity of cDNAs, resulting in a biased representation of expressed gene sequences. (iii) Genes with low expression levels are rarely detected so that sequences of low abundance mRNAs are underrepresented in conventional datasets. (iv) 5'-expressed sequence tags are somewhat random because of premature termination of the reverse transcription reaction. Figure 4.81 shows an idealised schematic view of expressed sequence tags generated by various public projects (Strausberg & Riggins, 2000).

Recently, Camargo *et al.* (2001) described the open reading from expressed sequence tags (ORESTES) strategy. This approach was developed by a collaboration between the FAPESP/LICR Human Cancer Genome Project contributed by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; São Paulo, Brazil) and the Ludwig Institute for Cancer Research (LICR; São Paulo, Brazil) (Camargo *et al.*, 2001). The advantage of ORESTES is that a high proportion of

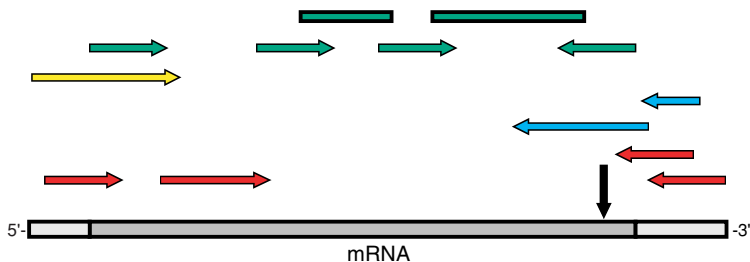


Figure 4.81 Idealised schematic view of expressed sequence tags (ESTs) generated by various public projects. A full-length cDNA is represented by the grey rectangle, with 5' and 3'-untranslated regions indicated by the light grey sections. The red arrows depict 5' and 3'-expressed sequence tags based on the Merck Gene Index (MGI) approach (see also table 4.12); blue arrows depict sequence tags derived from the Cancer Genome Anatomy Project (CGAP) 3'-approach. Note that at both the 5' and 3'-ends alternate expressed sequence tag positions are possible based on transcript variations or incomplete cDNA synthesis. The purple arrow indicates the Mammalian Gene Collection (MGC) expressed sequence tag

strategy in which 5'-expressed sequence tags are generated to search for full-open reading frame clones. ORESTES tags are shown by the green arrows, which are spaced more evenly throughout the cDNA sequence. The green bars denote regions where sequence gaps exist that might be subject to the transcript-finishing approach of Camargo *et al.* (2000). The black arrow indicates where a SAGE tag might be located. This arrow is vertical to indicate that SAGE tags are located at a precise site within a transcript. The figure was adapted from Strausberg & Riggins (2001) with permission from *Proceedings of the National Academy of Sciences of the USA*.

the sequence tags are in the coding regions of transcripts. Moreover, the approach has a normalisation effect for a broader sampling of many different transcripts, with less dependence on expression levels (Strausberg & Riggins, 2001).

4.4.5.2 Principle and basics

As originally described, the ORESTES approach was applied to total RNA isolated from excised tissues enriched in tumour cells using microdissection. DNase I treatment was performed to remove genomic DNA contamination (Dias Neto *et al.*, 2000). After control of RNA, quality by Northern blot analysis and confirmation of absence of contaminating genomic DNA by PCR analysis, poly(A)⁺ mRNA was isolated from the samples. Following reverse transcription of mRNA into cDNA using randomly selected primers (see Dias Neto *et al.*, 2000, for details on selection of primers for selective amplification of fragments of the open reading frames), diluted single-stranded cDNA samples were amplified by PCR using the same or a single, alternative pair of primers. The resulting PCR product pools were then checked on silver-stained polyacrylamide gels for complexity. Amplification pools with a single, predominant product reflecting the amplification of a highly abundant gene were not processed further. Pools showing multiple bands were cloned into plasmids for sequencing. See figure 4.82 for a flowchart of the ORESTES procedure.

In ORESTES, the experimental sequence data is analysed by an automated protocol that (i) assesses the quality of sequences, (ii) trims vector and plasmid sequence, (iii) removes undesirable sequences such as those sequences derived from bacterial, mitochondrial, and ribosomal RNA, (iv) masks repetitive elements, and (v) undertakes serial basic local alignment sequence tool (BLAST) searches against existing databases (Dias Neto *et al.*, 2000).

4.4.5.3 Typical results and examples

The potential of the ORESTES technology has been demonstrated in three steps. In a first step, a database of 10,000 sequences from human breast tumours was created to demonstrate the ability of the ORESTES method (Dias Neto *et al.*, 2000). These sequences were generated from PCR fragments derived from the central, coding regions of the transcripts as mentioned in the principles and basics section. This study showed the potential of ORESTES to act as the basis of a shotgun approach to the sequencing of both known and novel human transcripts.

In a second step, a set of 250,000 tags was constructed (de Souza *et al.*, 2000). These tags were assembled in 81,429 contigs and compared to the complete sequence of human chromosome 22. Exactly 1,181 contigs matched to chromosome 22 and allowed the identification of 162 of the 247 known genes on this chromosome. Sixty-seven of the 150 related genes and of 45 of the 148 genes predicted from expressed sequence tags were also identified. In addition, a further 219 previously non-annotated transcribed sequences were characterised. Of these, 171 were defined by expressed sequence tags or cDNA sequences available in public

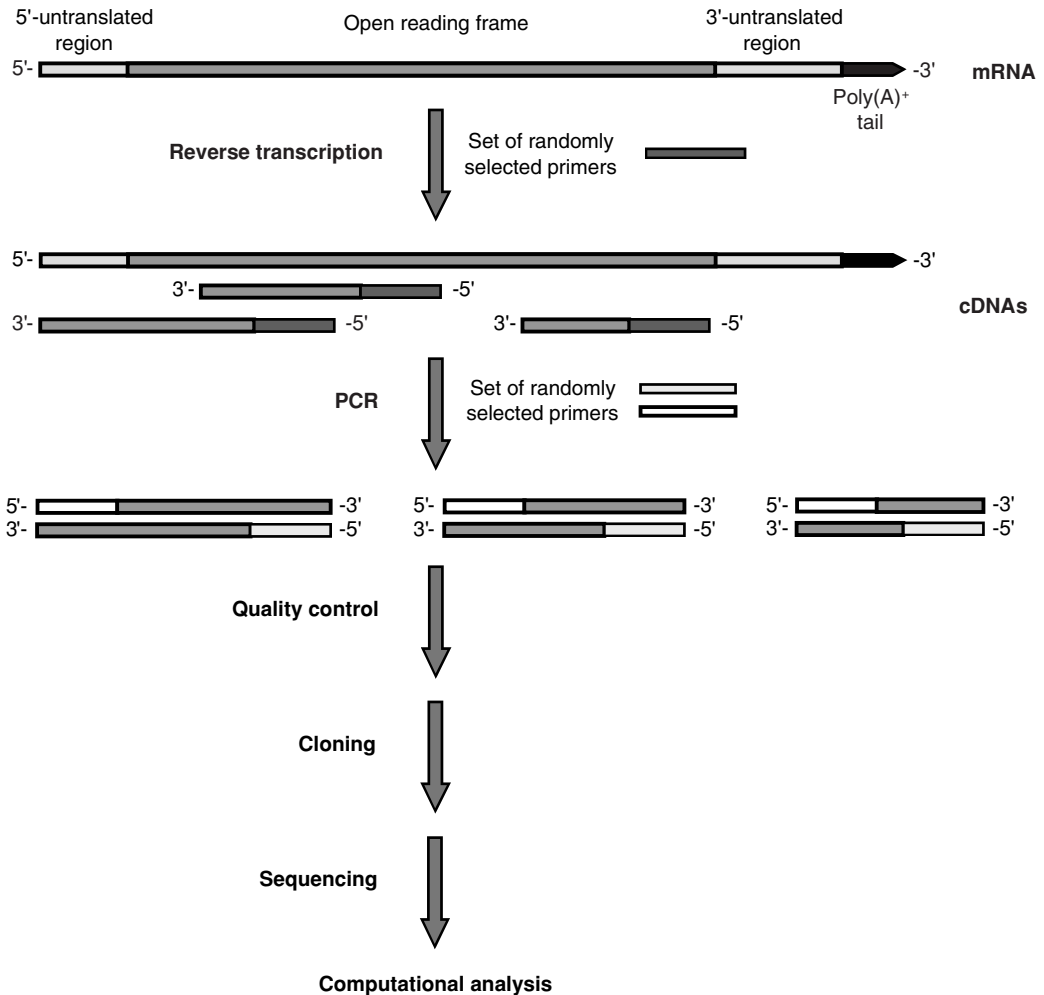


Figure 4.82 Principle of the open reading frame expressed sequence tags (ORESTES) approach. See text for details on the ORESTES procedure.

database but not utilised in the initial annotation of the first human chromosome 22 sequence. ORESTES also allowed the identification of further 48 transcribed sequences on this chromosome not defined by other sequences in public databases. All of these sequences defined using the ORESTES approach coincided with DNA regions which have been predicted to encode exons.

In a third step, a set of nearly 700,000 ORESTES sequences was constructed from samples derived from 24 human tissues (Camargo *et al.*, 2001). Using a subset of these data compared to a set of about 15,00 full-length mRNA sequences, the authors showed that the ORESTES technology made it possible to identify over 80 percent of the highly and moderately expressed genes of a tissue. Between 40 and

50 percent of the rarely expressed genes were also found. In this respect, the capacity of the ORESTES approach for gene discovery and shotgun transcript sequence generation seems to exceed that of conventional expressed sequence tag generation.

4.4.5.4 Discussion

In the ORESTES approach, low-stringency PCR conditions are used to produce cDNA libraries from which a relatively small number of individual clones are sequenced. To cover an almost complete transcriptome several thousand ORESTES libraries are produced, each with different primers, such that each library is expected to contain unique cDNA sequences.

Compared to other approaches, ORESTES seems to have the following advantages. (i) As described by Camargo *et al.* (2001), experimental results confirm the theoretical expectations of ORESTES in that sequences identified by sequencing are spaced throughout the transcripts, thereby providing a scaffold to complete full-length transcript sequences. (ii) The normalisation effect of the ORESTES approach for a broader sampling of many transcripts with less dependence on transcript abundance facilitates the discovery of genes with low expression levels. However, the normalisation effect of the ORESTES approach that allows the identification of rarely expressed genes makes it impossible to use the ORESTES technology for differential gene expression analysis (Camargo *et al.*, 2001). As with all methods, ORESTES also has its disadvantages. (i) The approach is extremely time and labour-intensive, a feature it shares with most high-throughput sequencing approaches. (ii) Due to the low-stringency of the RT-PCR used in the ORESTES approach, the use of very pure and high quality RNA is required.

Efforts to sequence and annotate cDNA with a view to identify the genes encoded in the human genome are of great current importance. The ORESTES approach represents a valuable addition to such efforts (table 4.12).

Transcriptome analysis differs from genome sequencing in that the 'biological space' of the transcriptome still remains to be defined. In the past, much effort has been directed to the molecular characterisation of cancer, but it is evident that comprehensive study of the transcriptome will require more substantial study of normal human tissues and cells as well as tissues from a variety of disease states. ORESTES is a valuable tool to provide more detailed information on the entire transcriptome.

Table 4.12 Major database approaches providing data on cDNA structures and their distribution (see also chapter 7).

Database	Internet link	References	Description
Cancer Genome Anatomy Project (CGAP)	http://cgap.nci.nih.gov/	Riggins & Strausberg (2001)	The Cancer Genome Anatomy Project (CGAP) is a collaborative network of cancer researchers with a common goal: to decipher the genetic changes that occur during cancer formation and progression. The project brings together several recent technologies capable of high-throughput analysis to help achieve this goal and offers several tools that are freely distributed.
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/	Boguski <i>et al.</i> (1993)	dbEST is a division of GenBank that contains a vast collection of expressed sequence tags (ESTs). Most ESTs are 300 to 500 base pairs reads from cDNAs derived from mRNA expressed in a variety of cell types and organisms. dbEST is enriched in 3'-ends, many of which do not contain much coding sequence (see chapter 7 for details).
German Human cDNA Project	http://www2.mips.bio-chem.mpg.de/proj/cDNA/index.html	Wiemann <i>et al.</i> (2001)	The German Human cDNA Project was formed by a consortium of eight experienced DNA sequencing laboratories in the frame of the German Human Genome Project (DHGP) with the aim to characterise a large set of cDNA clones not yet identified in other projects, and to contribute a part towards the systematic identification and functional characterisation of the human genes.
IMAGE consortium	http://image.llnl.gov/	Lennon <i>et al.</i> (1996)	The IMAGE resource shares high-quality, arrayed cDNA libraries and places sequence, map, and expression data on the clones in these arrays into the public domain. The human and mouse genomes were the first to be studied, and the collection now contains clones from rat, zebrafish, <i>Xenopus laevis</i> and rhesus macaque.

Table 4.12 (continued)

Database	Internet link	References	Description
Mammalian Gene Collection (MGC) project	http://mgc.nci.nih.gov	Strausberg <i>et al.</i> (1999)	The MGC project is an effort to generate full-length cDNA publicly accessible resources. The MGC project entails the production of libraries, sequencing, database and repository development, and the support of library construction, sequencing, and analytic technologies to obtain a full set of human and other mammalian full-length sequences and clones of expressed genes.
Merck Gene Index (MGI)	–	Eckman <i>et al.</i> (1998); Williamson (1999)	The Merck Gene Index project fills a niche in the Human Genome Project by directly identifying genes through sequences of their transcripts.
RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/	Maglott <i>et al.</i> (2000); Pruitt & Maglott (2001); Pruitt <i>et al.</i> (2000)	RefSeq provides reference sequence standards for genomes, transcripts and proteins; human, mouse and rat mRNA. Together with LocusLink, RefSeq provides a non-redundant view of genes and other loci.
RIKEN Mouse Gene Encyclopaedia Project	http://www.gsc.riken.go.jp/	Kawai <i>et al.</i> (2001)	The RIKEN Mouse Gene Encyclopaedia Project is a systematic approach to determining the full coding potential of the mouse genome and involves collection and sequencing of full-length cDNAs and physical mapping of the corresponding genes to their genome.
Serial analysis of gene expression tag to gene mapping database (SAGEmap)	http://www.ncbi.nlm.nih.gov/SAGE/	Lal <i>et al.</i> (1999)	The public SAGEmap was created as a component of CGAP to provide a central location for depositing, retrieving, and analysing human gene expression data. This database uses SAGE to quantify transcript levels in both malignant and normal human tissues.
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/	Wheeler <i>et al.</i> (2000, 2001 and 2002)	UniGene groups the sequences of dbEST into organism specific clusters, each cluster representing a unique gene. Currently, UniGene is available for six vertebrates (human, rat, mouse, cow, zebrafish and frog) and five plants (thale cress, wheat, barley, rice and corn).

4.4.6

PCR-SAGE and SAGE-Lite**4.4.6.1 Introduction**

Serial analysis of gene expression (SAGE) has proved to be an invaluable method for describing the transcriptome, both by revealing those genes expressed in a given tissue or cell type and by quantifying their relative expression levels. It also offers an avenue for discovery of novel genes. These aspects of SAGE have been discussed in detail in previous sections of this book.

One limitation of SAGE is its requirement for relatively large quantities of RNA. The original SAGE protocol developed by Velculescu *et al.* (1995) recommended using two to five micrograms of starting mRNA (approximately 50 to 100 micrograms of total RNA). Developing a protocol referred to as miniSAGE, Qing-Ye *et al.* (2000) reduced the amount of total RNA needed to as little as one microgram. However, for many biological samples it is not possible to obtain even this small amount of RNA. This has led to the independent development of two closely related modifications of the SAGE technique: SAGE-Lite (Peters *et al.*, 1999) and PCR-SAGE (Neilson *et al.*, 2000).

PCR-SAGE and SAGE-Lite both utilise the polymerase chain reaction (PCR) to amplify the transcriptome as a first step, lifting the amount of starting material for SAGE from picogram to the requisite microgram levels. Using PCR-SAGE, it was possible to produce a SAGE catalogue from nine human germinal vesicle oocytes (approximately 900 picograms of mRNA) and to identify transcripts required by the egg during the late stage of maturation (Neilson *et al.*, 2000). In a similar vein, Peters *et al.* (2001) were able to develop a SAGE catalogue from cerebral artery biopsies to reveal transcriptome changes during intracranial aneurysm using SAGE-Lite.

4.4.6.2 Principle of the PCR-SAGE procedure

For PCR-SAGE, total RNA is extracted using Trizol (Invitrogen Ltd., Paisley, United Kingdom) in accordance with the manufacturers protocols. For this RNA to be amplified by PCR, it must subsequently be modified to contain priming sites at both the 3' and 5'-ends of each molecule. The resulting cDNA must also be biotinylated at the end corresponding to the poly(A)⁺ tail of the original mRNA to provide for capture of the 3'-most *Nla* III fragment in the SAGE step. The PCR-SAGE technique achieves these requirements by modifying the Clontech SMART II cDNA Synthesis System (Clontech Laboratories, Inc., Palo Alto, California, USA).

The modified SMART II system uses a modified oligo(dT) primer (5'-AAG CAG TGG TAA CAA CGC AGG CTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TVN-3'; where V denotes adenine, cytosine, or guanine and N denotes adenine, cytosine, guanine, or thymine) to initiate conventional first strand cDNA synthesis from the poly(A)⁺ tail of mRNA. The modified primer ensures the integration of a PCR priming site at the 3'-end of each cDNA. Clontech's SMART II system completes cDNA synthesis at the 5'-end of the mRNA by adding a short tract of cyto-

sines. The ability to add cytosines is a function of Invitrogen's SuperScript II (Karlsruhe, Germany) reverse transcriptase. This tract can hybridise with an oligonucleotide containing a short tract of guanines. Following hybridisation with this nucleotide the SuperScript II reverse transcriptase switches strands and extends the cDNA using the hybridised oligonucleotide sequence as a template. This process incorporates a second priming sequence at the 5'-end of each cDNA molecule. The technology is frequently described as 'strand switching technology' and is now widely used to capture small amounts of mRNA.

To ensure that the final SAGE template is biotinylated at the poly(A)⁺ end only, cDNA synthesis at the 5'-end of the mRNA is completed by the incorporation of a modified SMART II oligonucleotide containing a *Sap* I site (5'-AAG CAG TGG TAA CAA CGC AGG CTC TTC GGG-3'; *Sap* I site underlined) not present in the 3'-cDNA synthesis primer. *Sap* I is a restriction enzyme that recognises a seven-base sequence and, as such, is expected to cut DNA on average only once every 16,384 base pairs.

Second strand cDNA synthesis occurs during the first cycle of PCR. PCR-SAGE primer design means that both cDNA strands have identical 3'-termini, allowing amplification with a single primer (5'-AAG CAG TGG TAA CAA CGC AGG CT-3'). PCR amplification maintains a *Sap* I site at the 5'-end of the cDNA but does not introduce this restriction enzyme site at the 3'-end corresponding to the poly(A)⁺ tail. Using a 5'-biotinylated primer, PCR generates cDNA with biotin tags at both ends. Digestion of PCR-amplified cDNA with *Sap* I removes the biotin tag and a small amount of attached sequence only from the 5'-end of the mRNA. The cleaved 5'-ends are small (31 base pairs in length) and removed by size fractionation to yield cDNAs biotinylated solely at the end corresponding to the poly(A)⁺ tail of the original mRNA.

4.4.6.3 Principle of the SAGE-Lite procedure

SAGE-Lite (Peters *et al.*, 1999) is based on the same principle as PCR-SAGE but utilises slightly different primer sequences for cDNA synthesis (5'-AAG CAG TGG TAA CAA CGC AGA GTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TVN-3'; 5'-AAG CAG TGG TAA CAA CGC AGA GTA CGC GGG-3'). Biotinylation of cDNA at the end corresponding to the poly(A)⁺ tail is achieved by performing PCR with the same primers used to synthesise cDNA. In this case, only the oligo(dT) primer is biotinylated. The result is an amplified SAGE template that is biotinylated at one end only.

cDNA amplified by PCR-SAGE and SAGE-Lite produces two to five micrograms of material, enough for the construction of a SAGE catalogue (Peters *et al.*, 1999, Neilson *et al.*, 2000) using any of the standard protocols now available (see previous chapters related to SAGE).

4.4.6.4 PCR amplification and SAGE

The main criticism aimed at PCR-SAGE and SAGE-Lite is based on possible distortion of transcript representation within the transcriptome due to use of the PCR amplification. The kinetics of PCR amplification contains two phases. In the initial phase of amplification, the quantity of DNA produced increases exponentially. As the process continues through successive cycles of PCR amplification, the amount of material generated reaches a plateau where DNA concentration does not continue to increase (phase two). This occurs as a result of reaction components, such as primer, becoming limiting. During phase two, the limitation of reaction components leads to competition between molecules at each amplification cycle. Since individual cDNA species are not amplified with equal probabilities and efficiencies, the result is a product mix not representative of the original mRNA/cDNA sample.

During the exponential phase of DNA amplification, by contrast, reaction components present in excess and optimum enzyme activity ensure that each DNA species is copied *pro rata*, thus maintaining the ratios of individual DNA species to others in the mixture. As long as DNA for use in SAGE is collected while PCR is in the exponential phase of amplification, the relative proportions of each species within the mixture is maintained. These principles of PCR kinetics form the basis of quantitative PCR and have been shown to be valid and robust in many studies.

Peters *et al.* (1999) tested their amplification protocol for PCR generated distortions of the transcriptome using RNA extracted from human fibrosarcoma HT-1080 cells in culture. The RNA sample was pooled so that one aliquot was converted to cDNA without being amplified. A 200 nanograms sample of total RNA from the same pool was also converted to cDNA and then amplified by PCR. Samples were dot blotted onto nylon membrane and hybridised with radioactive probes for five different transcripts. With 22 cycles of amplification, the ratios of amplified to non-amplified signal for all five transcripts were equivalent, indicating that the relative proportions of each of the transcripts in the two samples had been maintained despite the PCR amplification process. With increased cycle number, this relationship was not maintained, indicating that over-amplification of the template distorts cDNA species representation.

A similar control was not undertaken for the human germinal vesicle oocyte catalogue. However, analysis of its PCR-SAGE catalogue suggests that PCR had not greatly distorted the relative abundance of cDNA species in the original sample. For example, the rank order of abundance for cytoskeletal proteins is generally consistent with accepted expression levels. Moreover, although quantitative gene expression data from the human oocyte is limited, it was possible to compare some measurements with those in the mouse. Rothstein *et al.* (1992) estimated that actin makes up 0.11 percent of all transcripts in the mouse oocyte, an estimate that was derived from a hybridisation study of a mouse oocyte cDNA library that was not amplified by PCR prior to construction. By contrast, the level of actin gene expression in the human germinal vesicle oocyte from PCR-SAGE was 0.073 percent of transcripts. However, if total transcript population in the mouse

egg is adjusted to compensate for the differences in cortical area (the site of most cytoskeletal actin), it equates to a comparable level of about 0.085 percent. The authors also generated a comprehensive gene expression profile from mouse oocyte cDNA libraries in which PCR was not used to amplify the initial template (Stanton & Green, 2001). In this study, the authors found that the level of vimentin expression in the mouse oocyte library is 0.027 percent compared to 0.030 percent of the total transcriptome respectively in the human oocyte library.

Other data also point to a low impact of PCR distortion. For example, tags for transcripts of *zona pellucida* protein are present in the PCR-SAGE catalogue at low levels. However, these proteins are expressed at their highest level in the oocyte during early stages of egg development and, in the germinal vesicle oocyte, production of the *zona pellucida* is complete and transcripts for these proteins are down regulated (Epifano *et al.*, 1995). Therefore, this is consistent with low SAGE tag numbers for *zona pellucida* transcripts in the germinal vesicle oocyte catalogue. As another example, the *ataxia telangiectasia* gene is known to be highly expressed in oocytes (Barlow *et al.*, 1998). This is borne out by SAGE tags for this gene at the high level of 0.105 percent of total transcripts in the human germinal vesicle catalogue.

The final qualitative evidence for the complexity of the human germinal vesicle oocyte PCR-SAGE catalogue is demonstrated by inspecting whole classes of RNA transcripts. As indicated earlier, inspection of the SAGE catalogue for cytoskeletal genes shows they are distributed in broadly the same way as other cell types. In addition, there are approximately 40 ribosomal proteins in the catalogue. This is comparable to other, non-amplified SAGE catalogues (Welle *et al.*, 1999). It seems unlikely that random PCR induced distortion of the initial sample would maintain such complex patterns of transcript representation. The maintenance of transcriptome complexity following PCR amplification of small RNA samples has also been demonstrated recently by Peters *et al.* (2001) through comparison of SAGE-Lite catalogues constructed from cranial artery biopsies.

4.4.6.5 Conclusion

The impossibility of acquiring large quantities of RNA from many biological samples has lead to the independent development of two amplifying techniques for SAGE. These are SAGE-Lite (Peters *et al.*, 1999) and PCR-SAGE (Neilson *et al.*, 2000). Both procedures rely on PCR to amplify cDNA synthesised from less than 50 nanograms of total RNA and both utilise strand switching by reverse transcriptase to introduce universal priming sites at each end of the cDNA. SAGE-Lite and PCR-SAGE differ only in the mechanisms used to ensure unilateral biotinylation of cDNA. Although it is possible to eliminate cDNA species representation by PCR of SAGE templates, this occurs only if the sample is over-amplified. Relative transcript concentrations maintained, the number of PCR cycles is limited.

With the advent of new technologies, such as laser capture microdissection, for harvesting individual cells from tissue, the demand for transcriptome catalogues

from small quantities of RNA will increase. SAGE-Lite and PCR-SAGE offer robust solutions to this problem.

4.4.7

SADE – SAGE adaption for downsized extracts

4.4.7.1 Introduction

The method of serial analysis of gene expression (SAGE) consists in analysing short diagnostic cDNA tags to establish quantitative gene expression profiles in tissues or cells (Velculescu *et al.*, 1995). SAGE rests on two principles: (i) a ten-nucleotide sequence tag isolated from a defined region of a transcript contains sufficient information for its unequivocal identification; indeed, there are 4^{10} (1,048,576) combinations for ten-nucleotides long sequences, whereas the human genome contains between 30,000 and 40,000 genes; (ii) concatenation of several tags (*e.g.*, ten to fifty) into a single DNA molecule, characterised by sequencing, greatly increases the throughput of data acquisition. In a series of elegant studies, Velculescu *et al.* (1995 and 1997) corroborated the validity of these principles, and demonstrated that tags abundance in a SAGE library is proportional to the expression level of their corresponding mRNA, thus establishing the quantitative feature of SAGE data.

Considering that SAGE only allows the analysis of fairly large amounts (Velculescu *et al.*, 1995) of biological material (five micrograms of poly(A)⁺ mRNA, corresponding to about 25×10^6 mammalian cells), a much more sensitive assay was set up. SADE, a SAGE adaptation for downsized extracts (Virlon *et al.*, 1999), makes it possible to generate a library of cDNA tags from 30,000 cells (Cheval *et al.*, 2000; Virlon *et al.*, 1999; Zakin *et al.*, 2000).

4.4.7.2 Principles and basics

The principle of the SADE procedure (see figure 4.83) is similar to the original SAGE protocol described in another section of this book. However, it is essentially a single-tube assay from obtaining of tissue lysate to cDNA tag recovery.

The SADE procedure includes the following modifications compared to the original SAGE protocol: (i) A single-step method for mRNA extraction from tissue lysate; (ii) use of a reverse transcriptase lacking ribonuclease H activity; (iii) use of a different anchoring enzyme; (iv) modification of procedure for blunt-ending cDNA tags; and (v) design of modified primers for PCR amplification.

In the first step of the SADE procedure, poly(A)⁺ mRNAs are isolated directly from the tissue lysate through binding to oligo(dT) primers covalently bound to magnetic beads, and the cDNA is synthesised immediately. Following second strand synthesis, the double-stranded cDNA fraction still present on beads is digested with the anchoring enzyme (*e.g.*, *Sau* 3AI). The remaining 3'-ends of bound cDNA are divided into two fractions. Each fraction is ligated to either linker primer A or linker primer B and then digested with *Bsm* FI. The released cDNA tags are blunt-ended and the two fractions are then ligated to each other. Following

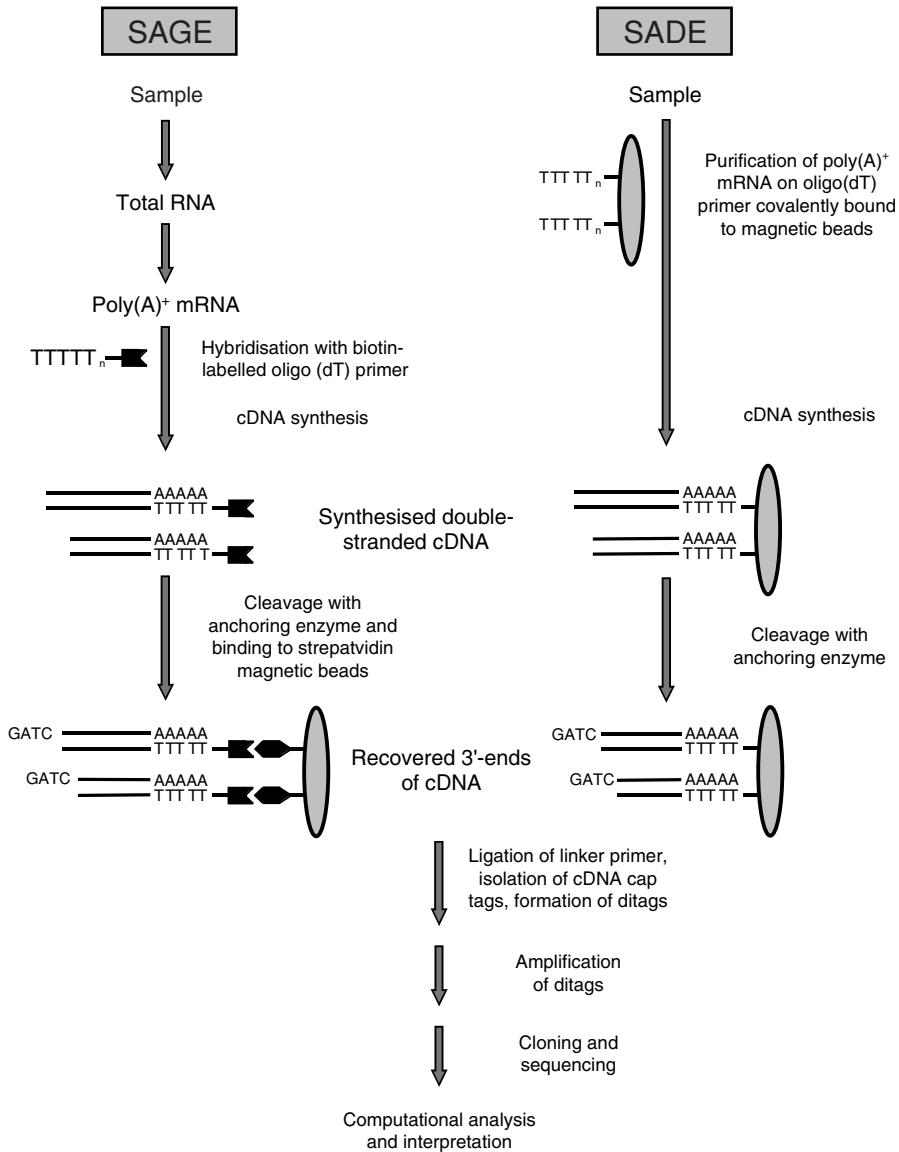


Figure 4.83 Schematic representation of the SAGE adaption for downsized extracts (SADE) procedure compared to the original serial analysis of gene expression (SAGE) procedure. See text for details.

PCR amplification of the resulting ditag, the sample is digested with *Sau* 3AI, and ditags are purified and concatenated as described in the original SAGE method. Concatemers of more than 350 base pairs are recovered, cloned into plasmids and sequenced (Virlon *et al.*, 1995).

4.4.7.3 Typical results

The major advantage of the SADE method is its increased efficiency that results in a lower amount of starting material required. At the stage of synthesis of double-stranded cDNA, the SADE procedure increases the amount of cDNA four-fold, as compared with the original SAGE method, suggesting improved yield for mRNA recovery. At the stage of anchoring enzyme digested cDNA, a 400-fold difference is observed for SADE and SAGE yields.

The SADE method has been successfully applied to three independent experimental set-ups. In the original publication in which SADE was established, the method was used for quantitative mRNA profiling in the mouse kidney (Virlon *et al.*, 1999). Compared to the original SAGE protocol (Velculescu *et al.*, 1995), SADE enabled a 1,000-fold reduction of the amount of starting material (about 50,000 cells). The potential of the SADE approach was evaluated by generating SADE tag libraries from microdissected nephron segments. The medullary thick ascending limb of Henle's loop (MTAL), which is nearly homogeneous at the cell level, and the outer medullary collecting duct (OMCD), which consists of three cell types (principal, α -, and β -intercalated cells in a 6:3:1 ratio) were studied in detail. Libraries were obtained from 150 millimetres of kidney tubules (about 50,000 cells), and 7,500 tags were sequenced in each case.

Compared to a SAGE tag library derived from whole kidney tissue, in the MTAL library, the amount of mitochondrial tags was even higher than in the kidney library. Thus, all nine most abundant tags corresponded to transcripts encoded by the mitochondrial genome. Turning to tags for nuclear transcripts, three tags corresponded to uromodulin which is synthesised specifically in the thick ascending limb of Henle's loop (Hession *et al.*, 1987) and known as the major protein in normal urine. Several abundant tags corresponded to genes that allow, either directly or indirectly, a high rate of sodium chloride re-absorption in the MTAL. The former class included transporters (*e.g.*, several Na/K-ATPase subunits and the kidney chloride channel ClC-K1) and the latter enzymatic activities that are linked to ATP production (ATP synthase and cytochrome c oxidase subunits, creatine kinase B, adenine nucleotide translocase-2). This gene expression pattern agrees with the fact that MTAL cells display a very high rate of sodium chloride re-absorption and have especially high Na/K-ATPase activity (Katz *et al.*, 1979).

In the OMCD library, mitochondrial tags were also abundant, with seven of the eight most abundant tags corresponding to mitochondrial transcripts. Among the most abundant tags for nuclear transcripts, some matched mRNAs encoding proteins essential for collecting duct functions, *i.e.*, regulation of water re-absorption by vasopressin and of sodium transport by aldosterone: the vasopressin-sensitive water channel aquaporin-2, type 2 11β -hydroxysteroid dehydrogenase (responsible for aldosterone-dependent regulation of sodium transport by transforming glucocorticoids into inactive compounds). The kidney-specific cadherin (a basolateral membrane protein in a subpopulation of collecting duct cells), and the band 3 anion exchanger (a marker for collecting duct α -intercalated cells) were also much more abundant than in the whole kidney library (Virlon *et al.*, 1999).

In a further experiment, SADE was used to compare the gene expression profiles in wild-type and *Otx2*(-/-) 6.5 days *post coitum* embryos (Zakin *et al.*, 2000).¹⁾ About 30,000 cells were processed for each library and a total of 27,100 and 21,443 tags were sequenced from the wild-type and *Otx2*(-/-) library, corresponding to 11,256 and 8,893 different transcripts, respectively. In total, 141 tags corresponding to differentially expressed genes were identified in the two libraries. Of these, 55 matched to known cDNAs, 27 corresponded to ESTs, and 59 were totally unknown. Most of the 141 tags corresponding to known genes were expressed at high levels and were involved in basic cell functions not specifically related to development.

Whole mount *in situ* hybridisation experiments were performed on wild-type and mutant embryos at 6.5 days *post coitum* using probes directed to differentially expressed mRNAs identified by SADE in the wild-type and *Otx2*(-/-) library (among others the protease inhibitor cystatin B, the cysteine-rich secreted glycoprotein *Wnt4*, and embryonic ectoderm development factor, *eed*). These experiments confirmed the data generated by SADE and revealed that *Otx2*(-/-) may be involved in the regulation of expression of these genes. Taken together, these studies suggest that the abnormalities presented by the *Otx2* mutant embryos are probably because of the defective migratory properties of the visceral endoderm tissue as a whole. This may result specifically in the wrong localisation of the cells fated to form the anterior visceral endoderm, leading to an ineffective head organiser.

In a third approach, SADE was used to analyse the transcriptome of a highly differentiated mouse clonal kidney cortical collecting duct principal cell line (mpkCCD(cl4)), and the changes in the transcriptome induced by aldosterone and vasopressin (Robert-Nicoud *et al.*, 2001). SADE analysis was performed on untreated cells and on cells treated with either aldosterone or vasopressin for four hours. In total, 169,721 transcript tags (39,091 tags, 62,796 tags, and 67,834 tags, respectively) were sequenced from the corresponding three SADE libraries. Even considering only tags that occurred twice or more, this analysis revealed that the 169,721 tags corresponded to 14,654 different tags of which 3,642 did not match to known mouse sequences.

A comparison of the three SAGE libraries revealed 34 aldosterone-induced transcripts, 29 aldosterone-repressed transcripts, 48 vasopressin-induced transcripts and eleven vasopressin-repressed transcripts. A subset of the differentially-expressed transcripts was studied by Northern blot analysis or semi-quantitative reverse transcription-based PCR to further validate the SADE data. These analyses confirmed the modified expression of mRNAs corresponding to the selected tags. These studies identified new sets of aldosterone and vasopressin-regulated genes.

Data generated using SADE are available on the internet at <http://www-dsv.cea.fr/thema/get/sade.html>. This website outlines the principle of SADE, as well as the differences between SAGE and SADE. For each published library (Robert-Nicoud *et al.*, 2001; Virlon *et al.*, 1999; Zakin *et al.*, 2000), the complete list of tags detected twice or more is provided. When tags match a registered cDNA at the

1) *Otx2* is a homeobox transcription factor.

appropriate position, a link ('annotation') is available. This makes possible, for all annotated tags, to immediately obtain gene names and GenBank accession numbers, which themselves provide links to PubMed.

4.4.7.4 Summary and discussion

Compared to the original SAGE procedure, SADE includes the following modifications: (i) a single-step method for mRNA extraction from tissue lysate; (ii) use of a reverse transcriptase lacking ribonuclease H activity; (iii) use of a different anchoring enzyme (see above); (iv) modified procedure for generation of blunt-ended cDNA tags; and (v) design of optimised primers for PCR amplification.

However, each of these modifications is not of equal importance in generating a library from a small amount of cells. For example, Datson *et al.* (1999) generated a SAGE library from about 10^5 cells by changing only the mRNA extraction method. In this case, streptavidin-coated tubes were used to recover cDNAs primed with a biotinylated oligo(dT) primer. However, this procedure is barely used, suggesting it is not robust or sensitive enough for certain requirements of a library although the studies performed by Virlon *et al.* (1999) showed that their modified SAGE procedure increases the efficiency four-fold at the stage of synthesis of double-stranded cDNA and about 400-fold at the stage of anchoring enzyme digested cDNA.

The modifications integrated in SADE allow the construction of tag libraries from a starting material of 50,000 cells, but was also successfully performed using 15,000 cells (Virlon *et al.*, 1999). In addition, the SADE method is flexible enough to also allow the analysis of macro-amounts of cells (Robert-Nicoud *et al.*, 2001).

4.4.8

TALEST – Tandem arrayed ligation of expressed sequence tags

4.4.8.1 Introduction

The completion of a draft human genome sequence by public and private consortia has been of enormous value to the biological and pharmaceutical sciences and has ignited much interest in 'functional genomics', *i.e.* the assignment of cellular function to each of the about 30,000 to 40,000 expressed genes estimated to reside in the genome. Because only a fraction of the full genetic repertoire is expressed in any given cell at any given time, and because the degree of gene expression can dramatically influence cellular phenotype, the first step in the task is the analysis of the transcriptome – the assessment of global gene expression at both qualitative and quantitative levels.

A number of methods have been developed to assess and quantify global gene expression. The technique currently most *in vogue* is hybridisation of labelled RNA or cDNA to high-density microarrays of oligonucleotides or cDNAs (so-called 'gene chips'). The availability of a draft sequence of the human genome coupled with technology to produce and analyse increasingly dense arrays allows researchers to quickly evaluate much of the human transcriptome in any cell or tissue. Nevertheless, technical hurdles remain and of course the method is applicable

only to those genes whose sequences are already available (and are printed *a priori* on the array), precluding opportunities for new gene discovery and limiting the utility of the method in species whose genomes are less characterised.

An alternative approach to generating quantitative gene expression profiles is the sequencing of random isolates from cDNA libraries to generate expressed sequence tags (ESTs) (Adams *et al.*, 1991). Within appropriately constructed libraries, frequency distributions of cDNA clones are largely proportional to steady-state transcript levels in the RNA population from which the library was derived, rendering the technique reasonably quantitative (Lee *et al.*, 1995; Okubo *et al.*, 1992). See also chapter 7 for a more detailed description of this technique. A variation of the EST sequencing approach called serial analysis of gene expression (SAGE) was described by Velculescu *et al.* (1995) and is now widely used. The SAGE technique involves the use of type II restriction endonucleases to generate short but positionally defined sequences from cDNAs which are randomly ligated in a tail-to-tail fashion and amplified by PCR to form so-called 'ditags' – essentially pairs of short ESTs. These ditags are concatenated into arrays which are cloned and analysed by DNA sequencing. Because each sequencing template contains multiple ESTs, throughput is significantly enhanced allowing the generation of an expression profile from a few thousand templates.

A drawback to the SAGE technique is its reliance on PCR amplification to generate ditags which compromises the quantitative aspects of the method. Owing to the exponential amplification of templates by PCR, even very minor variations in amplification efficiency of template sequences can give rise to dramatic differences in quantity of PCR product. As it was demonstrated previously (Spinella *et al.*, 1999), in the SAGE technique, the frequency of tag isolation is influenced not only by the starting frequency of the mRNA templates, but also by the intrinsic amplification efficiency of the individual tag sequences within each ditag. For most SAGE applications that compare profiles of two or more sample cell lines or tissues, this quantitative problem is of minor importance because the amplification biases introduced by PCR are the same for a given tag in each profile. The flaw becomes more problematic when attempting to ascertain relative levels of transcription among genes *within* any cell or tissue.

In order to circumvent this problem, a cDNA tag-based technique called TALEST for tandem arrayed ligation of expressed sequence tags was developed that does not rely on PCR amplification to generate tag arrays (Spinella *et al.*, 1999). The TALEST technique retains the advantages of SAGE with respect to throughput but avoids quantitative biases associated with PCR by relying on traditional bacterial amplification of clones rather than *in vitro* amplification by PCR. This section will describe the TALEST technique and discuss some of the considerations for data analysis inherent in TALEST, SAGE, or indeed any EST counting approach to transcriptional profiling.

4.4.8.2 Principles and basics

Figure 4.84 depicts a schematic diagram of the TALEST technique. Double-stranded cDNA is prepared from the target mRNA pool by standard methods using a biotinylated oligo(dT) primer. The cDNA is treated with *Eco* R1 methylase to protect internal *Eco* R1 sites and then digested with a four-base cutting restric-

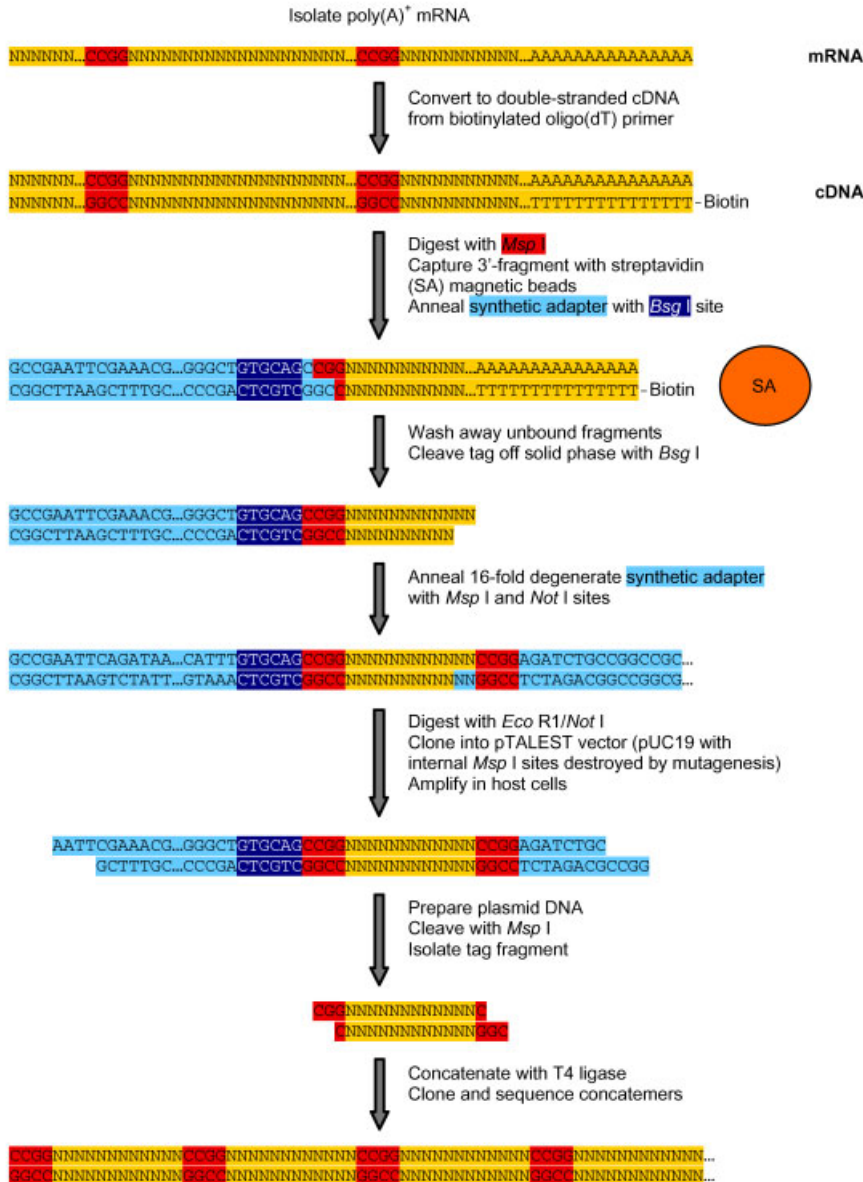


Figure 4.84 Schematic representation of the tandem arrayed ligation of expressed sequence tags (TALEST) process. See text for details.

tion enzyme (referred to as the punctuating enzyme or PE – typically *Msp*I). The 3′-most fragment is then isolated by affinity capture on streptavidin-coated magnetic particles and is extensively washed to remove upstream *Msp* I fragments. A double-stranded adapter molecule with a 5′-overhang compatible with *Msp* I is then ligated to the solid-phase cDNA. The adapter introduces a type II restriction site (*Bsg* I) immediately 5′ to the ligated cDNA and contains an *Eco* R1 site at its 5′-terminus to facilitate later cloning. Digestion of the adapter-ligated, solid-phase cDNA with *Bsg* I releases into the solution phase a linear DNA fragment consisting of the adapter itself and an additional ten to twelve nucleotides of unknown cDNA sequence separated from the adapter by the *Msp* I punctuation sequence as shown in figure 4.84. This fragment is then ligated to a second biotinylated adapter molecule containing a *Not* I site at its 5′-end and a 16-fold degenerate 3′-overhang which renders it compatible with all possible cDNA sequences released by *Bsg* I. This adapter introduces a second *Msp* I site to the 3′-end of the original DNA fragment such that all the molecules contain a twelve-base cDNA-derived tag sequence flanked at both ends by *Msp* I punctuation sites. The resulting molecule is double-digested with *Eco* R1/*Not* I and the biotinylated *Not* I fragment is removed by affinity capture on streptavidin-coated magnetic particles. The remaining *Eco* R1/*Not* I fragment is isolated by acrylamide gel electrophoresis to resolve it from dimers of the adapter. The result of these manipulations is an *Eco* R1/*Not* I-tailed DNA fragment containing a twelve base pairs cDNA tag flanked at both ends by the *Msp* I punctuation sequence. The fragment is eluted from the gel and recovered by ethanol precipitation in preparation for cloning into a pUC19 vector in which endogenous *Msp* I sites have been destroyed by site-directed mutagenesis (pTALEST vector).

The recombinant plasmids are transformed into competent *Escherichia coli* in order to generate a tag library. Plasmid DNA is prepared from the library and digested with *Msp* I to release the tags. Each tag is DNA fragment consisting of a twelve base-pair sequence derived from the cDNA flanked at both ends by GC clamp sequences which prevents the melting of tags at ambient temperatures and attendant bias against AT-rich sequences (confirmed by the normal distribution of GC-content in the tags around the expected mean of six, as well as the presence of numerous tags containing only adenines and thymines) in addition to single-stranded *Msp* I 5′-overhangs. The tag fragments are purified away from the plasmid backbone by polyacrylamide gel electrophoresis and ligated together to form concatemers. Concatemers of sufficient minimal length are isolated by agarose gel electrophoresis and cloned into standard pUC19 plasmid for DNA sequence analysis. Each array consists of 20 to 60 twelve-base tag sequences separated from each other and from the plasmid backbone by the defined four-base punctuating enzyme sequence, 5′-CCGG-3′. Because any individual tag sequence can ligate into an array in either sense or antisense orientation, tag sequences which are reverse complements of each other must be considered identical. This is accomplished *in silico* using software that compares all tag sequences with their reverse complements as they are found, writing only the alphabetically primary sequence to the tag database – regardless of the actual orientation in which the tag occurs in the array.

Both the TALEST and SAGE techniques rely on the ligation of a DNA adapter molecule containing a type II's restriction enzyme site which allows the generation of fragments containing several nucleotides of positionally defined cDNA-derived sequence (essentially small ESTs). The major difference is that TALEST does not utilise *in vitro* amplified ditags but, as with traditional EST approaches, relies on simple bacterial amplification of individual clones in a library to generate sufficient material for analysis.

4.4.8.3 Quantitative considerations

In order for an expressed gene to be 'tagged' by TALEST (or SAGE), the gene must contain the four-base punctuating restriction enzyme site. Since these occur, on average, once every 256 base pairs, the frequency of cDNAs which lack any given random four-base sequence is expected to be relatively low. In order to empirically estimate the frequency of 'untagged' genes, the entire RefSeq database of 14,038 full-length human mRNA transcripts was searched for the subset which lack an *Msp* I (5'-CCGG-3') site and found 1,234 such sequences (8.8 percent) – a coverage of 91.2 percent.

An under-appreciated problem with any profiling technique that relies upon restriction enzyme digestion of cDNA (TALEST, SAGE, AFLP, *etc.*) derives from the fact that the initial reverse transcriptase reaction often terminates prematurely and therefore does not always produce full-length copies of the starting mRNA template (particularly when the template is long). This implies that the greater the distance of the 3'-most punctuation enzyme site from the oligo(dT) priming site in the mRNA, the less the probability that it will be contained in any given cDNA copy. Hence, all other things being equal, tags located at sites far distal to the 3'-end of their gene would be expected to be somewhat under-represented relative to tags located more proximal to the 3'-end of their gene.

Another concern about the use of short tag sequences to identify genes is the risk that a given tag sequence will not be uniquely associated with a single gene. The theoretical problems inherent in tag-to-gene assignments have been discussed elsewhere in the context of SAGE (Lash *et al.*, 2000). In order to empirically assess the problem, the authors returned to the RefSeq database to provide a representative data sample. Of the 14,038 genes in the database that contain an *Msp* I restriction site, there were 356 cases (5.8 percent) in which two or more distinct genes would be expected to generate the same tag sequence. The vast majority of these represent cases in which a prospective TALEST tag fails to discriminate between alternative splice variants of the same gene or between individual members of highly related multi-gene families.

4.4.8.4 Examples and results

The TALEST technique was used to discover differentially expressed genes between a pair of cloned cell lines (designated 4A4 and 2C5, respectively) derived from a single polyclonal breast tumour but having different metastatic properties upon xe-

nograft. Tag libraries were prepared from both cell lines using about ten micrograms of mRNA from each. The primary TALEST libraries each contained more than one million clones. Plasmid DNA was prepared from about 13,000 independent colonies and subjected to automated DNA sequence analysis. A frequency distribution of tags from both cell lines was generated and searched against the RefSeq database to generate an expression profile. The 4A4 profile contained 211,455 unambiguous (no uncalled bases) tags representing 24,456 independent sequences ranging from the most abundant sequence (corresponding to mitochondrial cytochrome c oxidase) observed 6,214 times, to 10,880 singlets (*i.e.* tags observed only once). The 2C5 profile contained 211,647 tags representing 23,508 independent sequences. The most abundant tag (again, mitochondrial cytochrome c oxidase) was observed 4,272 times and there were 9,108 singlets. The number of unique tag sequences observed across both cell lines was 35,241 of which 1,952 were differentially expressed by at least five-fold. The purpose of the remainder of this chapter is not to discuss individual genes, their validation by other techniques, or their potential role in tumour metastasis (all of which will be discussed in forthcoming manuscripts). Rather, the biological and statistical interpretation of these data from a more generic perspective will be discussed.

4.4.8.5 From tags to transcripts

One of the most striking items in these data is the sheer number of unique tag sequences obtained – more than 35,000 from both cell lines. If one accepts that each tag sequence defines a unique transcript, one is forced to conclude that the vast majority of the entire human genome is actively expressed in these cells. This is of course unlikely and indeed, more than half of the unique tag sequences were seen only once suggesting that many represent artefacts of one sort or another that do not correspond to unique transcripts. Such artifactual tags can be generated in a variety of ways. Any partially digested cDNAs in which the 3'-most *Msp* I site was 'missed' by the restriction enzyme would produce a tag sequence flanking upstream *Msp* I sites. Heteronuclear or unspliced RNAs in the original mRNA pool could give rise to tags that are actually located in 3'-introns. Moreover, cDNA priming from sites other than the poly(A) tail of an mRNA could also result in artifactual tags that flank *Msp* I sites other than the 3'-most site in the transcript. Finally, since the actual sequences are single-pass and are not manually checked, sequencing errors will also give rise to spurious tags. A miscall rate of one percent per base would be expected to give rise to more than 4,000 spurious tags in a profile this large. While many of the singlet tags seen in the profiles undoubtedly do correspond to very rare mRNAs, the unavoidable presence of artefacts significantly inflates the apparent number of transcripts expressed. As a proportion of the total number of sequences found, such artefacts can be quite abundant. However, as a proportion of the total number of tags isolated, the number is not especially onerous. Even supposing that every one of the singlet tags represents an artefact sequence, they represent less than five percent of the total number of tags sequenced.

4.4.8.6 Discussion – Significance of differential expression

In considering whether any tag whose count differs between two samples represents a truly differentially expressed gene, one must make both biological and statistical decisions. From a biological perspective, one must decide *a priori* what level of differential gene expression is likely to have a phenotypic consequence. This is true regardless of what method is employed to assess expression, and there are no hard and fast rules governing the process. There are some genes in which a relatively minor change in steady state transcript level can give rise to a major phenotypic effect on cells. On the other hand, cells from carriers of many recessive mutations typically produce half the amount of wild-type transcript as do normal homozygotes suggesting that for these genes, a two-fold difference at least in steady-state transcription level has a rather minor effect on cellular phenotype. If one is choosing among a large set of differentially expressed genes and trying to ascertain which of them play a causal role in some phenotypic difference, one typically sets an arbitrary difference threshold as a condition for follow-up. As with many biological experiments, the actual decision point requires balancing the effort required to analyse and eliminate false-positives with avoidance of false-negatives. In the end, the decision is often based on the resources available to follow-up individual genes. Spinella and colleagues have arbitrarily set five-fold as a threshold for biological significance and in their profiles, 1,952 tags meet this criterion. This includes tags that were observed five times in one profile and not at all in the other. However, it excludes tags observed four or fewer times in one profile and zero in the other – even though these may in fact represent qualitatively differentially expressed genes.

In contrast to hybridisation-based approaches to profiling, methods like SAGE or TALEST that rely on sampling from an arbitrarily large pool of ESTs are susceptible to sampling error. The appropriate methodology for formal statistical analysis of SAGE (and by extension, TALEST) data has been a subject of some debate and the reader is referred to several seminal articles discussing the topic (Audic & Claverie, 1997; Chen *et al.*, 1998; Kal *et al.*, 1999; Lash *et al.*, 2000; Man *et al.*, 2000; Stollberg *et al.*, 2000). Spinella *et al.* use the Bayesian approach developed by Stephen F. Altschul and his colleagues (Lash *et al.*, 2000), which forms the basis for the differential expression data available through SAGEmap at the National Center for Biotechnology Information. Regardless of the statistical methodology employed, it is intuitively obvious that the greater the number of observations of a given tag sequence in one or both profiles, the greater the likelihood that observed differences in the sample reflect real population differences in the starting mRNA pools. This fact results in a number of somewhat under-recognised issues in the analysis of SAGE or TALEST data.

First, given sufficient number of observations, even subtle differences in relative expression levels (which may be biologically meaningless) can achieve statistical significance. For example, in this analysis the tag corresponding to mitochondrial cytochrome c oxidase was observed 6,214 times in the 4A4 profile and 4,272 times in the 2C5 profile. Regardless of the analytical method chosen, this frequency difference is highly statistically significant and likely reflects a real difference in the

two cell populations at the time they were harvested for mRNA extraction. However, it is important to note that transcription is not a static process and even within an individual cell, the level of any particular transcript varies over time around some steady state mean. The isolation of RNA from a population of cells essentially provides a snapshot in time of a dynamic transcriptome. Hence, this 1.4-fold difference seems unlikely to be biologically important regardless of the statistical significance. Conversely, tags that are observed relatively infrequently may not achieve statistical significance despite much larger apparent differences in relative expression level. In profiles of more than 400,000, even a tag that is observed as many as ten times in one population and only once in the other – a ten-fold apparent difference in expression level, does not achieve statistical significance at 95 percent confidence. Simple sampling error will frequently give rise differences of this magnitude among low abundance tags whose parent transcripts do not actually differ at all.

The Bayesian statistical analysis used to examine differential expression provides a probability estimate that, given observed tag sampling frequencies, the actual transcript population frequencies lie within a certain range. If one chooses a probability cut-off (say $P < 0.01$) and a minimum difference threshold (say five-fold), one can select that subset of tags for follow-up in which both biological and statistical criteria are met. In profiles examined by Spinella et al., 1,952 tags met the biological constraint of greater than five-fold difference. Of these however, only 88 also met the statistical criteria. What is perhaps more instructive is to recognise that, even in the (unusually large by SAGE standards) profile of more than 400,000 tags, a tag that is completely absent in one profile must be observed at least 14 times in the alternative profile in order to meet the $P < 0.01$ cutoff for significance. Of the about 35,000 tag sequences in both profiles (some of which were artefacts) fewer than 4,200 were observed at least 14 times. Even supposing that half of the total independent tag sequences represent artefacts, this still implies that significant gene expression differences can be detected only among transcripts in the top 25 percent of abundance. One can of course increase this proportion by setting less stringent selection criteria (with the attendant increase in false-positives) – or by performing still larger profiles. Nevertheless, it must be recognised that EST counting approaches in general are substantially biased toward detecting differences in the high abundance transcript pool while real differences in expression among transcripts in the mid to low abundance pool – the majority of all transcripts – get lost in the statistical noise.

4.5 Summary

The methods described in this chapter present the investigator with an embarrassment of riches. For that reason, particular care should be paid to the design of studies of gene expression and to the selection of the correct experimental approach. Many of the methods described are extremely powerful, but even the most power-

ful method will only yield a meaningful biological result within the context of an intelligent experiment. Thus, as noted elsewhere, the investigator must take special care to ensure an adequate number of cases and controls for each data point. Also, it should be borne in mind that all biological processes are dynamic in nature and that almost all are part of regulatory networks. Thus, particular attention should also be paid to deciding on measurement time points and to the consideration of counter-regulatory mechanisms which may influence the gene or genes under investigation.

In general, the methods described in this chapter refer to an experimental approach that was slow in gaining acceptance. Just a few years ago, the use of such techniques such as differential display RT-PCR to simultaneously screen for the regulation of hundreds of genes was likened in some quarters to going on a “fishing expedition” and was discussed as being of “low scientific worth”. This view has now changed. From a Popperian perspective, the performance of an experiment with no predefined hypothesis is a scientifically futile exercise. However, the experimental questions that can be addressed by the techniques described in this chapter do not lack a prior hypothesis: only the nature of the hypothesis is unfamiliar, but is a reflection of our current appreciation of the complex web of genetic interaction that exists in living organisms.

As noted in the introduction to this chapter, one of the most important considerations in choosing the appropriate gene analysis method is the issue of cost. Indeed, this limitation makes a rigorous experimental design all the more necessary.

However, even for the cash-strapped researcher light is on the horizon. It may be expected that at the current pace of development the cost of many high-throughput methods will fall sharply in the near future. In fact, the process of price reduction is already well advanced for DNA microarray technology, for example.

Second only in importance to adequate experimental design is the issue of competent data analysis. In order to extract meaningful information and ultimately valuable knowledge from high-throughput gene expression analyses, rigorous and professional bioinformatics is required. This topic is dealt with in detail in chapter 7.

4.6

References

Internet resources

- Affymetrix, Inc. (Santa Clara, California, USA); available at <http://www.affymetrix.com/>.
- Array Designer Software (TeleChem International, Inc./arrayit.com, Sunnyvale, California, USA); available <http://arrayit.com/>.
- Biofrontera Pharmaceuticals AG (Leverkusen, Germany); available at <http://www.biofrontera.com/>.
- BioSap, Blast Integrated Oligonucleotide Selection Accelerator Package (The Institute for Systems Biology, Seattle, Washington, USA); available at <http://biosap.sourceforge.net/>.
- British Biocell International, Ltd. (Cardiff, United Kingdom); <http://www.british-biocell.co.uk>.
- Capital Genomix, Inc. (Gaithersburg, Maryland, USA); available at <http://www.capital-genomix.com/>.
- CGAP (Cancer Genome Anatomy Project) at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/>.
- Chugai Biopharmaceuticals, Inc. (San Diego, California USA); available at <http://www.chugaibio.com/>.
- Clondia Chip Technologies, Inc. (Jena, Germany); <http://www.clondia.com/>.
- Compugen (Tel Aviv, Israel); available at <http://www.cgen.com/>.
- CuraGen Corporation (New Haven, Connecticut, USA); available at <http://www.curagen.com/>.
- dbEST (Database of Expressed Sequence Tags) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/dbEST/>.
- DEODAS, DEgenerate Oligo Nucleotide Design and Analysis System (The Institute for Systems Biology, Seattle, Washington, USA); available at <http://deodas.sourceforge.net/>.
- Digital Gene Technologies, Inc. (La Jolla, California, USA), available at <http://www.dgt.com/>.
- Featurama/ProbePicker (The Institute for Systems Biology, Seattle, Washington, USA); available at <http://propepicker.sourceforge.net/>.
- GenHunter Corporation (Nashville, Tennessee, USA); available at <http://www.genhunter.com/>.
- Gene Logic, Inc. (Gaithersburg, Maryland, USA); available at <http://www.genelogic.com/>.
- Genzyme Molecular Oncology (Framingham, Massachusetts, USA); available at <http://www.genzymemolecularoncology.com/>.
- German Human cDNA Project (a consortium of eight experienced DNA sequencing laboratories; detailed information is available on the homepage); available at <http://www2.mips.biochem.mpg.de/proj/cDNA/index.html>.
- IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium; available at <http://image.llnl.gov/>.
- Interactiva (Thermo Hybaid GmbH, Ulm, Germany); available at <http://www.interactiva.de>.
- Lynx Therapeutics, Inc. (Hayward, California, USA); available at <http://www.lynxgen.com/>.
- Mammalian Gene Collection (MGC) project (National Institutes of Health, NIH; Bethesda, Maryland, USA); available at <http://mgc.nci.nih.gov>.
- Nanogen corporation (San Diego, California, USA); available at <http://www.nanogen.com>.
- Nanoprobe, Inc. (Yaphank, New York, USA); available at <http://www.nanoprobe.com/>.
- OligoArray (Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan, USA); available at <http://berry.engin.umich.edu/oligoarray/>.
- Oligonucleotide Probe Sequence Design Software for Genetic Analysis (Olympus Optical Co. Europe, Hamburg, Germany); available at <http://www.olympus.co.jp/>.
- Oligos4Array (MWG-Biotech GmbH, Ebersberg, Germany); available at <http://www.mwg-biotech.com/>.
- R&D Systems, Inc. (Minneapolis, Minnesota, USA); available at <http://www.rndsystems.com/>.
- RefSeq (National Institute of Biotechnology Information, NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/LocusLink/>.
- RIKEN Mouse Gene Encyclopaedia Project (RIKEN Yokohama Institute, 1-7-22 Suehirocho, Tsurumi-ku, Yokohama City, Kanagawa, Japan); available at <http://www.gsc.riken.go.jp/>.

- SADE homepage (Département de Biologie Joliot-Curie, Commissariat à l'Énergie Atomique, CEA, Saclay, Gif-sur-Yvette, France); available at <http://www-dsv.cea.fr/thema/get/sade.html>.
- SAGEmap (serial analysis of gene expression tag to gene mapping database) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/SAGE/>.
- The MGuide version 2.0 (the Brown Lab's complete guide to microarraying for the molecular biologist (Patrick O. Brown's homepage, Department of Biochemistry, Howard Hughes Medical Institute, School of Medicine at Stanford University, Stanford, California, USA); accessible at <http://cmgm.stanford.edu/pbrown/mguide/index.html>.
- UniGene website at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/UniGene/>.
- Website of the Department of Molecular Biophysics and Biochemistry and the Department of Genetics (Yale University, School of Medicine, New Haven, Connecticut, USA) at <http://bioinfo.mbb.yale.edu/expression/neutrophil/>.
- Cited publications**
- Aach, J., Rindone, W., and Church, G. M. Systematic management and analysis of yeast gene expression data. *Genome Research* **2000**; *10*(4):431-445.
- Abdi, F., Bradbury, E. M., Doggett, N., and Chen, X. Rapid characterization of DNA oligomers and genotyping of single nucleotide polymorphism using nucleotide-specific mass tags. *Nucleic Acids Research* **2001**; *29*(13):E61.
- Aebersold, R., Hood, L. E., and Watts, J. D. Equipping scientists for the new biology. *Nature Biotechnology* **2000**; *18*(4):359.
- Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C., and Venter, J. C. Sequence identification of 2,375 human brain genes. *Nature* **1992**; *355*(6361):632-634.
- Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., and White, O. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **1995**; *377*(6547):3-17.
- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merrill, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science* **1991**; *252*(5013):1651-1656.
- Aittokallio, T., Ojala, P., Nevalainen, T. J., and Nevalainen, O. Analysis of similarity of electrophoretic patterns in mRNA differential display. *Electrophoresis* **2000**; *21*(14):2947-2956.
- Aittokallio, T., Ojala, P., Nevalainen T. J., and Nevalainen, O. Automated detection of differentially expressed fragments in mRNA differential display. *Electrophoresis* **2001**; *22*(10):1935-1945.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. In: Molecular biology of the cell. Garland Publishing, New York, New York, USA **1998**:528.
- Alivisatos, A. P., Johnsson, K. P., Peng, X., Wilson, T. E., Loweth, C. J., Bruchez, M. P., Jr., and Schultz, P. G. Organization of 'nanocrystal molecules' using DNA. *Nature* **1996**; *382*(6592):609-611.
- Allawi, H., Bartholomay, C. T., Chehak, L., Curtis, M. L., Eis, P. S., Hall, J. G., Ip, H. S., Kaiser, M., Kwiatkowski, R. W., Jr., Lukowiak, A. A., Lyamichev, V., and Ma, W. Detection of RNA. *WO190337* **2001**.
- Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the USA* **2000**; *97*(18):10101-10106.
- Alwine, J. C., Kemp, D. J., and Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the USA* **1977**; *74*(12):5350-5354.
- Amati, B., and Land, H. Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Current Opinion in Genetics and Development* **1994**; *4*(1):102-108.
- Angelastro, J. M., Klimaschewski, L. P., and Vitolo, O. V. Improved *Nla*III digestion of PAGE-purified 102 bp ditags by addition of

- single purification step in both the SAGE and microSAGE protocols. *Nucleic Acids Research* **2000**; 28(12):e62.
- Antequera, F., and Bird, A. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the USA* **1993**; 90(24):11995-11999.
- Aparicio, A. A. J. R. How to count... human genes. *Nature Genetics* **2000**; 25(2):129-130.
- Argani, P., Rosty, C., Reiter, R. E., Wilentz, R. E., Murugesan, S. R., Leach, S. D., Ryu, B., Skinner, H. G., Goggins, M., Jaffee, E. M., Yeo, C. J., Cameron, J. L., Kern, S. E., and Hruban, R. H. Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is over-expressed in pancreatic adenocarcinoma. *Cancer Research* **2001**; 61(11):4320-4324.
- Aristarkhov, A., Martin, C., and Palmer, M. A. J. High throughput and high sensitivity detection assays. *WO0060116* **2000**.
- Audic, S. and Claverie, J. M. The significance of digital gene expression profiles. *Genome Research* **1997**; 7(10):986-995.
- Authier, L., Grossiord, C., and Brossier, P. Gold nanoparticle-based quantitative electrochemical detection of amplified human cytomegalovirus DNA using disposable microband electrodes. *Analytical Chemistry* **2001**; 73(18):4450-4456.
- Bachem, C. W. B., van der Hoeven, R. S., de Bruijn, S. M., Vreugdenhil, D., Zabeau, M., and Visser, R. G. F. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: Analysis of gene expression during potato tuber development. *Plant Journal* **1996**; 9(5):745-753.
- Baldi, P., and Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **2001**; 17(6):509-519.
- Barié, N., Rapp, M., and Ache, H. J. UV cross-linked polysiloxanes as new coating materials for SAW devices with long term stability. *Sensors and Actuators B: Chemical* **1998**; 46(2):97-103.
- Barlow, C., Liyanage, M., Moens, P. B., Tarounas, M., Nagashima, K., Brown, K., Rottinghaus, S., Jackson, S. P., Tagle, D., Ried, T., and Wynshaw-Boris, A. *Atm* deficiency results in severe meiotic disruption as early as leptotema of prophase I. *Development* **1998**; 125(20):4007-4017.
- Barzen, C., Brecht, A., and Gauglitz, G. Optical multiple-analyte immunosensor for water pollution control. *Biosensors and Bioelectronics* **2002**; 17(4):289-295.
- Bassett, D. E., Jr., Eisen, M. B., and Boguski, M. S. Gene expression informatics – it's all in your mine. *Nature Genetics* **1999**; 21(Supplement 1):51-55.
- Basson, M. D., Liu, Y. W., Hanly, A. M., Emenaker, N. J., Shenoy, S. G., and Gould-Rothberg, B. E. Identification and comparative analysis of human colonocyte short-chain fatty acid response genes. *Journal of Gastrointestinal Surgery* **2000**; 4(5):501-512.
- Battaglia, C., Salani, G., Consolandi, C., Bernardi, L. R., and De Bellis, G. Analysis of DNA microarrays by non-destructive fluorescent staining using SYBR green II. *Biotechniques* **2000**; 29(1):78-81.
- Becker-Andre, M., and Hahlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Research* **1989**; 17(22):9437-9446.
- Beier, M., and Hoheisel, J. D. Versatile derivatization of solid support media for covalent bonding on DNA-microchips. *Nucleic Acids Research* **1999**; 27(9):1970-1979.
- Beier, M., and Hoheisel, J. D. Production by quantitative photolithographic synthesis of individually quality checked DNA microarrays. *Nucleic Acids Research* **2000**; 28(4):E11.
- Beissbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J., Hauser, N. C., Scheideler, M., Hoheisel, J. D., Schutz, G., Poustka, A., and Vingron, M. Processing and quality control of DNA array hybridization data. *Bioinformatics* **2000**; 16(11):1014-1022.
- Binnig, G., Rohrer, H., Gerber, C. and Weibel, E. Surface studies by scanning tunneling microscope. *Physical Review Letters* **1982**; 49(1):57-61.
- Binnig, G., Quate, C. F. and Gerber, C. Atomic force microscope. *Physical Review Letters* **1986**; 56(9):930-933.
- Birkert, O., Haake, H. M., Schutz, A., Mack, J., Brecht, A., Jung, G., and Gauglitz, G. A streptavidin surface on planar glass substrates for the detection of biomolecular interaction. *Analytical Biochemistry* **2000**; 282(2):200-208.

- Blanchard, A. P., Kaiser, R. J., and Hood, L. E. High-density oligonucleotide arrays. *Biosensors and Bioelectronics* **1996**; 11(6-7):687-690.
- Blanchard, A. P. In: Genetic engineering: principles and methods. Setlow, J. K., editor. Plenum Press, New York, New York, USA **1998**; 20:111-124.
- Blanchard, A. P., and Friend, S. H. Cheap DNA arrays – it's not all smoke and mirrors. *Nature Biotechnology* **1999**; 17(10):953.
- Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST – database for “expressed sequence tags”. *Nature Genetics* **1993**; 4(4): 332-333.
- Bonaldo, M. F., Lennon, G., and Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research* **1996**; 6(9):791-806.
- Boon, K., Caron, H. N., van Asperen, R., Valentijn, L., Hermus, M. C., van Sluis, P., Roobeek, I., Weis, I., Voute, P. A., Schwab, M., and Versteeg, R. *N-myc* enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis. *EMBO Journal* **2001**; 20(6):1383-1393.
- Borovikova, L. V., Ivanova, S., Zhang, M., Yang, H., Botchkina, G. I., Watkins, L. R., Wang, H., Abumrad, N., Eaton, J. W., and Tracey, K. J. Vagus nerve stimulation attenuates the systemic inflammatory response to endotoxin. *Nature* **2000**; 405(6785):458-462.
- Braun, A., Little, D. P., Reuter, D., Müller-Mysok, B., and Köster, H. Improved analysis of microsatellites using mass spectrometry. *Genomics* **1997**; 46(1):18-23.
- Bray, M. S., Boerwinkle, E., and Doris, P. A. High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise. *Human Mutations* **2001**; 17(4):296-304.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* **2001**; 29(4):365-371.
- Brennan, T. M. Method and apparatus for conducting an array of chemical reactions on a support surface. *US6210894* **2001**.
- Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M., DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. Gene expression analysis by massively parallel signature sequencing (MPSS™) on microbead arrays. *Nature Biotechnology* **2000a**; 18(6):630-634.
- Brenner, S., Williams, S., Vermass, E. H., Storck, T., Moon, K., McCollum, C., Mao, J., Luo, S., Kirchner, J. J., Eletr, S., DuBridge, R. B., Burcham, T., and Albrecht, G. *In vitro* cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. *Proceedings of the National Academy of Sciences of the USA* **2000b**; 97(4): 1665-1670.
- Brockman, J. M., Frutos, A. G., and Corn, R. M. A multistep chemical modification procedure to create DNA arrays on gold surfaces for the study of Protein-DNA interactions with surface plasmon resonance imaging. *Journal of the American Chemical Society* **1999**; 121(35):8044-8051.
- Brockman, J. M., Nelson, B. P., and Corn, R. M. Surface plasmon resonance imaging measurements of ultrathin organic films. *Annual Review of Physical Chemistry* **2000**; 51:41-63.
- Broude, N. E., Woodward, K., Cavallo, R., Cantor, C. R., and Englert, D. DNA microarrays with stem-loop DNA probes: preparation and applications. *Nucleic Acids Research* **2001**; 29(19):E92.
- Brown, T. Analysis of DNA sequences by blotting and hybridization. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1993a**:2.9.1-2.9.20.
- Brown, T. Hybridization analysis of DNA blots. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1993b**:2.10.1-2.10.16.
- Brown, T., and Mackey, K. Analysis of RNA by Northern and slot blot hybridization. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and

- Struhl, K., editors. John Wiley & Sons, New York, New York, USA 1997:4.9.1-4.9.16.
- Brown, C. S., Goodwin, P. C., and Sorger, P. K. Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the USA* 2001; 98(16):8944-8949.
- Bruce, W., Folkerts, O., Garnaat, C., Crasta, O., Roth, B., and Bowen, B. Expression profiling of the maize flavonoid pathway genes controlled by estradiol-inducible transcription factors CRC and P. *Plant Cell* 2000; 12(1): 65-80.
- Bruce, W., Desbons, P., Crasta, O., and Folkerts, O. Gene expression profiling of two related maize inbred lines with contrasting root-lodging traits. *Journal of Experimental Botany* 2001; 52(Roots special issue):459-468.
- Buess, M., Moroni, C., and Hirsch, H. H. Direct identification of differentially expressed genes by cycle sequencing and cycle labelling using the differential display PCR primers. *Nucleic Acids Research* 1997; 25(11):2233-2235.
- Buetow, K. H., Edmonson, M., MacDonald, R., Clifford, R., Yip, P., Kelley, J. Little, D. P., Strausberg, R., Koester, H., Cantor, C. R., and Braun, A. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* 2001; 98(2):581-584.
- Burgener, M., Sanger, M., and Candrian, U. Synthesis of a stable and specific surface plasmon resonance biosensor surface employing covalently immobilized peptide nucleic acids. *Bioconjugate Chemistry* 2000; 11(6):749-754.
- Butler, J. M., Jiang-Baucom, P., Huang, M., Belgrader, P., and Girard, J. Peptide nucleic acid characterisation by MALDI-TOF mass spectrometry. *Analytical Chemistry* 1996; 68(18):3283-3287.
- Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., Andrade, L. E., Carrer, H., El-Dorry, H. F., Espreafico, E. M., Habr-Gama, A., Giannella-Neto, D., Goldman, G. H., Gruber, A., Hackel, C., Kimura, E. T., Maciel, R. M., Marie, S. K., Martins, E. A., Nobrega, M. P., Paco-Larson, M. L., Pardini, M. I., Pereira, G. G., Pesquero, J. B., Rodrigues, V., Rogatto, S. R., da Silva, I. D., Sogayar, M. C., Sonati, M. F., Tajara, E. H., Valentini, S. R., Alberto, F. L., Amaral, M. E., Aneas, I., Arnaldi, L. A., de Assis, A. M., Bengtson, M. H., Bergamo, N. A., Bombonato, V., de Camargo, M. E., Canevari, R. A., Carraro, D. M., Cerutti, J. M., Correa, M. L., Correa, R. F., Costa, M. C., Curcio, C., Hokama, P. O., Ferreira, A. J., Furuzawa, G. K., Gushiken, T., Ho, P. L., Kimura, E., Krieger, J. E., Leite, L. C., Majumder, P., Marins, M., Marques, E. R., Melo, A. S., Melo, M., Mestriner, C. A., Miracca, E. C., Miranda, D. C., Nascimento, A. L., Nobrega, F. G., Ojopi, E. P., Pandolfi, J. R., Pessoa, L. G., Prevedel, A. C., Rahal, P., Rainho, C. A., Reis, E. M., Ribeiro, M. L., da Ros, N., de Sa, R. G., Sales, M. M., Sant'anna, S. C., dos Santos, M. L., da Silva, A. M., da Silva, N. P., Silva, W. A., Jr., da, Silveira, R. A., Sousa, J. F., Stecconi, D., Tsukumo, F., Valente, V., Soares, F., Moreira, E. S., Nunes, D. N., Correa, R. G., Zalcberg, H., Carvalho, A. F., Reis, L. F., Brentani, R. R., Simpson, A. J., and de Souza, S. J. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proceedings of the National Academy of Sciences of the USA* 2001; 98(21):12103-12108.
- Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., and Versteeg, R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001; 291(5507):1289-1292.
- Caruthers, M. H., Barone, A. D., Beaucage, S. L., Dodds, D. R., Fisher, E. F., McBride, L. J., Matteucci, M., Stabinsky, Z., and Tang, J. Y. Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. *Methods in Enzymology* 1987; 154:287-313.
- Chaudhari, N., and Hahn, W. E. Genetic expression in the developing brain. *Science* 1983; 220(4600):924-928.
- Chen, H., Centola, M., Altschul, S. F., and Metzger, H. Characterization of gene expression in resting and activated mast cells. *Journal of Experimental Medicine* 1998; 188(9):1657-1668.
- Chen, J. J., Rowley, J. D., and Wang, S. M. Generation of longer cDNA fragments from

- serial analysis of gene expression tags for gene identification. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(1):349-353.
- Chen, Z. J., Shen, H., and Tew, K. D. Gene expression profiling using a novel method: amplified differential gene expression (ADGE). *Nucleic Acids Research* **2001**; 29(10):e46.
- Cheng, J., Sheldon, E. L., Wu, L., Uribe, A., Gerrue, L. O., Carrino, J., Heller, M. J., and O'Connell, J. P. Preparation and hybridization analysis of DNA/RNA from *E. coli* on microfabricated bioelectronic chips. *Nature Biotechnology* **1998**; 16(6):541-546.
- Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., and Childs, G. Making and reading microarrays. *Nature Genetics* **1999**; 21(Supplement 1):15-19.
- Cheval, L., Virlon, B., and Elalouf, J.-M. SADE: a microassay for serial analysis of gene expression. In: Functional genomics. A practical approach. Hunt, S. P., and Livesey, J. P., editors. Oxford University Press, Oxford, United Kingdom **2000**:139-163.
- Chicurel, M. Faster, better, cheaper genotyping. *Nature* **2001**; 412(6847):580-582.
- Cho, Y., Meade, J., Walden, J., Guo, Z., and Liang, P. Multi-color fluorescent differential display. *Biotechniques* **2001**; 30(3):562-572.
- Chomczynski, P., and Sacchi, N. Single-step RNA isolation from cultured cells or tissues. In: Current protocols in molecular biology. Ausubel, F. A., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K., editors. John Wiley & Sons, New York, New York, USA **1997**:4.2.1-4.2.9.
- Chudin, E., Walker, R., Kosaka, A., Wu, S. X., Rabert, D., Chang, T. K., and Kreder, D. E. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biology* **2002**; 3(1):RESEARCH0005.1-RESEARCH0005.10.
- Clark, M. D., Panopoulou, G. D., Cahill, D. J., Bussow, K., and Lehrach, H. Construction and analysis of arrayed cDNA libraries. *Methods in Enzymology* **1999**; 303:205-233.
- Clark, M. D., Panopoulou, G. D., Cahill, D. J., Bussow, K., and Lehrach, H. Construction and analysis of arrayed cDNA libraries. *Methods in Enzymology* **1999**; 303:205-233.
- Clark, M. D., Hennig, S., Herwig, R., Clifton, S. W., Marra, M. A., Lehrach, H., Johnson, S. L., and the WU-GSC EST Group. An oligonucleotide fingerprint normalised and expressed sequence tag characterized zebrafish cDNA library. *Genome Research* **2001**; 11(9):1594-1602.
- Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **1999**; 8(10):1821-1832.
- Coille, I., Gauglitz, G., and Hoebeke, J. Characterisation of antibodies and analytes by surface plasmon resonance for the optimisation of a competitive immunoassay based on energy transfer. *Analytical and Bioanalytical Chemistry* **2002a**; 372(2):293-300.
- Coille, I., Reder, S., Bucher, S., and Gauglitz, G. Comparison of two fluorescence immunoassay methods for the detection of endocrine disrupting chemicals in water. *Biomolecular Engineering* **2002**; 18(6):273-280.
- Costello, C. E. Bioanalytic applications of mass spectrometry. *Current Opinion in Biotechnology* **1999**; 10(1):22-28.
- Craig, A. G., Nizetic, D., Hoheisel, J. D., Zehetner, G., and Lehrach, H. Ordering of cosmid clones covering the *Herpes simplex* virus type I (HSV-I) genome: a test case for fingerprinting by hybridisation. *Nucleic Acids Research* **1990**; 18(3):2653-2660.
- Crain, P. F., and McCloskey, J. A. Applications of mass spectrometry to the characterization of oligonucleotides and nucleic acids. *Current Opinion in Biotechnology* **1998**; 9(1):25-34.
- Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W., and Weissenbach, J. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genetics* **2000**; 25(2): 235-238.
- Crow, J. F., and Kimura, M. In: Introduction to population genetics theory. Harper & Row, New York, New York, USA **1970**.
- Dansch, G. Autometallography. A new technique for light and electron microscopic visualization of metals in biological tissues (gold, silver, metal sulphides and metal selenides). *Histochemistry* **1984**; 81(4):331-335.
- Datson, N. A., van der Perk-de Jong, J., van den Berg, M. P., de Kloet, E. R., and Vreugdenhil, E. MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Research* **1999**; 27(5):1300-1307.

- Datson, N. A., van der Perk, J., de Kloet, E. R., and Vreugdenhil, E. Expression profile of 30,000 genes in rat hippocampus using SAGE. *Hippocampus* **2001**; *11*(4):430-444.
- De Lecea, L., Kilduff, T. S., Peyron, C., Gao, X., Foye, P. E., Danielson, P. E., Fukuhara, C., Battenberg, E. L., Gautvik, V. T., Bartlett, F. S., 2nd, Frankel, W. N., van den Pol, A. N., Bloom, F. E., Gautvik, K. M., and Sutcliffe, J. G. The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proceedings of the National Academy of Sciences of the USA* **1998**; *95*(1):322-327.
- De Lumley-Woodyear, T. Campbell, C., and Heller, A. Direct Enzyme-amplified electrical recognition of a 30-base model oligonucleotide. *Journal of the American Chemical Society* **1996**; *118*(23):5504-5508.
- De Souza, S. J., Camargo, A. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., Andrade, L. E., Carrer, H., El-Dorry, H. F., Espreafico, E. M., Habr-Gama, A., Giannella-Neto, D., Goldman, G. H., Gruber, A., Hackel, C., Kimura, E. T., Maciel, R. M., Marie, S. K., Martins, E. A., Nobrega, M. P., Paco-Larson, M. L., Pardini, M. I., Pereira, G. G., Pesquero, J. B., Rodrigues, V., Rogatto, S. R., da Silva, I. D., Sogayar, M. C., de Fatima Sonati, M., Tajara, E. H., Valentini, S. R., Acencio, M., Alberto, F. L., Amaral, M. E., Aneas, I., Bengtson, M. H., Carraro, D. M., Carvalho, A. F., Carvalho, L. H., Cerutti, J. M., Correa, M. L., Costa, M. C., Curcio, C., Gushiken, T., Ho, P. L., Kimura, E., Leite, L. C., Maia, G., Majumder, P., Marins, M., Matsukuma, A., Melo, A. S., Mestriner, C. A., Miracca, E. C., Miranda, D. C., Nascimento, A. N., Nobrega, F. G., Ojopi, E. P., Pandolfi, J. R., Pessoa, L. G., Rahal, P., Rainho, C. A., da Ros, N., de Sa, R. G., Sales, M. M., da Silva, N. P., Silva, T. C., da Silva, W., Jr., Simao, D. F., Sousa, J. F., Stecconi, D., Tsukumo, F., Valente, V., Zalcberg, H., Brentani, R. R., Reis, F. L., Dias-Neto, E., and Simpson, A. J. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences of the USA* **2000**; *97*(23):12690-12693.
- Deforce, D. L., and van den Eeckhout, E. G. Analysis of oligonucleotides by ESI-MS. *Advances in Chromatography* **2000**; *40*:539-566.
- Delamarque, E., Bernard, A., Schmid, H., Michel, B., and Biebuyck, H. Patterned delivery of immunoglobulins to surfaces using microfluidic networks. *Science* **1997**; *276*(5313):779-781.
- DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**; *278*(5338):680-686.
- Dias Neto, E., Garcia Correa, R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr., Zago, M. A., Bordin, S., Costa, F. F., Goldman, G. H., Carvalho, A. F., Matsukuma, A., Baia, G. S., Simpson, D. H., Brunstein, A., de Oliveira, P. S., Bucher, P., Jongeneel, C. V., O'Hare, M. J., Soares, F., Brentani, R. R., Reis, L. F., de Souza, S. J., and Simpson, A. J. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences of the USA* **2000**; *97*(7):3491-3496.
- Diatchenko, L., Lau, Y., Campbell, A., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E., and Siebert, P. Suppression subtractive hybridisation: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences of the USA* **1996**; *93*(12):6025-6030.
- Dieterle, F., Nopper, D., and Gauglitz, G. Quantification of butanol and ethanol in aqueous phases by reflectometric interference spectroscopy – different approaches to multivariate calibration. *Fresenius' Journal of Analytical Chemistry* **2001**; *370*(6):723-730.
- Draghici, S., Kuklin, A., Hoff, B., and Shams, S. Experimental design, analysis of variance and slide quality assessment in gene expression arrays. *Current Opinion in Drug Discovery and Development* **2001**; *4*(3):332-337.
- Drmanac, R., Labat, I., Brukner, I., and Crkvenjakov, R. Sequencing of megabase plus DNA by hybridisation: theory of the method. *Genomics* **1989**; *4*(2):114-128.
- Drmanac, R., Strezoska, Z., Labat, I., Drmanac, S., and Crkvenjakov, R. Reliable hybridisation of oligonucleotides as short as six nucleotides. *DNA and Cell Biology* **1990**; *9*(7):527-534.
- Drmanac, S., Stavropoulos, N. A., Labat, I., Vonau, J., Hauser, B., Soares, M. B., and Drmanac, R. Gene-representing cDNA clusters defined by hybridisation of 57,419 clones from infant brain libraries with short oligo-

- nucleotide probes. *Genomics* **1996**; 37(1): 29-40.
- Drobyshev, A. L., Zasedatelev, A. S., Yershov, G. M., and Mirzabekov, A. D. Massive parallel analysis of DNA-Hoechst 33258 binding specificity with a generic oligodeoxyribonucleotide microchip. *Nucleic Acids Research* **1999**; 27(20):4100-4105.
- Dubiley, S., Kirillov, E., Lysov, Y., and Mirzabekov, A. Fractionation, phosphorylation and ligation on oligonucleotide microchips to enhance sequencing by hybridization. *Nucleic Acids Research* **1997**; 25(12):2259-2265.
- Dubiley, S., Kirillov, E., and Mirzabekov, A. Polymorphism analysis and gene detection by minisequencing on an array of gel-immobilized primers. *Nucleic Acids Research* **1999**; 27(18):e19.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. Expression profiling using cDNA microarrays. *Nature Genetics* **1999**; 21(Supplement 1):10-14.
- Dunham, I. The gene guessing game. *Yeast* **2000**; 17(3):218-224.
- Dunphy, J. C., Busch, K. L., and Buchanan, M. V. Rapid extraction and structural characterization of biomolecules in agarose gels by laser desorption Fourier transform mass spectrometry. *Analytical Chemistry* **1993**; 65(10):1329-1335.
- Dyson, N. J. Immobilization of nucleic acids and hybridization analysis. In: *Essential molecular biology: A practical approach*. Brown, T. A., editor. IRL Press at Oxford University Press, Oxford, United Kingdom **1991**; 2:111-156.
- Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the USA* **1992**; 89(7):3010-3014.
- Eckman, B. A., Aaronson, J. S., Borkowski, J. A., Bailey, W. J., Elliston, K. O., Williamson, A. R., and Blevins, R. A. The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining. *Bioinformatics* **1998**; 14(1):2-13.
- Edman, C. F., Raymond, D. E., Wu, D. J., Tu, E., Sosnowski, R. G., Butler, W. F., Nerenberg, M., and Heller, M. J. Electric field directed nucleic acid hybridization on microchips. *Nucleic Acids Research* **1997**; 25(24): 4907-4914.
- Egholm, M., Buchardt, O., Christensen, L., Behrens, C., Freier, S. M., Driver, D. A., Berg, R. H., Kim, S. K., Norden, B., and Nielsen, P. E. PNA hybridizes to complementary oligonucleotides obeying the Watson-Crick hydrogen-bonding rules. *Nature* **1993**; 365(6446):566-568.
- Eickhoff, B., Korn, B., Schick, M., Poustka, A., and van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Research* **1999**; 27(22):e33.
- Eickhoff, H., Schuchhardt, J., Ivanov, I., Meier-Ewert, S., O'Brien, J., Malik, A., Tandon, N., Wolski, E. W., Rohlf, E., Nyarsik, L., Reinhardt, R., Nietfeld, W., and Lehrach, H. Tissue gene expression analysis using arrayed normalised cDNA libraries. *Genome Research* **2000**; 10(8):1230-1240.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(25): 14863-14868.
- Epifano, O., Laing, L., Familiari, M., Moos, M. C., Jr., and Dean, J. Coordinate expression of the three zona pellucida genes during mouse oogenesis. *Development* **1995**; 121(7):1947-1956.
- Ermantraut, E., Schulz, T., Tuchscheerer, J., Wölfel, S., Saluz, H.-P., Thallner, E., and Köhler, J. M. Building highly diverse arrayed substance libraries by micro offset printing. In: *Proceedings of micro total analysis systems '98*. Harrison, D. J., and van den Berg, A., editors. Kluwer Scientific Publishing, Dordrecht, The Netherlands **1998**:213-216.
- Evertsz, E., Starink, P., Gupta, R., and Watson, D. Technology and applications of gene expression microarrays. In: *Microarray biochip technology*. BioTechniques Books. Schena, M., editor. Eaton Publishing Corporation, Natick, Massachusetts, USA **2000**:149-166.
- Evertsz, E. M., Au-Young, J., Ruvolo, M. V., Lim, A. C., and Reynolds, M. A. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Bio-techniques* **2001**; 31(5):1182-1192.
- Ewing, B., and Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics* **2000**; 25(2):232-234.
- Fei, Z., and Smith, L. M. Analysis of single nucleotide polymorphisms by primer exten-

- sion and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **2000**; 14(11):950-959.
- Fei, Z., Ono, T., and Smith, L. M. MALDI-TOF mass spectrometric typing of single nucleotide polymorphisms with mass-tagged ddNTPs. *Nucleic Acids Research* **1998**; 26(11):2827-2828.
- Felder, S., and Kris, R. M. High throughput assay system. *US6238869* **2001**.
- Feriotto, G., Ferlini, A., Ravani, A., Calzolari, E., Mischiati, C., Bianchi, N., and Gambari, R. Biosensor technology for real-time detection of the cystic fibrosis W1282X mutation in CFTR. *Human Mutation* **2001a**; 18(1):70-81.
- Feriotto, G., Corradini, R., Sforza, S., Bianchi, N., Mischiati, C., Marchelli, R., and Gambari, R. Peptide nucleic acids and biosensor technology for real-time detection of the cystic fibrosis W1282X mutation by surface plasmon resonance. *Laboratory Investigation* **2001b**; 81(10):1415-1427.
- Fields, C., Adams, M. D., White, O., and Venter, J. C. How many genes in the human genome? *Nature Genetics* **1994**; 7(3):345-346.
- Fischer, A., Saedler, H., and Theissen, G. Restriction fragment length polymorphism-coupled domain-directed differential display. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(12):5331-5335.
- Fitzgerald, M. C., Parr, G. R., and Smith, L. M. Basic matrices for the matrix-assisted laser desorption/ionization mass spectrometry of proteins and oligonucleotides. *Analytical Chemistry* **1993**; 65(22):3204-3211.
- Fitzgerald, M. C., and Smith, L. M. Mass spectrometry of nucleic acids: the promise of matrix-assisted laser desorption-ionization (MALDI) mass spectrometry. *Annual Review in Biophysics and Biomolecular Structure* **1995**; 24:117-140.
- Fotin, A. V., Drobyshv, A. L., Proudnikov, D. Y., Perov, A. N., and Mirzabekov, A. D. Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucleic Acids Research* **1998**; 26(6):1515-1521.
- Freeman, W. M., Robertson, D. J., and Vrana, K. E. Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques* **2000**; 29(5):1042-1046, 1048-1055.
- Frohman, M. A. Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: thermal RACE. *Methods in Enzymology* **1993**; 218:340-356.
- Frohman, M. A. On beyond classic RACE (rapid amplification of cDNA ends). *PCR Methods and Applications* **1994**; 4(1):S40-S58.
- Fu, D. J., Broude, N. E., Köster, H., Smith, C. L., and Cantor, C. R. A DNA sequencing strategy that requires only five bases of known terminal sequence for priming. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(22):10162-10166.
- Fu, D. J., Broude, N. E., Köster, H., Smith, C. L., and Cantor, C. R. Efficient preparation of short DNA sequence ladders potentially suitable for MALDI-TOF DNA sequencing. *Genetic Analysis* **1996**; 12(3-4):137-142.
- Fu, D. J., Köster, H., Smith, C. L., and Cantor, C. R. Sequencing double-stranded DNA by strand displacement. *Nucleic Acids Research* **1997**; 25(3):677-679.
- Fu, D. J., Tang, K., Braun, A., Reuter, D., Darnhofer-Demar, B., Little, D. P., O'Donnell, M. J., Cantor, C. R., and Köster, H. Sequencing exons 5 to 8 of the p53 gene by MALDI-TOF mass spectrometry. *Nature Biotechnology* **1998**; 16(4):381-384.
- Garcia, B. A., Heaney, P. J., and Tang, K. Improvement of the MALDI-TOF analysis of DNA with thin-layer matrix preparation. *Analytical Chemistry* **2002**; 74(9):2083-2091.
- Garner, H. R. Microarray robotics and instrumentation systems. In: *Proceedings of lab chips and microarrays for biotechnical applications*. Cambridge Healthtech Institute, CHI; Newton Upper Falls, Massachusetts, USA **1999**.
- Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C., Yates, J. R., 3rd. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Analytical Chemistry* **2000**; 72(4):757-763.
- Gauglitz, G., Brecht, A., Kraus, G., and Nahm, W. Chemical and biochemical sensors based on interferometry at thin (multi-)layers. *Sensors and Actuators B: Chemical* **1993**; 11(1):21-27.
- Gevaert, K., and Vandekerckhove, J. Protein identification methods in proteomics. *Electrophoresis* **2000**; 21(6):1145-1154.
- Gibbs, W. W. Shrinking to enormity: DNA microarrays are reshaping basic biology – but

- scientists fear they may soon drown in the data. *Scientific American* **2001**; 284(1):33-34.
- Gilles, P. N., Wu, D. J., Foster, C. B., Dillon, P. J., and Chanock, S. J. Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. *Nature Biotechnology* **1999**; 17(4):365-70.
- Gotoh, M., Hasegawa, Y., Shinohara, Y., Shimizu, M., and Tosu, M. A new approach to determine the effect of mismatches on kinetic parameters in DNA hybridization using an optical biosensor. *DNA Research* **1995**; 2(6):285-293.
- Green, C. D., Simons, J. F., Taillon, B. E., and Lewin, D. A. Open systems: panoramic views of gene expression. *Journal of Immunological Methods* **2001**; 250(1-2):67-79.
- Greene, G. L., and Letsinger, R. L. Formation of internucleotide 3'-5' phosphoramidate links by direct coupling of phosphoryl and amino groups. *Nucleic Acids Research* **1975**; 2(7):1123-1127.
- Griffin, T. J., Hall, J. G., Prudent, J. R., and Smith, L. M. Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(11):6301-6306.
- Griffin, T. J., and Smith, L. M. Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends in Biotechnology* **2000**; 18(2):77-84.
- Griffiths, W. J. Nanospray mass spectrometry in protein and peptide chemistry. *EXS* **2000**; 88:69-79.
- Gubler, U., and Hoffman, B. J. A simple and very efficient method for generating cDNA libraries. *Gene* **1983**; 25(2-3):263-269.
- Guedon, P., Livache, T., Martin, F., Lesbre, F., Roget, A., Bidan, G., and Levy, Y. Characterization and optimization of a real-time, parallel, label-free, polypyrrole-based DNA sensor by surface plasmon resonance imaging. *Analytical Chemistry* **2000**; 72(24):6003-6009.
- Guerasimova, A., Nyarsik, L., Girnus, I., Steinfath, M., Wruck, W., Griffiths, H., Herwig, R., Wierling, C., O'Brien, J., Eickhoff, H., Lehrach, H., and Radelof, U. New tools for oligonucleotide fingerprinting. *Biotechniques* **2001**; 31(3):490-495.
- Guilhaus, M., Selby, D., and Mlynski, V. Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrometry Reviews* **2000**; 19(2):65-107.
- Gullberg, U., Bengtsson, N., Bulow, E., Garwicz, D., Lindmark, A., and Olsson, I. Processing and targeting of granule proteins in human neutrophils. *Journal of Immunological Methods* **1999**; 232(1-2):201-210.
- Guschin, D., Yershov, G., Zaslavsky, A., Gemmell, A., Shick, V., Proudnikov, D., Arenkov, P., and Mirzabekov, A. Manual manufacturing of oligonucleotide, DNA, and protein microchips. *Analytical Biochemistry* **1997**; 250(2):203-211.
- Hacia, J. G., Edgemon, K., Sun, B., Stern, D., Fodor, S. P., and Collins, F. S. Two color hybridization analysis using high density oligonucleotide arrays and energy transfer dyes. *Nucleic Acids Research* **1998**; 26(16):3865-3866.
- Hainfeld, J. F., and Powell, R. D. Nanogold technology: New frontiers in gold labeling. *Cell Vision* **1997**; 4:408-432.
- Hainfeld, J. F., and Powell, R. D. New frontiers in gold labeling. *Journal of Histochemistry and Cytochemistry* **2000**; 48(4):471-480.
- Harris, T. M., Massimi, A., and Childs, G. Injecting new ideas into microarray printing. *Nature Biotechnology* **2000**; 18(4):384-385.
- Hashimoto, K., Ito, K., and Ishimori, Y. Sequence-specific gene detection with a gold electrode modified with DNA probes and an electrochemically active dye. *Analytical Chemistry* **1994**; 66(21):3830-3833.
- Hashimoto, K., Ito, K., and Ishimori, Y. Novel DNA sensor for electrochemical gene detection. *Analytica Chimica Acta* **1994b**; 286(2):219-224.
- Hassan, A., Stengele, K. P., Giegrich, H., Cornwell, P., Isham, K. R., Sachleben, R., and Pfeleiderer, W. Photolabile protecting groups for nucleosides: Synthesis and photodeprotection rates. *Tetrahedron* **1997**; 53(12):4247-4264.
- Hastie, N. D., and Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **1976**; 9(4 Part 2):761-774.
- Hayat, M. A. Immunogold silver staining. 1st edition. Hayat, M. A., editor. CRC Press, Inc., London, United Kingdom **1995**.
- Heaton, R. J., Peterson, A. W., and Georgiadis, R. M. Electrostatic surface plasmon resonance: direct electric field-induced hybridization and denaturation in monolayer nucleic acid films and label-free discrimination of base mismatches. *Proceedings of the National*

- Academy of Sciences of the USA **2001**; 98(7):3701-3704.
- Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J. E., Snesrud, E., Lee, N., and Quackenbush, J. A concise guide to cDNA microarray analysis. *Biotechniques* **2000**; 29(3):548-562.
- Heller, M. J., and Tu, E. Active programmable electronic devices for biomolecules. *US5605662* **1993**.
- Herrler, M. Use of SMART-generated cDNA for differential gene expression studies. *Journal of Molecular Medicine* **2000**; 78(7):B23.
- Herrmann, J. L., Rastelli, L., Burgess, C. E., Fernandes, E. F., Rothberg, B. E., Rothberg, J. M., and Shimkets, R. A. Implications of oncogenomics for cancer research and clinical oncology. *Cancer Journal* **2001**; 7(1):40-51.
- Herwig, R., Poustka, A. J., Muller, C., Bull, C., Lehrach, H., and O'Brien, J. Large-scale clustering of cDNA-fingerprinting data. *Genome Research* **1999**; 9(11):1093-1105.
- Herwig, R., Schmitt, A. O., Steinfath, M., O'Brien, J., Seidel, H., Meier-Ewert, S., Lehrach, H., and Radelof, U. Information theoretical probe selection for hybridisation experiments. *Bioinformatics* **2000**; 16(10): 890-898.
- Herwig, R., Aanstad, P., Clark, M., and Lehrach, H. Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Research* **2001**; 29(23):E117.
- Hession, C., Decker, J. M., Sherblom, A. P., Kumar, S., Yue, C. C., Mattaliano, R. J., Tizard, R., Kawashima, E., Schmeissner, U., Heletky, S., Chow, E. P., Burne, C. A., Shaw, A., and Muchmore, A. V. Uromodulin (Tamm-Horsfall glycoprotein): a renal ligand for lymphokines. *Science* **1987**; 237(4821):1479-1484.
- Higgins, G. S., Little, D. P., and Köster, H. Competitive oligonucleotide single-base extension combined with mass spectrometric detection for mutation screening. *Biotechniques* **1997**; 23(4):710-714.
- Hill, A. A., Brown, E. L., Whitley, M. Z., Tucker-Kellogg, G., Hunter, C. P., and Slonim, D. K. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biology* **2001**; 2(12):RESEARCH0055.1- RE-SEARCH0055.13.
- Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chisoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., and Marra, M. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research* **1996**; 6(9):807-828.
- Hoffmann, R., Zwilling, S., and Luebbert, H. Neues Verfahren zur Identifikation und Charakterisierung von mRNA-Molekülen. *DE19806431* **1998**.
- Hoheisel, J. D., Lennon, G. G., Zehetner, G., and Lehrach, H. Use of high coverage reference libraries of *Drosophila melanogaster* for relational data analysis. A step towards mapping and sequencing of the genome. *Journal of Molecular Biology* **1991**; 220(4):903-914.
- Holmlin, R. E., Dandliker, P. J., and Barton, J. K. Charge transfer through the DNA base stack. *Angewandte Chemie International Edition English* **1997**; 36(24):2714-2718.
- Hu, G. K., Madore, S. J., Moldover, B., Jatke, T., Balaban, D., Thomas, J., and Wang, Y. Predicting splice variant from DNA chip expression data. *Genome Research* **2001**; 11(7):1237-1245.
- Huang, T. J., Liu, M., Knight, L. D., Grody, W. W., Miller, J. F., and Ho, C. M. An electrochemical detection scheme for identification of single nucleotide polymorphisms using hairpin-forming probes. *Nucleic Acids Research* **2002**; 30(12):e55.
- Hubank, M., and Schatz, D. Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Research* **1994**; 22(25):5640-5648.
- Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephanian, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **2001**; 19(4):342-347.
- Huth-Fehre, T., Gosine, J. N., Wu, K. J., and Becker, C. H. Matrix-assisted laser desorption mass spectrometry of oligodeoxythymidyl

- acids. *Rapid Communications in Mass Spectrometry* **1992**; 6(3):209-213.
- Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology* **2000**; 7(6):805-817.
- Irie, T., Oshida, T., Hasegawa, H., Matsuoka, Y., Li, T., Oya, Y., Tanaka, T., Tsujimoto, G., and Kambara, H. Automated DNA fragment collection by capillary array gel electrophoresis in search of differentially expressed genes. *Electrophoresis* **2000**; 21(2):367-374.
- Ito, T., Kito, K., Adati, N., Mitsui, Y., Hagiwara, H., and Sakaki, Y. Fluorescent differential display: arbitrarily primed RT-PCR fingerprinting on an automated DNA sequencer. *Federation of European Biochemical Societies Letters* **1994**; 351(2):231-236.
- Ito, T., Ikeda, U., Shimpō, M., Yamamoto, K., and Shimada, K. Serotonin increases interleukin-6 synthesis in human vascular smooth muscle cells. *Circulation* **2000**; 102(20):2522-2527.
- Ivanova, N. B., and Belyavsky, A. V. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research* **1995**; 23(15):2954-2958.
- Ivashuta, S., Uchiyama, K., Gau, M., and Shimamoto, Y. Linear amplification coupled with controlled extension as a means of probe amplification in a cDNA array and gene expression analysis during cold acclimation in alfalfa (*Medicago sativa* L.). *Journal of Experimental Botany* **2002**; 53(367):351-359.
- Jackson, P. E., Scholl, P. F., and Groopman, J. D. Mass spectrometry for genotyping: an emerging tool for molecular medicine. *Molecular Medicine Today* **2000**; 6(7):271-276.
- Janshöff, A., Galla, H.-J., and Steinem, C. Piezoelectric mass-sensing devices as biosensors – an alternative to optical biosensors? *Angewandte Chemie International Edition English* **2000**; 39(22):4004-4032.
- Jin-Lee, H., Goodrich, T. T., and Corn, R. M. SPR imaging measurements of 1-D and 2-D DNA microarrays created from microfluidic channels on gold thin films. *Analytical Chemistry* **2001**; 73(22):5525-5531.
- Johnston, R. F., Pickett, S. C., and Barker, D. L. Autoradiography using storage phosphor technology. *Electrophoresis* **1990**; 11(5):355-360.
- Johnston, D. H., Glasgow, K. and Thorp, H. H. Electrochemical measurement of the solvent accessibility of nucleobases using electron transfer between DNA and metal complexes. *Journal of the American Chemical Society* **1995**; 117(35):8933-8938.
- Johnston, M. V. Sampling and analysis of individual particles by aerosol mass spectrometry. *Journal of Mass Spectrometry* **2000**; 35(5):585-595.
- Jung, A., Stemmler, I., Brecht, A., and Gugglitz, G. Covalent strategy for immobilization of DNA-microspots suitable for microarrays with label-free and time-resolved optical detection of hybridization. *Fresenius' Journal of Analytical Chemistry* **2001**; 371(2):128-136.
- Jurinke, C., van den Boom, D., Cantor, C. R., and Köster, H. Automated genotyping using the DNA MassArray technology. *Methods in Molecular Biology* **2001**; 170:103-116.
- Jurinke, C., van den Boom, D., Cantor, C. R., and Köster, H. Automated genotyping using the DNA MassArray technology. *Methods in Molecular Biology* **2002**; 187:179-192.
- Kadota, K., Miki, R., Bono, H., Shimizu, K., Okazaki, Y., and Hayashizaki, Y. Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. *Physiological Genomics* **2001**; 4(3):183-188.
- Kahn, J., Mehraban, F., Ingle, G., Xin, X., Bryant, J. E., Vehar, G., Schoenfeld, J., Grimaldi, C. J., Peale, F., Draksharapu, A., Lewin, D. A., and Gerritsen, M. E. Gene expression profiling in an *in vitro* model of angiogenesis. *American Journal of Pathology* **2000**; 156(6):1887-1900.
- Kal, A. J., van Zonneveld, A. J., Benes, V., van den Berg, M., Koerkamp, M. G., Albermann, K., Strack, N., Ruijter, J. M., Richter, A., Dujon, B., Ansorge, W., and Tabak, H. F. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular Biology of the Cell* **1999**; 10(6):1859-1872.
- Kamb, A., and Ramaswami, M. A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnology* **2001**; 1(1):8.
- Kambhampati, D., Nielsen, P. E., and Knoll, W. Investigating the kinetics of DNA-DNA and PNA-DNA interactions using surface plas-

- mon resonance-enhanced fluorescence spectroscopy. *Biosensors and Bioelectronics* **2001**; 16(9-12):1109-1118.
- Kampke, T., Kieninger, M., and Mecklenburg, M. Efficient primer design algorithms. *Bioinformatics* **2001**; 17(3):214-225.
- Kane, M. D., Jatke, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research* **2000**; 28(22):4552-4557.
- Kaplan, B. B., Schachter, B. S., Osterburg, H. H., de Vellis, J. S., and Finch, C. E. Sequence complexity of polyadenylated RNA obtained from rat brain regions and cultured rat cells of neural origin. *Biochemistry* **1978**; 17(25):5516-5524.
- Karas, M., and Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* **1988**; 60(20):2299-2301.
- Karsten, S. L., Van Deerlin, V. M., Sabatti, C., Gill, L. H., and Geschwind, D. H. An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis. *Nucleic Acids Research* **2002**; 30(2):E4.
- Kato, K. Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Research* **1995**; 23(18):3685-3690.
- Kato, K. Adapter-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Research* **1997**; 25(22):4694-4696.
- Kato, T., Imatani, T., Miura, T., Minaguchi, K., Saitoh, E., and Okuda, K. Cytokine-inducing activity of family 2 cystatins. *Biological Chemistry* **2000**; 381(11):1143-1147.
- Katz, A. I., Doucet, A., and Morel, F. Na-K-ATPase activity along the rabbit, rat, and mouse nephron. *American Journal of Physiology* **1979**; 237(2):F114-F120.
- Kaufmann, R. Matrix-assisted laser desorption ionization (MALDI) mass spectrometry: a novel analytical tool in molecular biology and biotechnology. *Journal of Biotechnology* **1995**; 41(2-3):155-175.
- Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyo-sawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojohori, T., Bono, H., Kasu-kawa, T., Saito, R., Kadota, K., Matsuda, H., Ashburner, M., Batalov, S., Casavant, T., Fleischmann, W., Gaasterland, T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S., Matsuo, Y., Nikaido, I., Pesole, G., Quackenbush, J., Schriml, L. M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bjunga, N., Carninci, P., de Bonaldo, M. F., Brownstein, M. J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill, D., Hofmann, M., Hume, D. A., Kamiya, M., Lee, N. H., Lyons, P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring, B., Ringwald, M., Rodriguez, I., Sakamoto, N., Sasaki, H., Sato, K., Schonbach, C., Seya, T., Shibata, Y., Storch, K. F., Suzuki, H., Toyooka, K., Wang, K. H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-Boris, A., Yoshida, K., Hasegawa, Y., Kawaji, H., Kohtsuki, S., and Hayashizaki, Y. Functional annotation of a full-length mouse cDNA collection. *Nature* **2001**; 409(6821):685-690.
- Kawamoto, S., Ohnishi, T., Chisaka, O., and Okubo, K. Expression profiling by 'cDNA-AFLP': A PCR based method for genome-wide gene expression profiling. *Genome Research* **1999**; 9(12):1305-1312.
- Kawamoto, S., Yoshii, J., Mizuno, K., Ito, K., Miyamoto, Y., Ohnishi, T., Matoba, R., Hori, N., Matsumoto, Y., Okumura, T., Nakao, Y., Yoshii, H., Arimoto, J., Ohashi, H., Nakanishi, H., Ohno, I., Hashimoto, J., Shimizu, K., Maeda, K., Kuriyama, H., Nishida, K., Shimizu-Matsumoto, A., Adachi, W., Ito, R., Kawasaki, S., and Chae, K. S. BodyMap: a collection of 3' ESTs for analysis of human gene expression information. *Genome Research* **2000**; 10(11):1817-1827.
- Kenten, J. H., Gudibande, S. R., Link, J., Willey, J., Curfman, B., Major, E. O., and Massay, R. J. Improved electrochemiluminescent label for DNA probe assays: rapid quantitative assays of HIV-1 polymerase chain reaction products. *Clinical Chemistry* **1992**; 38(6):873-879.
- Kenzelmann, M., and Muhlemann, K. Substantially enhanced cloning efficiency of SAGE (serial analysis of gene expression) by adding a heating step to the original protocol. *Nucleic Acids Research* **1999**; 27(3):917-918.

- Kerr, M. K., and Churchill, G. A. Statistical design and the analysis of gene expression microarray data. *Biostatistics* **2001a**; 2(2): 183-201.
- Kerr, M. K., and Churchill, G. A. Statistical design and the analysis of gene expression microarray data. *Genetical Research* **2001b**; 77(2):123-128.
- Kerr, M. K., Martin, M., and Churchill, G. A. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **2000**; 7(6):819-837.
- Kessler, C. Nonradioactive analysis of biomolecules. 2nd edition. Kessler, C., editor. Springer Verlag, Berlin, Germany **2000**.
- King, H. C., and Sinha, A. A. Gene expression profile analysis by DNA microarrays: promise and pitfalls. *Journal of the American Medical Association* **2001**; 286(18):2280-2288.
- Knight, J. When the chips are down. *Nature* **2001a**; 410(6831):860-861.
- Knight, J. Geneticists' work in disarray as DNA-chip producer pulls the plug. *Nature* **2001b**; 414(6860):135-136.
- Koester, H., Tang, K., Fu, D. J., Braun, A., van den Boom, D., Smith, C. L., Cotter, R. J., and Cantor, C. R. A strategy for rapid and efficient DNA sequencing by mass spectrometry. *Nature Biotechnology* **1996** 14(9):1123-1128.
- Kohroki, J., Tsuchiya, M., Fujita, S., and Nakanishi, T. A novel strategy for identifying differential gene expression: an improved method of differential display analysis. *Biochemical and Biophysical Research Communications* **1999**; 262(2):365-367.
- Kornberg, A. Why purify enzyme? *Methods in Enzymology* **1990**; 182:1-5.
- Kornmann, B., Preitner, N., Rifat, D., Fleury-Olela, F., and Schibler, U. Analysis of circadian liver gene expression by ADDER, a highly sensitive method for the display of differentially expressed mRNAs. *Nucleic Acids Research* **2001**; 29(11):E51.
- Korri-Youssef, H., Garnier, F., Srivastava, P., Godillot, P., and Yassar, A. Toward bioelectronics: Specific DNA recognition based on an oligonucleotide-functionalized polypyrrole. *Journal of the American Chemical Society* **1997**; 119(31):7388-7391.
- Kricka, L. J. Nucleic acid detection technologies – labels, strategies, and formats. *Clinical Chemistry* **1999**; 45(4):453-458.
- Kricka L. J. Stains, labels and detection strategies for nucleic acids assays. *Annals of Clinical Biochemistry* **2002**; 39(Part 2):114-129.
- Kukanskis, K., Elkind, J., Melendez, J., Murphy, T., Miller, G., and Garner, H. Detection of DNA hybridization using the TISPR-1 surface plasmon resonance biosensor. *Analytical Biochemistry* **1999**; 274(1):7-17.
- Kumar, A., and Whitesides, G. M. Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol "ink" followed by chemical etching. *Applied Physics Letters* **1993**; 63(14):2002-2004.
- Kwiatkowski, R. W., Lyamichev, V., de Arruda, M., and Neri, B. Clinical, genetic, and pharmacogenetic applications of the invader assay. *Molecular Diagnosis* **1999**; 4(4):353-364.
- Lackie, P. M. Immunogold silver staining for light microscopy. *Histochemistry and Cell Biology* **1996**; 106(1):9-17.
- Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. A public database for gene expression in human cancers. *Cancer Research* **1999**; 59(21):5403-5407.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., Le Vine, R., McEwan, P., McKernan, K., Meldrum, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chisoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L.,

- Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Honkamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Petersen, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., and Szustakowski, J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**; 409(6822):860-921.
- Larsen, L. A., Christiansen, M., Vuust, J., and Andersen, P. S. Recent developments in high-throughput mutation screening. *Pharmacogenomics* **2001**; 2(4):387-399.
- Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. SAGEmap: A public gene expression resource. *Genome Research* **2000**; 10(7):1051-1060.
- Lawson, N. D., Krause, D. S., and Berliner, N. Normal neutrophil differentiation and secondary granule gene expression in the EML and MPRO cell lines. *Experimental Hematology* **1998**; 26(12):1178-1185.
- Lee, S., and Tomasetto, C. R. S. Positive selection of candidate tumor-suppressor genes by subtractive hybridisation. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(7):2825-2829.
- Lee, N. H., Weinstock, K. G., Kirkness, E. F., Earle-Hughes, J. A., Fuldner, R. A., Marmaros, S., Glodek, A., Gocayne, J. D., Adams, M. D., Kerlavage, A. R., Fraser, C. M., and Venter, J. C. Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(18): 8303-8307.
- Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(18):9834-9839.
- Lee, T. Y., and Shim, Y. B. Direct DNA hybridization detection based on the oligonucleotide-functionalized conductive polymer. *Analytical Chemistry* **2001**; 73(22):5629-5632.
- Lee, S., Chen, J., Zhou, G., and Wang, S. M. Generation of high-quantity and quality tag/ditag cDNAs for SAGE analysis. *Biotechniques* **2001**; 31(2):348-350, 352-354.
- Lennon, G. G., and Lehrach, H. Hybridization analyses of arrayed cDNA libraries. *Trends in Genetics* **1991**; 7(10):314-317.
- Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M. B. The I. M. A. G. E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **1996**; 33(1):151-152.
- Leushner, J. MALDI TOF mass spectrometry: an emerging platform for genomics and

- diagnostics. *Expert Review in Molecular Diagnostics* **2001**; 1(1):11-18.
- Leuschner, J., and Chiu, N. H. Automated mass spectrometry: a revolutionary technology for clinical diagnostics. *Molecular Diagnosis* **2000**; 5(4):341-348.
- Lewin, B. Gene expression. Lewin, B., editor. Wiley, New York, New York, USA **1980**.
- Lewin, B. Genes IV. Lewin, B., editor. Oxford University Press, Oxford, United Kingdom **1990**:466-481.
- Li, F., and Stormo, G. D. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **2001**; 17(11):1067-1076.
- Li, C., and Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2001**; 2(8):RESEARCH0032.1-RESEARCH0032.11.
- Li, Q., Luo, G., Feng, J., Cai, D., and Qi, O. Photoelectrochemistry as a novel strategy for DNA hybridization detection. *Analyst* **2000a**; 125(11):1908-1910.
- Li, L., Garden, R. W., and Sweedler, J. V. Single-cell MALDI: a new tool for direct peptide profiling. *Trends in Biotechnology* **2000b**; 18(4):151-160.
- Lian, Z., Wang, L., Yamaga, S., Bonds, W., Beazer-Barclay, Y., Kluger, Y., Gerstein, M., Newburger, P. E., Berliner, N., and Weissman, S. M. Genomic and proteomic analysis of the myeloid differentiation program. *Blood* **2001**; 98(3):513-524.
- Liang, P. Gene discovery using differential display. *Genetic Engineering News* **2000**; 20:37.
- Liang, P., and Pardee, A. B. Differential display of eukaryotic mRNA by means of the polymerase chain reaction. *Science* **1992**; 257(5072):967-971.
- Liang, P., Averboukh, L., and Pardee, A. B. Method of differential display. In: *Methods in molecular genetics*. Adolph, K. W., editor. Academic Press, San Diego, California, USA **1994**:3-16.
- Liang, P., and Pardee, A. B. Recent advances in differential display. *Current Opinion in Immunology* **1995**; 7(2):274-280.
- Liang, P., Bauer, D., Averboukh, L., Warthoe, P., Rohrwild, M., Muller, H., Strauss, M., and Pardee, A. B. Analysis of altered gene expression by differential display. *Methods in Enzymology* **1995**; 254:304-321.
- Liang, F., Holt, I., Perteau, G., Karamycheva, S., Salzberg, S. L., and Quackenbush, J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics* **2000**; 25(2):239-240.
- Liang, S., Rossby, S. P., Liang, P., Shelton, R. C., Manier, D. H., Chakrabarti, A., and Sulser, F. Detection of an mRNA polymorphism by differential display. *Molecular Biotechnology* **2001**; 19(2):121-124.
- Ligler, F. S., Rowe, C. A., Balderson, S., Feldstein, M., and Golden, J. P. Fluorescence array biosensor – Biochemistry and application. In: *Proceedings of micro total analysis systems '98*. Harrison, D. J., and van den Berg, A., editors. Kluwer Scientific Publishing, Dordrecht, The Netherlands **1998**: 217-220.
- Limbach, P. A., Crain, P. F., and McCloskey, J. A. Characterization of oligonucleotides and nucleic acids by mass spectrometry. *Current Opinion in Biotechnology* **1995**; 6(1):96-102.
- Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nature Genetics* **1999**; 21(1 Supplement):20-24.
- Lisitsyn, N., Lisitsyn N, and Wigler, M. Cloning the differences between two complex genomes. *Science* **1993**; 259(5097):946-951.
- Little, D. P., Braun, A., Darnhofer-Demar, B., and Köster, H. Identification of apolipoprotein E polymorphisms using temperature cycled primer oligo base extension and mass spectrometry. *European Journal of Clinical Chemistry and Clinical Biochemistry* **1997**; 35(7):545-548.
- Little, D. P., Braun, A., O'Donnell, M. J., and Köster, H. Mass spectrometry from miniaturized arrays for full comparative DNA analysis. *Nature Medicine* **1997a**; 3(12): 1413-1416.
- Little, D. P., Braun, A., Darnhofer-Demar, B., Frilling, A., Li, Y., McIver, R. T., Jr., and Köster, H. Detection of RET proto-oncogene codon 634 mutations using mass spectrometry. *Journal of Molecular Medicine* **1997b**; 75(10):745-750.
- Lo, D., Hilbush, B., and Sutcliffe, J. G. TOGA analysis of gene expression to accelerate target development. *European Journal of Pharmaceutical Sciences* **2001**; 14(3):191-196.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. Expression monitoring by hybridization to high-density oligonucleo-

- tide arrays. *Nature Biotechnology* **1996**; 14(13):1675-1680.
- Lockhart, D. J., and Winzler, E. A. Genomics, gene expression and DNA arrays. *Nature* **2000**; 405(6788):827-836.
- Long, A. D., Mangalam, H. J., Chan, B. Y., Toller, L., Hatfield, G. W., and Baldi, P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *Journal of Biological Chemistry* **2001**; 276(23):19937-19944.
- Lukyanov, K., Diatchenko, L., Chenchik, A., Nanisetti, A., Siebert, P., Usman, N., Matz, M., and Lukyanov, S. Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect. *Biochemical and Biophysical Research Communications* **1997**; 230(2):285-288.
- Lysov, I. P., Florent'ev, V. L., Khorlin, A. A., Khrapko, K. R., and Shik, V. V. Determination of the nucleotide sequence of DNA using hybridisation with oligonucleotides. A new method. *Doklady Akademii Nauk SSSR* **1988**; 303(6):1508-1511.
- MacLeod, M. C., Gaddis, S. S., and Aldaz, C. M. Combinatorial oligonucleotide PCR: a method for rapid, global expression analysis. *US6221600* **2001**.
- Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Research* **2000**; 28(1):126-128.
- Mahadevappa, M., and Warrington, J. A. A high-density probe array sample preparation method using 10- to 100-fold fewer cells. *Nature Biotechnology* **1999**; 17(11):1134-1136.
- Maier, E., Meier-Ewert, S., Ahmadi, A. R., Curtis, J., and Lehrach, H. Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. *Journal of Biotechnology* **1994**; 35(2-3): 191-203.
- Man, M. Z., Wang, X., and Wang, Y. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* **2000**; 16(11):953-959.
- Manduchi, E., Grant, G. R., McKenzie, S. E., Overton, G. C., Surrey, S., and Stoekert, C. J., Jr. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* **2000**; 16(8):685-698.
- Margulies, E. H., and Innis, J. W. eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics* **2000**; 16(7):650-651.
- Margulies, E. H., Kardia, S. L., and Innis, J. W. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Research* **2001**; 29(12):e60.
- Marshall, E. Affymetrix settles suit, fixes mouse chips. *Science* **2001**; 291(5513): 2535.
- Mathieu-Daude, F., Welsh, J., Vogt, T., and McClelland, M. DNA rehybridization during PCR: the 'Cot effect' and its consequences. *Nucleic Acids Research* **1996**; 24(11): 2080-2086.
- Matsumura, H., Nirasawa, S., and Terauchi, R. Technical advance: transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). *The Plant Journal* **1999**; 20(6):719-726.
- Matz, M., Usman, N., Shagin, D., Bogdanov, E., and Lukyanov, S. Ordered differential display: a simple method for systematic comparison of gene expression profiles. *Nucleic Acids Research* **1997**; 25(12):2541-2542.
- Matz, M. V., and Lukyanov, S. A. Different strategies of differential display: areas of application. *Nucleic Acids Research* **1998**; 26(24):5537-5543.
- Mazzola, L. T., and Fodor, S. P. A. Imaging biomolecule arrays by atomic force microscopy. *Biophysical Journal* **1995**; 68(5): 1653-1660.
- Mazzola, L. T., Frank, C. W., Fodor, S. P. A., Mosher, C., Lartius, R., and Henderson, E. Discrimination of DNA hybridization using chemical force microscopy. *Biophysical Journal* **1999**; 76(6):2922-2933.
- McGall, G., Labadie, J., Brock, P., Wallraff, G., Nguyen, T., and Hinsberg, W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(24):13555-13560.
- Mecklenburg, M. XNA on Gold™: A versatile DNA biochip. *Bio World* **1999a**; 4:1-3.
- Mecklenburg, M. XNA on Gold™: A versatile DNA biochip. *Nature Genetics* **1999b**; 23(3 Supplement):61-62.
- Meier-Ewert, S., Maier, E., Ahmadi, A., Curtis, J., and Lehrach, H. An automated approach to generating expressed sequence catalogues. *Nature* **1993**; 361(6410):375-376.

- Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., Freund, J., Elge, T., Mott, R., Herrmann, B., and Lehrach, H. Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Research* **1998**; 26(9):2216-2223.
- Mikkelsen, S. R. Electrochemical biosensors for DNA sequence detection. *Electroanalysis* **1996**; 8(1):15-19.
- Millan, K. M., and Mikkelsen, S. R. Sequence-selective biosensor for DNA based on electroactive hybridization indicators. *Analytical Chemistry* **1993**; 65(17):2317-2323.
- Millan, K. M., Saraullo, A., and Mikkelsen, S. R. Voltammetric DNA biosensor for cystic fibrosis based on a modified carbon paste electrode. *Analytical Chemistry* **1994**; 66(18):2943-2948.
- Milosavljevic, A., Strezoska, Z., Zeremski, M., Grujic, D., Paunesku, T., and Crkvenjakov, R. Clone clustering by hybridisation. *Genomics* **1995**; 27(1):83-89.
- Milosavljevic, A., Zeremski, M., Strezoska, Z., Grujic, D., Dyanov, H., Batus, S., Salbego, D., Paunesku, T., Soares, M. B., and Crkvenjakov, R. Discovering distinct genes represented in 29,570 clones from infant brain cDNA libraries by applying sequencing by hybridisation methodology. *Genome Research* **1996**; 6(2):132-141.
- Mir, K. U. The hypothesis is there is no hypothesis. The Microarray Meeting, Scottsdale, Arizona, USA, 22-25 September 1999. *Trends in Genetics* **2000**; 16(2):63-64.
- Müller, H. J. In: Heritage from Mendel. Brink, R. A., editor. University of Wisconsin Press, Madison, Wisconsin, USA **1967**:419.
- Moreno, J. C., Pauws, E., van Kampen, A. H., Jedlickova, M., de Vijlder, J. J., and Ris-Stalpers, C. Cloning of tissue-specific genes using serial analysis of gene expression and a novel computational subtraction approach. *Genomics* **2001**; 75(1-3):70-76.
- Morhard, F., Dahint, R., and Grunze, M. *In situ* detection of cells and biochemical reactions by optical diffraction. In: Proceedings of micro total analysis systems '98 Harrison, D. J., und van den Berg, A., editors. Kluwer Scientific Publishing, Dordrecht, The Netherlands **1998**:469-472.
- Morosetti, R., Park, D. J., Chumakov, A. M., Grillier, I., Shiohara, M., Gombart, A. F., Nakamaki, T., Weinberg, K., and Koeffler, H. P. A novel, myeloid transcription factor, C/EBP epsilon, is upregulated during granulocytic, but not monocytic, differentiation. *Blood* **1997**; 90(7):2591-2600.
- Müller, H. P., and Kern, H. Radiation resistance, content and base composition of DNA in some radiation-induced mutants of *Azotobacter chroococcum*. *Zeitschrift für Naturforschung. Teil B: Chemie, Biochemie, Biophysik, Biologie* **1967**; 22(12):1330-1336.
- Munasinghe, A., Patankar, S., Cook, B. P., Madden, S. L., Martin, R. K., Kyle, D. E., Shoaibi, A., Cummings, L. M., and Wirth, D. F. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. *Molecular and Biochemical Parasitology* **2001**; 113(1):23-34.
- Mutch, D. M., Berger, A., Mansourian, R., Rytz, A., and Roberts, M. A. Microarray data analysis: a practical approach for selecting differentially expressed genes. *Genome Biology* **2001**; 2(12):PREPRINT0009.1-PRE-PRINT0009.31.
- N. N. The nature of the number. *Nature Genetics* **2000**; 25(2):127-128.
- Naef, F., Hacker, C. R., Patil, N., and Magnasco, M. Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biology* **2002a**; 3(1):PREPRINT0001.1-PREPRINT0001.24.
- Naef, F., Hacker, C. R., Patil, N., and Magnasco, M. Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biology* **2002b**; 3(4):RESEARCH0018.1-RESEARCH0018.11.
- Nair, S. M., Werkman, T. R., Craig, J., Finnell, R., Joels, M., and Eberwine, J. H. Corticosteroid regulation of ion channel conductances and mRNA levels in individual hippocampal CA1 neurons. *Journal of Neuroscience* **1998**; 18(7):2685-2696.
- Nakatani, K., Sando, S., and Saito, I. Scanning of guanine-guanine mismatches in DNA by synthetic ligands using surface plasmon resonance. *Nature Biotechnology* **2001**; 19(1):51-55.
- Nallur, G., Luo, C., Fang, L., Cooley, S., Dave, V., Lambert, J., Kukanskis, K., Kingsmore, S., Lasken, R., and Schweitzer, B. Signal amplification by rolling circle amplification on DNA microarrays. *Nucleic Acids Research* **2001**; 29(23):E118.

- Nam, N. Q., McGall, G. H., and Rava, R. P. Photocleavable protecting groups and methods for their use. *WO9839348* **1998**.
- Nam, N. Q., McGall, G. H., and Rava, R. P. Photocleavable protecting groups and methods for their use. *WO9839348* **1998**.
- Neilson, L., Andalibi, A., Kang, D., Coutifaris, C., Strauss, J. F., 3rd, Stanton, J. A., and Green, D. P. Molecular phenotype of the human oocyte by PCR-SAGE. *Genomics* **2000**; 63(1):13-24.
- Nelson, B. P., Grimsrud, T. E., Liles, M. R., Goodman, R. M., and Corn, R. M. Surface plasmon resonance imaging measurements of DNA and RNA hybridization adsorption onto DNA microarrays. *Analytical Chemistry* **2001**; 73(1):1-7.
- Newburger, P. E., Subrahmanyam, Y. V., and Weissman, S. M. Global analysis of neutrophil gene expression. *Current Opinion in Hematology* **2000**; 7(1):16-20.
- Newton, M. A., Kendzierski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **2001**; 8(1):37-52.
- Nielsen, P. E., Egholm, M., Berg, R. H., and Buchardt, O. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science* **1991**; 254(5037):1497-1500.
- Nikura, K., Nagata, K., and Okahata, Y. Quantitative detection of protein binding onto DNA strands by using a quartz-crystal microbalance. *Chemistry Letters* **1996**; 9:863-864.
- Nordhoff, E., Ingendoh, A., Cramer, R., Overberg, A., Stahl, B., Karas, M., Hillenkamp, F., and Crain, P. F. Matrix-assisted laser desorption/ionization mass spectrometry of nucleic acids with wavelengths in the ultraviolet and infrared. *Rapid Communications in Mass Spectrometry* **1992**; 6(12):771-776.
- Okahata, Y., Matsunobu, Y., Ijiri, K., Mukai, M., Murakami, A., and Makino, K. Hybridization of nucleic acids immobilized on a quartz-crystal microbalance. *Journal of the American Chemical Society* **1992**; 114(21):8299-8300.
- Okamoto, T., Suzuki, T., and Yamamoto, N. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nature Biotechnology* **2000**; 18(4):438-441.
- Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics* **1992**; 2(3):173-179.
- Pabón, C., Modrusan, Z., Ruvolo, M. V., Coleman, I. M., Daniel, S., Yue, H., and Arnold, L. J., Jr. Optimized T7 amplification system for microarray analysis. *Biotechniques* **2001**; 31(4):874-879.
- Paleček, E., Fojta, M., Tomschick, M., and Wang, J. Electrochemical biosensors for DNA hybridization and DNA damage. *Biosensors and Bioelectronics* **1998**; 13(5):621-628.
- Paleček, E., Fojta, M., and Jelen, F. New approaches in the development of DNA sensors: hybridization and electrochemical detection of DNA and RNA at two different surfaces. *Bioelectrochemistry* **2002**; 56(1-2):85-90.
- Paracchini, S., Arredi, B., Chalk, R., and Tyler-Smith, C. Hierarchical high-throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Research* **2002**; 30(6):e27.
- Park, S. J., Taton, T. A., and Mirkin, C. A. Array-based electrical detection of DNA with nanoparticle probes. *Science* **2002**; 295(5559):1503-1506.
- Parle-McDermott, A., McWilliam, P., Tighe, O., Dunican, D., and Croke, D. T. Serial analysis of gene expression identifies putative metastasis-associated transcripts in colon tumour cell lines. *British Journal of Cancer* **2000**; 83(6):725-728.
- Patolsky, F., Katz, E., Bardea, A., and Willner, I. Enzyme-linked amplified electrochemical sensing of oligonucleotide-DNA interactions by means of the precipitation of an insoluble product and using impedance spectroscopy. *Langmuir* **1999**; 15(11):3703-3706.
- Pelizzari, C. A., Khodarev, N. N., Gupta, N., Calvin, D. P., and Weichselbaum, R. R. Quantitative analysis of DNA array autoradiographs. *Nucleic Acids Research* **2000**; 28(22):4577-4581.
- Peters, D. G., Kassam, A. B., Yonas, H., O'Hare, E. H., Ferrell, R. E., and Brufsky, A. M. Comprehensive transcript analysis in small quantities of mRNA by SAGE-lite. *Nucleic Acids Research* **1999**; 27(24):e39.
- Peters, D. G., Kassam, A. B., Feingold, E., Heideich-O'Hare, E., Yonas, H., Ferrell, R. E., and

- Brufsky, A. Molecular anatomy of an intracranial aneurysm. Coordinated expression of genes involved in wound healing and tissue remodeling. *Stroke* **2001**; 32(4):1036-1042.
- Peterson, A. W., Heaton, R. J., and Georgiadis, R. M. The effect of surface probe density on DNA hybridization. *Nucleic Acids Research* **2001**; 29(24):5163-5168.
- Phillips, J., and Eberwine, J. H. Antisense RNA amplification: A linear amplification method for analyzing the mRNA population from single living cells. *Methods* **1996**; 10(3): 283-288.
- Piehler, J., Brecht, A., Gauglitz, G., Maul, C., Grabley, S., and Zerlin, M. Specific binding of low molecular weight ligands with direct optical detection. *Biosensors and Bioelectronics* **1997a**; 12(6):531-538.
- Piehler, J., Brecht, A., Gauglitz, G., Zerlin, M., Maul, C., Thiericke, R., and Grabley, S. Label-free monitoring of DNA-ligand interactions. *Analytical Biochemistry* **1997a**; 249(1):94-102.
- Pieles, U., Züricher, W., Schär, M., and Moser, H. E. Matrix assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for mass and sequence analysis of natural and modified nucleotides. *Nucleic Acids Research* **1993**; 21(14):3191-3196.
- Pisano J. A golden opportunity. *The Scientist* **2001**; 15(15):24.
- Pollard-Knight, D., Hawkins, E., Yeung, D., Pashby, D. P., Simpson, M., McDougall, A., Buckle, P., and Charles, S. A. Immunoassays and nucleic acid detection with a biosensor based on surface plasmon resonance. *Annales de Biologie Clinique* **1990**; 48(9):642-646.
- Porter, D. A., Krop, I. E., Nasser, S., Sgroi, D., Kaelin, C. M., Marks, J. R., Riggins, G., and Polyak, K. A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Research* **2001**; 61(15):5697-5702.
- Poustka, A., Pohl, T., Barlow, D. P., Zehetner, G., Craig, A., Michiels, F., Ehrlich, E., Frischauf, A. M., and Lehrach, H. Molecular approaches to mammalian genetics. *Cold Spring Harbor Symposium in Quantitative Biology* **1986**; 51(Part 1):131-139.
- Poustka, A. J., Herwig, R., Krause, A., Hennig, S., Meier-Ewert, S., and Lehrach, H. Toward the gene catalogue of sea urchin development: the construction and analysis of an unfertilized egg cDNA library highly normalised by oligonucleotide fingerprinting. *Genomics* **1999**; 59(2):122-133.
- Powell, J. Enhanced concatemer cloning—a modification to the SAGE (serial analysis of gene expression) technique. *Nucleic Acids Research* **1998**; 26(14):3445-3446.
- Prashar, Y., and Weissman, S. M. Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(2): 659-663.
- Prashar, Y., and Weissman, S. M. READS: a method for display of 3' end fragments of restriction enzyme-digested cDNAs for analysis of differential gene expression. *Methods of Enzymology* **1999**; 303:258-272.
- Proudnikov, D., and Mirzabekov, A. Chemical methods of DNA and RNA fluorescent labeling. *Nucleic Acids Research* **1996**; 24(22):4535-4542.
- Proudnikov, D., Timofeev, E., and Mirzabekov, A. Immobilization of DNA in polyacrylamide gel for the manufacture of DNA and DNA-oligonucleotide microchips. *Analytical Biochemistry* **1998**; 259(1):34-41.
- Pruitt, K. D., and Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* **2001**; 29(1):137-140.
- Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends in Genetics* **2000**; 16(1): 44-47.
- Quackenbush, J. Computational analysis of microarray data. *Nature Reviews in Genetics* **2001**; 2(6):418-427.
- Qin, L., Prins, P., Jones, J. T., Popeijus, H., Smant, G., Bakker, J., and Helder, J. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. *Nucleic Acids Research* **2001**; 29(7):1616-1622.
- Raddatz, G., Dehio, M., Meyer, T. F., and Dehio, C. PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics* **2001**; 17(1):98-99.
- Radelof, U., Hennig, S., Seranski, P., Steinfath, M., Ramser, J., Reinhardt, R., Poustka, A., Francis, F., and Lehrach, H. Preselection of shotgun clones by oligonucleotide fingerprinting: an efficient and high throughput strategy to reduce redundancy in large-scale sequencing projects. *Nucleic Acids Research* **1998**; 26(23):5358-5364.
- Radtkey, R., Feng, L., Muralhidar, M., Duhon, M., Canter, D., DiPierro, D., Fallon, S., Tu, E.,

- McElfresh, K., Nerenberg, M., and Sosnowski, R. Rapid, high fidelity analysis of simple sequence repeats on an electronically active DNA microchip. *Nucleic Acids Research* **2000**; 28(7):E17.
- Ramdas, L., Coombes, K. R., Baggerly, K., Abruzzo, L., Highsmith, W. E., Krogmann, T. Hamilton, S. R., and Zhang, W. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology* **2001**; 2(11):RESEARCH0047.
- Reinartz, J. R., Bruyns, E., Lin, J., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik R. Massively parallel signature sequencing (MPSS™) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in Functional Genomics and Proteomics* **2002**; 1(1):95-104.
- Relógio, A., Schwager, C., Richter, A., Ansorge, W., and Valcarcel, J. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Research* **2002**; 30(11):E51.
- Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., and Blattner, F. R. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research* **1999**; 27(19):3821-3835.
- Rifkin, S. A., Atteson, K., and Kim, J. Constraint structure analysis of gene expression. *Functional and Integrative Genomics* **2000**; 1(3):174-185.
- Riggins, G. J., and Strausberg, R. L. Genome and genetic resources from the Cancer Genome Anatomy Project. *Human Molecular Genetics* **2001**; 10(7):663-667.
- Rininger, J. A., DiPippo, V. A., and Gould-Rothberg, B. E. Differential gene expression technologies for identifying surrogate markers of drug efficacy and toxicity. *Drug Discovery Today* **2000**; 5(12):560-568.
- Rocke, D. M., and Durbin, B. A model for measurement error for gene expression arrays. *Journal of Computational Biology* **2001**; 8(6):557-569.
- Roepstorff, P. MALDI-TOF mass spectrometry in protein chemistry. *EXS* **2000**; 88:81-97.
- Rothstein, J. L., Johnson, D., DeLoia, J. A., Skowronski, J., Solter, D., and Knowles, B. Gene expression during preimplantation mouse development. *Genes and Development* **1992**; 6(7):1190-1201.
- Robert-Nicoud, M., Flahaut, M., Elalouf, J.-M., Nicod, M., Salinas, M., Bens, M., Doucet, A., Wincker, P., Artiguenave, F., Horisberger, J. D., Vandewalle, A., Rossier, B. C., and Firsov, D. Transcriptome of a mouse kidney cortical collecting duct cell line: effects of aldosterone and vasopressin. *Proceedings of the National Academy of the Sciences of the USA* **2001**; 98(5):2712-2716.
- Roest, C. H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quetier, F., Saurin, W., and Weissenbach, J. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genetics* **2000**; 25(2):235-238.
- Rouillard, J. M., Herbert, C. J., and Zuker, M. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **2002**; 18(3):486-487.
- Rose, S. D. Application of a novel microarraying system in genomics research and drug discovery. *Journal of the Association for Laboratory Automation* **1998**; 3(3):53-56.
- Ross, P., Hall, L., Smirnov, I., and Haff, L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature Biotechnology* **1998**; 16(13):1347-1351.
- Ross, P., Hall, L., and Haff, L. A. Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. *Biotechniques* **2000**; 29(3):620-629.
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **1985**; 230(4732):1350-1354.
- Salin, H., Vujasinovic, T., Mazurie, A., Maitrejean, S., Menini, C., Mallet, J., and Dumas, S. A novel sensitive microarray approach for differential screening using probes labelled with two different radioelements. *Nucleic Acids Research* **2002**; 30(4):e17.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, New York, USA **1989**.
- Sauer, S., Lechner, D., Berlin, K., Lehrach, H., Escary, J. L., Fox, N., and Gut, I. G. A novel procedure for efficient genotyping of single nucleotide polymorphisms. *Nucleic Acids Research* **2000**; 28(5):E13.
- Sauer, S., Gelfand, D. H., Boussicault, F., Bauer, K., Reichert, F., and Gut, I. G. Facile method for automated genotyping of single

- nucleotide polymorphisms by mass spectrometry. *Nucleic Acids Research* **2002**; 30(5):e22.
- Sauerbrey, G. Verwendung von Schwingquarzen zur Wägung dünner Schichten und zur Mikrowägung. *Zeitschrift für Physik C* **1959**; 155:206-222.
- Schadt, E. E., Li, C., Su, C., and Wong, W. H. Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **2000**; 80(2):192-202.
- Schadt, E. E., Li, C., Ellis, B., and Wong, W. H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **2001**; Supplement 37:120-125.
- Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Analytical Biochemistry* **1995**; 227(2):255-273.
- Schena, M., and Davis, R. W. Gene, genomes, and chips. In: *DNA microarrays: A Practical Approach*. Schena, M., editor. Oxford University Press, New York, New York, USA **1999**:1-16.
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**; 270(5235):467-470.
- Schmid, C. W. Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. *Progress in Nucleic Acid Research and Molecular Biology* **1996**; 53:283-319.
- Schobel, U., Barzen, C., and Gauglitz, G. Immunanalytical techniques for pesticide monitoring based on fluorescence detection. *Fresenius Journal of Analytical Chemistry* **2000**; 366(6-7): 646-658.
- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzog, H. Normalization strategies for cDNA microarrays. *Nucleic Acids Research* **2000**; 28(10):E47.
- Schutzbank, T. E., and Smith, J. Detection of human immunodeficiency virus type 1 proviral DNA by PCR using an electrochemiluminescence-tagged probe. *Journal of Clinical Microbiology* **1995**; 33(8):2036-2041.
- Schwarz, T., Yeung, D., Hawkins, E., Heaney, P., and McDougall, A. Detection of nucleic acid hybridization using surface plasmon resonance. *Trends in Biotechnology* **1991**; 9(10):339-340.
- Seligmann, B. Improved genomics and proteomics array technology. *American Biotechnology Laboratory* **2001**; 19(11):34.
- Service, R. F. Analytical chemistry. New test could speed bioweapon detection. *Science* **2002**; 295(5559):1447.
- Shahgholi, M., Garcia, B. A., Chiu, N. H., Heaney, P. J., and Tang, K. Sugar additives for MALDI matrices improve signal allowing the smallest nucleotide change (A:T) in a DNA sequence to be resolved. *Nucleic Acids Research* **2001**; 29(19):E91.
- Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analysing complex DNA samples using two-color fluorescent probe hybridisation. *Genome Research* **1996**; 6(7):639-645.
- Shchepinov, M. S., Udalovala, I. A., Bridgman, A. J., and Southern, E. M. Oligonucleotide dendrimers: synthesis and use as polylabelled DNA probes. *Nucleic Acids Research* **1997**; 25(22):4447-4454.
- Shchepinov, M. S., Denissenko, M. F., Smylie, K. J., Worl, R. J., Leppin, A. L., Cantor, C. R., and Rodi, C. P. Matrix-induced fragmentation of P3'-N5' phosphoramidate-containing DNA: high-throughput MALDI-TOF analysis of genomic sequence polymorphisms. *Nucleic Acids Research* **2001**; 29(18):3864-3872.
- Shen, L., Sheridan, P., Cao, W. W., Dailey, P. J., Salazar-Gonzalez, J. F., Breen, E. C., Fahey, J. L., Urdea, M. S., and Kolberg, J. A. Quantification of cytokine mRNA in peripheral blood mononuclear cells using branched DNA (bdNA) technology. *Journal of Immunological Methods* **1998a**; 215(1-2):123-134.
- Shen, H., Schultz, M., Kruh, G. D., and Tew, K. D. Increased expression of DNA-dependent protein kinase confers resistance to adriamycin. *Biochimica et Biophysica Acta* **1998b**; 1381(2):131-138.
- Shimkets, R. A., Lowe, D. G., Tai, J. T., Sehl, P., Jin, H., Yang, R., Predki, P. F., Rothberg, B. E., Murtha, M. T., Roth, M. E., Shenoy, S. G., Windemuth, A., Simpson, J. W., Simons, J. F., Daley, M. P., Gold, S. A., McKenna, M. P., Hillan, K., Went, G. T., and Rothberg, J. M. Gene expression analysis by transcript profiling coupled to a gene database query. *Nature Biotechnology* **1999**; 17(8):798-780.
- Shoshan, A., Grebinskiy, V., Magen, A., Scolnicov, A., Fink, E., Lehari, D. and Wasserman, A. Designing oligo libraries taking al-

- ternative splicing into account. In: *Proceedings of SPIE – International Society for Optical Engineering. Microarrays: Optical technologies and informatics*. Bittner, M. L., Chen, Y., Dorsel, A. N. and Dougherty, E. D., editors **2001**; 4266(6):86-95.
- Sigurdsson, F., Khanna-Gupta, A., Lawson, N., and Berliner, N. Control of late neutrophil-specific gene expression: insights into regulation of myeloid differentiation. *Seminars in Hematology* **1997**; 34(4):303-310.
- Sinclair, B. Everything's great when it sits on a chip. *The Scientist* **1999**; 13(11):18-20.
- Singer-Sam, J., Robinson, M. O., Bellve, A. R., Simon, M. I., and Riggs, A. D. Measurement by quantitative PCR of changes in HPRT, PGK-1, PGK-2, APRT, MTase, and Zfy gene transcripts during mouse spermatogenesis. *Nucleic Acids Research* **1990**; 18(5):1255-1259.
- Singh-Gassom, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnology* **1999**; 17(10):974-978.
- Sloviter, R. S., Valiquette, G., Abrams, G. M., Ronk, E. C., Sollas, A. L., Paul, L. A., and Neubort, S. Selective loss of hippocampal granule cells in the mature rat brain after adrenalectomy. *Science* **1989**; 243(4890): 535-538.
- Smaglik, P. Researchers take a gamble on the human genome. *Nature* **2000**; 405(6784):264.
- Smith, E. A., Kyo, M., Kumasawa, H., Nakatani, K., Saito, I., and Corn, R. M. Chemically induced hairpin formation in DNA monolayers. *Journal of the American Chemical Society* **2002**; 124(24):6810-6811.
- Soares, M. B., Bonaldo, M. F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. Construction and characterisation of a normalised cDNA library. *Proceedings of the National Academy of Science of the USA* **1994**; 91(20):9228-9232.
- Sosnowski, R. G., Tu, E., Butler, W. F., O'Connell, J. P., and Heller, M. J. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proceedings of the National Academy of Sciences of the USA* **1997**; 94(4):1119-1123.
- Southern, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **1975**; 98(3):503-517.
- Speicher, M. R., Gwyn Ballard, S., and Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* **1996**; 12(4):368-375.
- Spengler, B., Pan, Y., Cotter, R. J., and Kan, L. S. Molecular weight determination of underivatized oligodeoxyribonucleotides by positive-ion matrix-assisted ultraviolet laser-desorption mass spectrometry. *Rapid Communications in Mass Spectrometry* **1990**; 4(4):99-102.
- Spinella, D. G., Bernardino, A. K., Redding, A. C., Koutz, P., Wei, Y., Pratt, E., Myers, K., Chappel, G., Gerken, S., and McConnell, S. J. Tandem Arrayed Ligation of Expressed Sequence Tags (TALEST): A rapid method for generating global gene expression profiles. *Nucleic Acids Research* **1999**; 27(18):e22.
- St. Croix, B., Rago, C., Velculescu, V., Traverso, G., Romans, K. E., Montgomery, E., Lal, A., Riggins, G. J., Lengauer, C., Vogelstein, B., and Kinzler, K. W. Genes expressed in human tumor endothelium. *Science* **2000**; 289(5482):1197-1202.
- Stanton, J. L., and Green, D. P. L. A set of 840 mouse oocyte genes with well-matched human homologues. *Molecular Human Reproduction* **2001**; 7(6):521-543.
- Stears, R. L., Getts, R. C., and Gullans, S. R. A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiological Genomics* **2000**; 3(2): 93-99.
- Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., and O'Brien, J. Automated image analysis for array hybridisation experiments. *Bioinformatics* **2001**; 17(7): 634-641.
- Stemmler, E. A., Hettich, R. L., Hurst, G. B., and Buchanan, M. V. Matrix-assisted laser desorption/ionization Fourier-transform mass spectrometry of oligodeoxyribonucleotides. *Rapid Communications in Mass Spectrometry* **1993**; 7(9):828-836.
- Stoerker, J., Mayo, J. D., Tetzlaff, C. N., Sarra-cino, D. A., Schwoppe, I., and Richert, C. Rapid genotyping by MALDI-monitored nuclease selection from probe libraries. *Nature Biotechnology* **2000**; 18(11):1213-1216.
- Stollberg, J., Urschitz, J., Urban, Z., and Boyd, C. D. A quantitative evaluation of SAGE. *Genome Research* **2000**; 10(8):1241-1248.
- Stomakhin, A. A., Vasiliskov, V. A., Timofeev, E., Schulga, D., Cotter, R. J., and Mirzabekov,

- A. D. DNA sequence analysis by hybridization with oligonucleotide microchips: MALDI mass spectrometry identification of 5mers contiguously stacked to microchip oligonucleotides. *Nucleic Acids Research* **2000**; 28(5):1193-1198.
- Strausberg, R. L., Feingold, E. A., Klausner, R. D., and Collins, F. S. The mammalian gene collection. *Science* **1999**; 286(5439): 455-457.
- Strausberg, R. L., and Riggins, G. J. Navigating the human transcriptome. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(21):11837-11838.
- Strausberg, R. L. The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. *Journal of Pathology* **2001**; 195(1):31-40.
- Su, H., Williams, P., and Thompson, M. Platinum anticancer drug binding to DNA detected by thickness-shear-mode acoustic wave sensor. *Analytical Chemistry* **1995**; 67(5): 1010-1013.
- Subrahmanyam, Y. V., Baskaran, N., Newburger, P. E., and Weissman, S. M. A modified method for the display of 3'-end restriction fragments of cDNAs: molecular profiling of gene expression in neutrophils. *Methods in Enzymology* **1999**; 303:272-297.
- Subrahmanyam, Y. V. B. K., Yamaga, S., Newburger, P. E., and Weissman, S. M. A modified approach for the efficient display of 3' end restriction fragments of cDNAs. In: Differential display: A practical approach. Practical approach series. Leslie, R. A., and Robertson, H. A., editors. Oxford University Press, Oxford, United Kingdom **2000**:101-129.
- Subrahmanyam, Y. V., Yamaga, S., Prashar, Y., Lee, H. H., Hoe, N. P., Kluger, Y., Gerstein, M., Goguen, J. D., Newburger, P. E., and Weissman, S. M. RNA expression patterns change dramatically in human neutrophils exposed to bacteria. *Blood* **2001**; 97(8): 2457-2468.
- Sun, X., Ding, H., Hung, K., and Guo, B. A new MALDI-TOF based mini-sequencing assay for genotyping of SNPs. *Nucleic Acids Research* **2000**; 28(12):E68.
- Sun, B., Xie, W., Yi, G., Chen, D., Zhou, Y., and Cheng, J. Microminiaturized immunoassays using quantum dots as fluorescent label by laser confocal scanning fluorescence detection. *Journal of Immunological Methods* **2001**; 249(1-2):85-89.
- Sutcliffe, J. G. mRNA in the mammalian central nervous system. *Annual Review of Neuroscience* **1988**; 11:157-198.
- Sutcliffe, J. G., Foye, P. E., Erlander, M. G., Hilbush, B. S., Bodzin, L. J., Durham, J. T., and Hasel, K. W. TOGA: an automated parsing technology for analyzing expression of nearly all genes. *Proceedings of the National Academy of Science of the USA* **2000**; 97(5):1976-1981.
- Sutcliffe, J. G. Open-system approaches to gene expression in the CNS. *Journal of Neurosciences* **2001**; 21(21):8306-8309.
- Takenaka, S., Uto, Y., Kondo, H., Ihara, T., and Takagi, M. Electrochemically active DNA probes: detection of target DNA sequences at femtomole level by high-performance liquid chromatography with electrochemical detection. *Analytical Biochemistry* **1994**; 218(2):436-443.
- Takenaka, S., Uto, Y., Saita, H., Yokoyama, M., Kondo, H., and Wilson, D. Electrochemically active threading intercalator with high double stranded DNA selectivity. *Chemical Communications* **1998**; 10:1111-1112.
- Tang, K., Allman, S. L., Jones, R. B., Chen, C. H., and Araghi, S. Laser mass spectrometry of oligonucleotides with isomer matrices. *Rapid Communications in Mass Spectrometry* **1993a**; 7(6):435-439.
- Tang, K., Allman, S. L., and Chen, C. H. Matrix-assisted laser desorption ionization of oligonucleotides with various matrices. *Rapid Communications in Mass Spectrometry* **1993b**; 7(10):943-948.
- Tang, K., Allman, S. L., Chen, C. H., Chang, L. Y., and Schell, M. Matrix-assisted laser desorption/ionization of restriction enzyme-digested DNA. *Rapid Communications in Mass Spectrometry* **1994**; 8(2):183-186.
- Tang, K., Fu, D., Kotter, S., Cotter, R. J., Cantor, C. R., and Köster, H. Matrix-assisted laser desorption/ionization mass spectrometry of immobilized duplex DNA probes. *Nucleic Acids Research* **1995**; 23(16):3126-3131.
- Tang, K., Fu, D. J., Julien, D., Braun, A., Cantor, C. R., and Köster, H. Chip-based genotyping by mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(18):10016-10020.
- Tang, K., Shahgholi, M., Garcia, B. A., Heaney, P. J., Cantor, C. R., Scott, L. G., and Williamson, J. R. Improvement in the apparent mass resolution of oligonucleotides by using 12C/

- 14N-enriched samples. *Analytical Chemistry* **2002**; 74(1):226-231.
- Taton, T. A., Mirkin, C. A., and Letsinger, R. L. Scanometric DNA array detection with nanoparticle probes. *Science* **2000**; 289(5485): 1757-1760.
- Tenen, D. G., Hromas, R., Licht, J. D., and Zhang, D. E. Transcription factors, normal myeloid development, and leukemia. *Blood* **1997**; 90(2):489-519.
- Theriault, T. P., Winder, S. C., and Gamble, R. C. Application of ink-jet printing technology to the manufacture of molecular arrays. In: DNA microarrays: A Practical Approach. Schena, M., editor. Oxford University Press, New York, New York, USA **1999**:101-120.
- Thomas, J. G., Olson, J. M., Tapscott, S. J., and Zhao, L. P. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* **2001**; 11(7): 1227-1236.
- Tomac, S., Sarkar, M., Ratilainen, T., Wittung, P., Nielsen, P. E., Norden, B., and Graslund, A. Ionic effects on the stability and conformation of peptide nucleic acid complexes. *Journal of the American Chemical Society* **1996**; 118(24):5544-5552.
- Tremain, N., Korkko, J., Ibberson, D., Kopen, G. C., DiGirolamo, C., and Phinney, D. G. MicroSAGE analysis of 2,353 expressed genes in a single cell-derived colony of undifferentiated human mesenchymal stem cells reveals mRNAs of multiple cell lineages. *Stem Cells* **2001**; 19(5):408-418.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**; 17(6):520-525.
- Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **2001**; 29(12):2549-2557.
- Tsodikov, A., Szabo, A., and Jones, D. Adjustments and measures of differential expression for microarray data. *Bioinformatics* **2002**; 18(2):251-260.
- Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(9):5116-5121.
- Tyagi, S. Taking a census of mRNA populations with microbeads. *Nature Biotechnology* **2000**; 18(6):597-598.
- Uhlmann, E. Oligonucleotide technologies: synthesis, production, regulations and applications. 29-30th November 2000, Hamburg, Germany. *Expert Opinion on Biological Therapy* **2001**; 1(2):319-328.
- Urdea, M. S. Branched DNA signal amplification. *Biotechnology* **1994**; 12(9):926-928.
- Van de Rijke, F., Zijlmans, H., Li, S., Vail, T., Raap, A. K., Niedbala, R. S., and Tanke, H. J. Up-converting phosphor reporters for nucleic acid microarray. *Nature Biotechnology* **2001**; 19(3):273-276.
- Van den Berg, A., van der, Leij J., and Poppe, S. Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. *Nucleic Acids Research* **1999**; 27(17):e17.
- Van Gemen, B., van Beunigen, R., Nabbe, A., van Strijp, D., Jurriaans, S., Lens, P., and Kievits, T. A one-tube quantitative HIV-1 RNA NASBA nucleic acid amplification assay using electrochemiluminescent (ECL) labelled probes. *Journal of Virological Methods* **1994**; 49(2):157-167.
- Van Gijlswijk, R. P., Talman, E. G., Janssen, P. J., Snoeijers, S. S., Killian, J., Tanke, H. J., and Heetebrij, R. J. Universal Linkage System: versatile nucleic acid labeling technique. *Expert Reviews in Molecular Diagnostics* **2001**; 1(1):81-91.
- Van Kampen, A. H., van Schaik, B. D., Pauws, E., Michiels, E. M., Ruijter, J. M., Caron, H. N., Versteeg, R., Heisterkamp, S. H., Leunissen, J. A., Baas, F., and van der Mee, M. USAGE: a web-based approach towards the analysis of SAGE data. *Serial Analysis of Gene Expression. Bioinformatics* **2000**; 16(10):899-905.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **1995**; 270(5235):484-487.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, A. M., Bassett, E. D., Hieter, P., Vogelstein, B., and Kinzler, P. Characterisation of the yeast transcription. *Cell* **1997**; 88(2):243-251.
- Velculescu, V. E., Zhang, L., Zhou, W., Traverso, G., St. Croix, B., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression: detailed protocol (version 1.0d), Johns Hopkins

- Oncology Center and Howard Hughes Medical Institute, Baltimore, Maryland, USA 1999.
- Velculescu, V. E., Vogelstein, B., and Kinzler, K. W. Analysing uncharted transcriptomes with SAGE. *Trends in Genetics* **2000**; 16(10):423-425.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huseon, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Murganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu X. The sequence of the human genome. *Science* **2001**; 291(5507): 1304-1351.
- Virlon, B., Cheval, L., Buhler, J. M., Billon, E., Doucet, A., and Elalouf, J. M. Serial microanalysis of renal transcriptomes. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(26):15286-15291.
- Vizi, E. S., and Kiss, J. P. Neurochemistry and pharmacology of the major hippocampal transmitter systems: synaptic and nonsynaptic interactions. *Hippocampus* **1998**; 8(6): 566-607.
- Vreeke, M., Rocca, P., and Heller, A. Direct electrical detection of dissolved biotinylated horseradish peroxidase, biotin, and avidin. *Analytical Chemistry* **1995**; 67(2):303-306.
- Wang, J. Towards genoelectronics: Electrochemical biosensing of DNA hybridization. *Chemistry – A European Journal* **1999**; 5(6): 1681-1685.
- Wang, J. From DNA biosensors to gene chips. *Nucleic Acids Research* **2000**; 28(16): 3011-3016.
- Wang, J., Rivas, G., Ozsoz, M., Grant, D., Cai, X., and Parrado, C. Microfabricated electro-

- chemical sensor for the detection of radiation-induced DNA damage. *Analytical Chemistry* **1997a**; 69(7):1457-1460.
- Wang, J., Nielsen, P. E., Jiang, M., Cai, X., Fernandes, J. R., Grant, D. H., Ozsoz, M., Beglieter, A., and Mowat, M. Mismatch-sensitive hybridization detection by peptide nucleic acids immobilized on a quartz crystal microbalance. *Analytical Chemistry* **1997b**; 69(24):5200-5202.
- Wang, J., Rivas, G., Fernandes, J., Paz, J. L., Jiang, M., and Waymire, R. Indicator-free electrochemical DNA hybridization biosensor. *Analytica Chimica Acta* **1998**; 375(3): 197-203.
- Wang, J., Jiang, M., and Palecek, E. Real-time monitoring of enzymatic cleavage of nucleic acids using a quartz crystal microbalance. *Bioelectrochemistry and Bioenergetics* **1999a**; 48(2):477-480.
- Wang, J., Jiang, M., Fortes, A., and Mukherjee, B. New label-free DNA recognition based on doping nucleic-acid probes within conducting polymer films. *Analytica Chimica Acta* **1999b**; 402(1-2):7-12.
- Wang, A., Pierce, A., Judson-Kremer, K., Gad-dis, S., Aldaz, C. M., Johnson, D. G., and MacLeod, M. C. Rapid analysis of gene expression (RAGE) facilitates universal expression profiling. *Nucleic Acids Research* **1999b**; 27(23):4609-4618.
- Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T., and Marincola, F. M. High-fidelity mRNA amplification for gene profiling. *Nature Biotechnology* **2000a**; 18(4):457-459.
- Wang, A., Schneider-Broussard, R., Kumar, A. P., MacLeod, M. C., and Johnson, D. G. Regulation of BRCA1 expression by the Rb-E2F pathway. *Journal of Biological Chemistry* **2000b**; 275(6):4532-4536.
- Wang, X., Ghosh, S., and Guo, S. W. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research* **2001a**; 29(15):e75.
- Wang, J., Kawde, A. N., Erdem, A., and Salazar, M. Magnetic bead-based label-free electrochemical detection of DNA hybridization. *Analyst* **2001b**; 126(11):2020-2024.
- Wang, J., Xu, D., Kawde, A. N., and Polsky, R. Metal nanoparticle-based electrochemical stripping potentiometric detection of DNA hybridization. *Analytical Chemistry* **2001c**; 73(22):5576-5581.
- Wang, A., Johnson, D. G., and MacLeod, M. C. Molecular cloning and characterisation of a novel mouse epidermal differentiation gene and its promoter. *Genomics* **2001d**; 73(3): 284-290.
- Wang, J., and Kawde, A. N. Amplified label-free electrical detection of DNA hybridization. *Analyst* **2002**; 127(3):383-386.
- Warrior, U., Fan, Y., David C. A., Wilkin, J. A., McKeegan, E. M., Kofron, J. L., and Burns, D. J. Application of Quantigene nucleic acid quantification technology for high throughput screening. *Journal of Biomolecular Screening* **2000**; 5(5):343-352.
- Watanabe, Y., Gould, E., and McEwen, B. S. Stress induces atrophy of apical dendrites of hippocampal CA3 pyramidal neurons. *Brain Research* **1992**; 588(2):341-345.
- Weiner, A. M., Deininger, P. L., and Efstratiadis, A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annual Review in Biochemistry* **1986**; 55: 631-661.
- Welle, S., Bhatt, K., and Thornton, C. A. Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Research* **1999**; 9(5):506-513.
- Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D., and McClelland, M. Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research* **1992**; 20(19):4965-4970.
- Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., and Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(1):334-339.
- Werle, E., Schneider, C., Renner, M., Volker, M., and Fiehn, W. Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Research* **1994**; 22(20):4354-4355.
- Westin, L., Xu, X., Miller, C., Wang, L., Edman, C. F., and Nerenberg, M. Anchored multiplex amplification on a microelectronic chip array. *Nature Biotechnology* **2000**; 18(2):199-204.
- Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2000**; 28(1): 10-14.

- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2001**; 29(1):11-16.
- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Research* **2002**; 30(1):13-16.
- Whitney, L. W., and Becker, K. G. Radioactive 33-P probes in hybridization to glass cDNA microarrays using neural tissues. *Journal of Neuroscience Methods* **2001**; 106(1):9-13.
- Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocker, M., Blocker, H., Bauersachs, S., Blum, H., Lauber, J., Dusterhoft, A., Beyer, A., Kohrer, K., Strack, N., Mewes, H. W., Ottenwalder, B., Obermaier, B., Tampe, J., Heubner, D., Wambutt, R., Korn, B., Klein, M., and Poustka, A. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Research* **2001**; 11(3):422-435.
- Williamson, A. R. The Merck Gene Index project. *Drug Discovery Today* **1999**; 4(3):115-122.
- Wodicka, L., Dong, H., Mittmann, H., Ho, M., and Lockhart, D. J. Genome wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnology* **1997**; 15(13):1359-1367.
- Wooster, R. Cancer classification with microarrays is less more? *Trends in Genetics* **2000**; 16(8):327-329.
- Wu, K. J., Steding, A., and Becker, C. H. Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. *Rapid Communications in Mass Spectrometry* **1993**; 7(2):142-146.
- Xu, X.-H., Yang, H. C., Mallouk, T. E., and Bard, A. J. Immobilization of DNA on an aluminum(III) alkanebisphosphonate thin film with electrogenerated chemiluminescent detection. *Journal of the American Chemical Society* **1994**; 116(18):8386-8387.
- Xu, X.-H., and Bard, A. J. Immobilization and hybridization of DNA on an aluminum(III) alkanebisphosphonate thin film with electrogenerated chemiluminescent detection. *Journal of the American Chemical Society* **1995**; 117(9):2627-2631.
- Yamamoto, M., Wakatsuki, T., Hada, A., and Ryo, A. Use of serial analysis of gene expression (SAGE) technology. *Journal of Immunological Methods* **2001**; 250(1-2):45-66.
- Yamanaka, R., Lekstrom-Himes, J., Barlow, C., Wynshaw-Boris, A., and Xanthopoulos, K. G. CCAAT/enhancer binding proteins are critical components of the transcriptional regulation of hematopoiesis. *International Journal of Molecular Medicine* **1998**; 1(1):213-221.
- Yamashita, T., Hashimoto, S., Kaneko, S., Nagai, S., Toyoda, N., Suzuki, T., Kobayashi, K., and Matsushima, K. Comprehensive gene expression profile of a normal human liver. *Biochemical and Biophysical Research Communications* **2000**; 269(1):110-116.
- Yang, M. C., Ruan, Q. G., Yang, J. J., Eckenrode, S., Wu, S., McIndoe, R. A., and She, J. X. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiological Genomics* **2001**; 7(1):45-53.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **2002**; 30(4):e15.
- Yates, J. R., 3rd. Mass spectrometry. From genomics to proteomics. *Trends in Genetics* **2000**; 16(1):5-8.
- Ye, S. Q., Zhang, L. Q., Zheng, F., Virgil, D., and Kwitrovich, P. O. MiniSAGE: gene expression profiling using serial analysis of gene expression from 1 microg total RNA. *Analytical Biochemistry* **2000**; 287(1):144-152.
- Yershov, G., Barsky, V., Belgovskiy, A., Kirillov, E., Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, A., Dubiley, S., and Mirzabekov, A. DNA analysis and diagnostics on oligonucleotide microchips. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(10):4913-4918.
- Young, B. D., and Birnie, G. D. Complexity and specificity of polysomal poly(A⁺) RNA in mouse tissues. *Biochemistry* **1976**; 15(13):2823-2829.
- Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., Sealfon, S. C. Accuracy and calibration

- of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research* **2002**; 30(10):e48.
- Zakin, L., Reversade, B., Virlon, B., Rusniok, C., Glaser, P., Elalouf, J.-M., and Brûlet, P. Gene expression profiles in normal and *Otx2*^{-/-} early gastrulating mouse embryos. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(26):14288-14293.
- Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., and Kinzler, K.W. Gene expression profiles in normal and cancer cells. *Science* **1997**; 276(5316):1268-1272.
- Zhang, K., and Zhao, H. Assessing reliability of gene clusters from gene expression data. *Functional and Integrative Genomics* **2000**; 1(3):156-173.
- Zhao, N., Hashida, H., Takahashi, N., Misumi, Y., and Sakaki, Y. High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. *Gene* **1995**; 156(2):207-213.
- Zhou, Y., Kalocsai, P., Chen, J., and Shams, S. Information processing issues and solutions associated with microarray technology. In: *Microarray biochip technology*. BioTechniques Books. Schena, M., Editor. Eaton Publishing Corporation, Natick, Massachusetts, USA **2000**:167-200.
- Zhumabayeva, B., Diatchenko, L., Chenchik, A., and Siebert, P.D. Use of SMART-generated cDNA for gene expression studies in multiple human tumors. *Biotechniques* **2001**; 30(1):158-163.
- Zhuo, D., Zhao, W.D., Wright, F.A., Yang, H.Y., Wang, J.P., Sears, R., Baer, T., Kwon, D.H., Gordon, D., Gibbs, S., Dai, D., Yang, Q., Spitzner, J., Krahe, R., Stredney, D., Stutz, A., and Yuan, B. Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Research* **2001**; 11(5):904-918.
- Zien, A., Aigner, T., Zimmer, R., and Lengauer, T. Centralization: a new method for the normalization of gene expression data. *Bioinformatics* **2001**; 17(Supplement 1):S323-S331.

5

Protein expression analysis

5.1

Introduction

The study of proteins has been compared to the fight of Ulysses against the demigod Proteus, in which Proteus constantly transforms himself in order to outwit the hero (Wickware & Smaglik, 2001). This comparison describes quite well the difficulties faced by scientists working on proteins. Although based on defined genomic information, proteins develop a life of their own. First of all, in eukaryotic cells each gene encodes not one but as many as six to eight different proteins (Strohman, 1994). This is due to post-translational modifications such as phosphorylation, glycosylation or cleavage (see also chapter 1). During the life of the cell, the expression of proteins is up or down-regulated, huge multi-protein complexes are formed and dissolved, metal ions are bound and released and the proteins move from one part of the cell to another. For this reason, a thorough description of the proteome, *i.e.* the complete set of all proteins and their modifications of a cell, seems not to be possible and only 'snapshots' of protein profiles under specific physiological conditions are available. The term proteomics has been introduced to describe the large-scale characterisation of these protein profiles using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (Wasinger *et al.*, 1995; Wilkins *et al.*, 1996).

It was not until the emergence of biological mass spectrometry in the 1990s that large-scale protein identification became feasible. In 1996, as many as 150 yeast proteins were analysed by two-dimensional polyacrylamide gel electrophoresis and direct mass spectrometric analysis of the excised protein spots for the first time (Shevchenko *et al.*, 1996). Today, this technique is routinely applied to the detection of thousands of proteins.

Proteomics has now become an industry in its own right. Special software was necessary to effectively mine the databases and, therefore, the field of bioinformatics grew alongside. Today, peptide masses are routinely sent to another continent over the internet and the result of a database query is available in seconds. Basic tasks such as similarity analysis or pattern and profile prediction, but also advanced calculations such as the modelling of the tertiary structure of proteins, can be performed with public available software from anywhere in the world with an internet connection. Three popular gateways to proteome and genome data, and software tools are described in table 5.1.

Table 5.1 Gateways to proteome and genome data, and software tools.

Database	Internet link	Additional information	References
WORLD-2DPAGE (Swiss Institute of Bioinformatics, SIB; Geneva, Switzerland)	http://www.expasy.org/ch2d/2d-index.html	Index of 2D-PAGE databases of specific cell or tissue types.	Appel <i>et al.</i> (1996)
National Center for Biotechnology Information (NCBI; National Library of Medicine, Bethesda, Maryland, USA).	http://www.ncbi.nlm.nih.gov/	Databases of complete genomes of various organisms as well as mRNA, EST and protein databases. The NCBI creates public databases, conducts research in computational biology, develops software tools for analysing genome data, and disseminates biomedical information.	Tatusova <i>et al.</i> (1999); Wheeler <i>et al.</i> (2000, 2001 and 2002)
ExPASy (Expert Protein Analysis System) Molecular Biology Server (Swiss Institute of Bioinformatics)	http://www.expasy.org/	Gateway to protein databases and software tools. The ExPASy server of the Swiss Institute of Bioinformatics is dedicated to the analysis of protein sequences and structures as well as 2D-PAGE.	Hoogland <i>et al.</i> (1999); Wilkins <i>et al.</i> (1999)

In contrast to the accumulation of genomic sequence information, which for the human genome will be finished at some point in the near future, the investigation of function on the protein level will continue for years to come.

5.2

Sample separation in proteomics

5.2.1

Introduction

The simple but most important question for the choice of procedure for the separation of proteins is: what has to be separated? Is it a single protein, a group of proteins or the whole proteome of a cell or tissue? The answer to that question defines: (i) The technique to be used. The three most relevant techniques are conventional slab gel electrophoresis including isoelectric focussing (IEF), sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and two-dimensional PAGE, high-performance liquid chromatography (HPLC) and capillary electrophoresis (CE). (ii) The separation principle, for example, ion exchange, size exclusion,

or affinity. (iii) The need to combine several successive methods to a multidimensional separation procedure.

The following sections give an overview about current procedures in the field of protein separation. Some commonly used applications are presented in more detail.

5.2.2

Conventional slab gel electrophoresis

5.2.2.1 Isoelectric focusing (IEF)

All proteins are amphotolytes. The net charge of each protein is the sum of the positive and negative charges of the amino groups and carboxy acid groups, respectively. For every protein there exists a pH at which the net charge is zero. This is called the isoelectric point and usually ranges between pH 3 and pH 11.

In isoelectric focusing (IEF), proteins migrate within an inert polyacrylamide matrix containing a stable gradient according to their net charge under denaturing conditions and stop migrating when they reach their individual isoelectric point (figure 5.1).

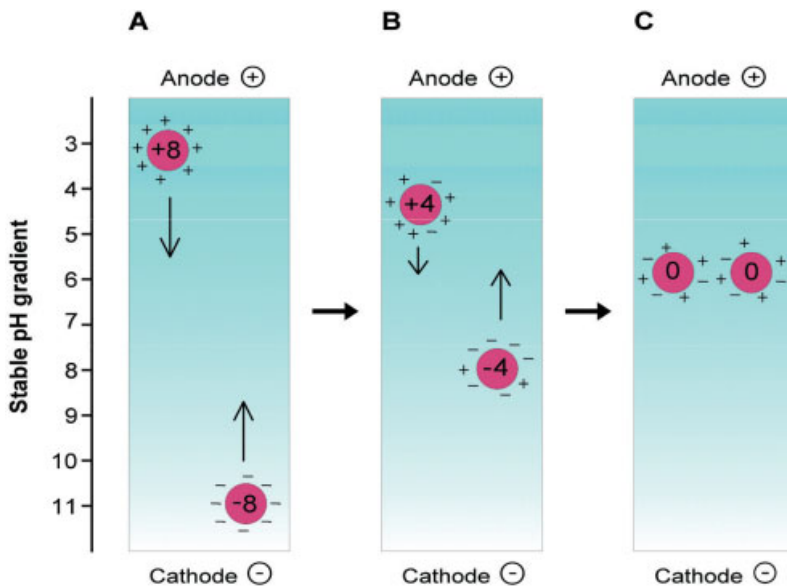


Figure 5.1 Isoelectric focusing (IEF). IEF is a method to separate proteins according to their net charge by electrophoresis. The gel matrix provides a stable and continuous pH gradient. **A)** At low pH, the amino groups of proteins are positively charged while the carboxylic acid groups are uncharged resulting in a positive net charge of the protein. At low pH, the situation

is reversed. The amino groups remain uncharged while the carboxylic acid groups carry a negative charge resulting in a negative net charge of the protein. **B)** The application of an electric field causes the migration of the charged protein through the gel matrix along the pH gradient. **C)** At the isoelectric pH, the protein has no net charge and accumulates to a sharp band.

Two approaches are available to provide a stable pH gradient within the matrix. The historical earlier method includes the use of carrier ampholytes (CA-IEF), initially described for slab gel electrophoresis by Robertson *et al.* (1987). In this method, a mixture of low molecular weight ampholytes covering the pH of interest is embedded into the carrier matrix of polyacrylamide. During a first gel run, these carrier ampholytes find their individual positions between cathode and anode where they supply a buffered pH that corresponds to their respective isoelectric point. A major problem with the use of carrier ampholytes is a time-dependent shift of the expected gradient called the 'cathodic drift' (Rilbe, 1977). This delay of the gradient generally decreases the reproducibility of CA-IEF.

The introduction of an immobilised pH gradient (IPG-IEF) has abolished this problem (Bjellquist *et al.*, 1982). In this variant, buffering acryloyl monomers are covalently bound to the acrylamide matrix during gel production. The buffering components are immobilised at fixed positions providing a stable and reproducible pH gradient. Another advantage of IPG-IEF is that large amounts of protein (up to 5 mg) can be applied so that more low abundance proteins become detectable. On the other hand, the resolution of some hydrophobic (*e.g.*, membrane) proteins is poor and others get lost when separated by immobilised pH gradient (Molloy, 2000). The most important characteristic of IEF is that it represents the first dimension of two-dimensional gels and is therefore an essential step in analysis of protein expression.

5.2.2.2 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and Western blotting

Proteins differ in three physical parameters: molecular weight, shape, and charge. SDS-PAGE is a common and simple method to determine the approximate molecular mass of a polypeptide or protein and its subunits or to verify the success of a protein preparation. This is possible as the influences of shape and charge on the migration behaviour of a polypeptide can be eliminated. SDS and β -mercaptoethanol are the chemical reagents that are used to ensure that the migration of proteins is proportional to the logarithm of their molecular masses (Weber & Osborn, 1972). β -mercaptoethanol is one component of the buffer in which the sample is heated before being applied to the gel. β -mercaptoethanol reduces disulphide bridges to sulfhydryl groups so that proteins lose their tertiary and quaternary structure. SDS is present in all components of the electrophoresis. The anionic detergent dodecyl sulphate has a negatively charged hydrophilic head region and a long hydrophobic tail region. The latter interacts with hydrophobic regions of a polypeptide and gives the surface an overall negative charge *via* its charged head regions that neutralises the positive charges of the polypeptide chain. The net effect of this treatment is to produce polypeptide or protein molecules whose migration behaviour is a pure function of their molecular mass. SDS-PAGE is not suitable for all proteins, however. For example, the molecular mass of some very large proteins and proteins containing large amounts of carbohydrates cannot be determined using this method.

Most commonly, SDS-PAGE is implemented as a discontinuous electrophoresis in slab gels (figure 5.2) according to the procedure described by Laemmli (1970). During the first part of the electrophoresis, the sample is concentrated in a small sharp band by means of a stacking gel. The stacked proteins then enter the separation gel at almost the same time. Both the stacking gel and the separation gel consists of a porous polyacrylamide matrix, but the pores in the separation gel are smaller. Usually, the stacking separation gel contains four (to separate very large proteins) to fifteen (to separate small proteins) percent acrylamide depending on the protein of interest. Polypeptides smaller than 15 kilodalton should be separated in tricine gels (Schägger & von Jagow, 1987). Proteins can be visualised by staining with Coomassie Blue dye or silver. Silver staining is more sensitive than other non-radioactive staining methods.

For the unambiguous identification of a polypeptide within a gel and for numerous downstream applications, it is often useful to transfer the separated polypeptides electrophoretically out of the gel matrix onto the surface of a membrane. This process is called Western blotting (Towbin *et al.*, 1979). The gel and the blotting buffer are first equilibrated in transfer buffer and the gel is laid in direct contact onto the blotting membrane that is usual of polyvinylidene difluoride (PVDF) or nitrocellulose. This transfer leads to the quantitative immobilisation of the previously separated proteins on the surface of the membrane. Usually, this is carried out electrophoretically in a submersed blotting chamber. The success of the transfer

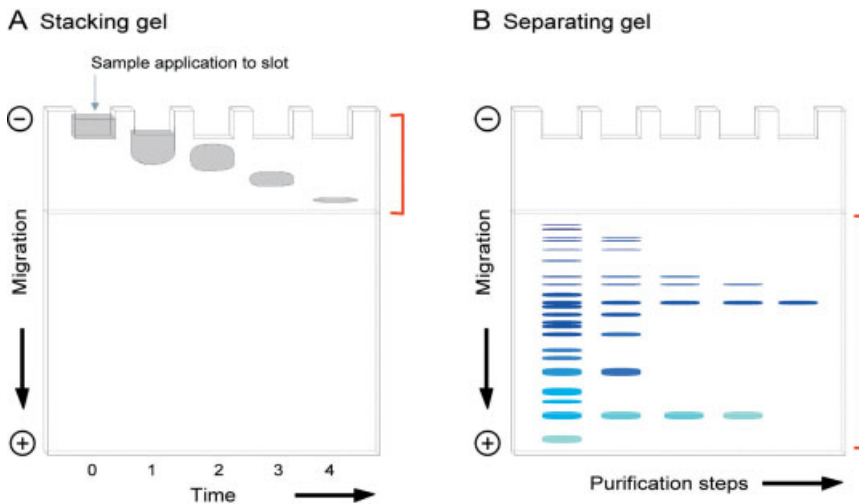


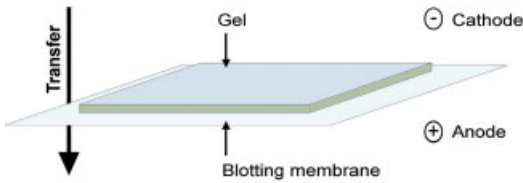
Figure 5.2 Sodium dodecyl sulfate-based polyacrylamide gel electrophoresis (SDS-PAGE).

A) Polypeptides applied to a gel slot accumulate to a small band during their migration through the electric field in the stacking gel forwarded by the electric field. **B)** In the presence of SDS, polypeptides are separated according to their molecular masses as the anionic detergent SDS

binds to polypeptide chains giving them an overall negative charge. Protein bands are usually stained with Coomassie Blue. The example shows the typical band patterns from crude cell extract (lane on the left) to the over 90 percent purified protein (lane on the right) occurring during a protein purification.

is checked by reversible staining of proteins on the membrane using Ponceau S red stain. Western blotting is also called immunoblotting^{3,4} as the most frequent downstream application is the immunochemical identification of relevant proteins using antibodies directed against the protein of interest that allow identification of the separated protein (figure 5.3).

A Blotting



B Temporary visualisation



All transferred polypeptides are visualised temporarily by red Ponceau S staining

C Immunochemical identification



Polypeptides of interest are visualised immunochemically

Figure 5.3 Western blotting procedure.

A) Blotting. All polypeptides separated by polyacrylamide gel electrophoresis are quantitatively transferred to the surface of a blotting membrane by electrophoresis. This procedure is called Western blotting. **B) Temporary**

visualisation. The efficiency of the transfer can be controlled by staining polypeptides reversibly using Ponceau S red stain.

C) Immunochemical detection. Polypeptides of interest are visualised and identified by specific antibodies.

5.2.2.3 Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)

One of the greatest challenges of the proteome analysis is the reproducible separation of the up to predicted 50,000 to 100,000 different cellular proteins while maintaining the relative amounts of these proteins, the abundance of which varies between ten and 10^6 copies in a single cell. The most widely used and most suitable method for quantitatively comparing changes in protein profiles of cells, tissues, or whole cell organisms is 2D-PAGE (Quadroni & James, 1999). Originally described by O'Farrell (1975), 2D-PAGE is a multi-dimensional separation method that requires two successive operations. In the first dimension, proteins are separated according to their charge by isoelectric focusing. Suitable precasted polyacrylamide gel strips containing an immobilised pH gradient (Bjellquist *et al.*, 1982) are available from several commercial suppliers. In the second dimension, that is carried out in right angle to the first dimension, the pre-separated proteins are further separated according to their molecular masses by SDS-PAGE (figure 5.4).

After staining, the spots can be excised from the gel, digested with the endoproteinase trypsin and the resulting peptides can be characterised by mass spectrometry. Proteins are then identified by comparing the mass of the peptide fragments with predictions from either mRNA sequences or protein databases (Yates, 1998).

Limited resolution is currently the main drawback of two-dimensional PAGE. Resulting spots are visualised by staining. Coomassie Blue facilitates only the visualisation of high abundant proteins, *i.e.* 10^5 copies per cell or approximately ten nanograms per square millimetre gel. Silver staining by contrast visualises about 0.01 nanograms per square millimetre gel (10^3 copies per cell). The most sensitive detection method is immunoblotting that may detect as little as ten protein copies per cell. The usefulness of immunoblotting in protein profiling is limited as it is a highly sequence specific procedure. A new generation of fluorescent

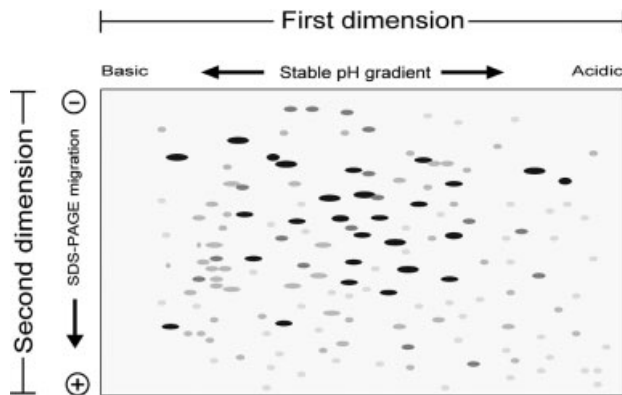


Figure 5.4 Two-dimensional polyacrylamide-gel electrophoresis (2D-PAGE). In the first dimension, polypeptides are separated according to their charge using isoelectric focusing (IEF). The following second dimension separates polypeptides according only to their molecular masses by sodium dodecyl sulphate-based polyacrylamide gel electrophoresis (SDS-PAGE).

dyes called SYPRO (SYPRO is an artificial word; S and Y are the initials of the person that invented the dye and PRO comes from protein) does not bind directly to the protein but to the protein-associated SDS and provides sensitivities comparable to those of silver staining (Berggren *et al.*, 2001). Even the most sensitive staining method available only allows the visualisation of about 10,000 spots per gel. Vuong *et al.* (2000) postulated that the detectable proteins are species represented with more than 1,000 copies per cell. Assuming that a single cell contains approximately 10^5 different mRNA molecules each of them encoding for in average five different proteins one can calculate that the 10,000 spots which are maximally visualised on a two-dimensional gel represent only about one fifth of the potential number of gene products (Vuong *et al.*, 2000). Staining methods must therefore be developed that allow the detection of low abundance proteins. The application of narrow pH range ultrazoom gels can enhance the resolution and improve the detection of low abundance proteins (Hoving *et al.*, 2000).

The methods mentioned and described so far can be summarised as the classical tools used for the analysis of differential protein expression.

5.2.2.4 Two-dimensional difference gel electrophoresis (2D-DIGE)

Two-dimensional difference gel electrophoresis (2D-DIGE) offers an improved comparability of quantitative changes of protein expression patterns in different samples. In principle, differences between two protein samples are detected by fluorescently tagging the samples with different dyes having distinguishable emission maxima as well as by co-separation and visualisation in a single conventional two-dimensional gel (Unlu *et al.*, 1997). Differences between the samples are analysed by overlaying the two false colour marked fluorescent channels.

5.2.3

High-performance liquid chromatography (HPLC)

High-performance liquid chromatography (HPLC) is a popular technique for the analysis of proteins and peptides because it is easy to use and is not limited by volatility or stability of the sample compound. HPLC was developed in the mid-1970s and quickly improved with the development of column packing materials and the additional convenience of online detectors. Computers and automation added to the convenience of HPLC during the last two decades. The past decade has seen a vast undertaking in the development of micro-columns and other specialised columns for separation. The dimensions of a typical HPLC column are at least 100 millimetres or more in length with an internal diameter of between three and five millimetres.

In principle, a sample is injected upstream of the column as part of the mobile phase and is pumped over a column that is composed of a solid phase and a mobile phase with a constant flow rate. Downstream of the column, the separated sample compounds are detected online using ultraviolet absorption, fluorescence or mass spectrometry and the eluate is collected in fractions (figure 5.5).

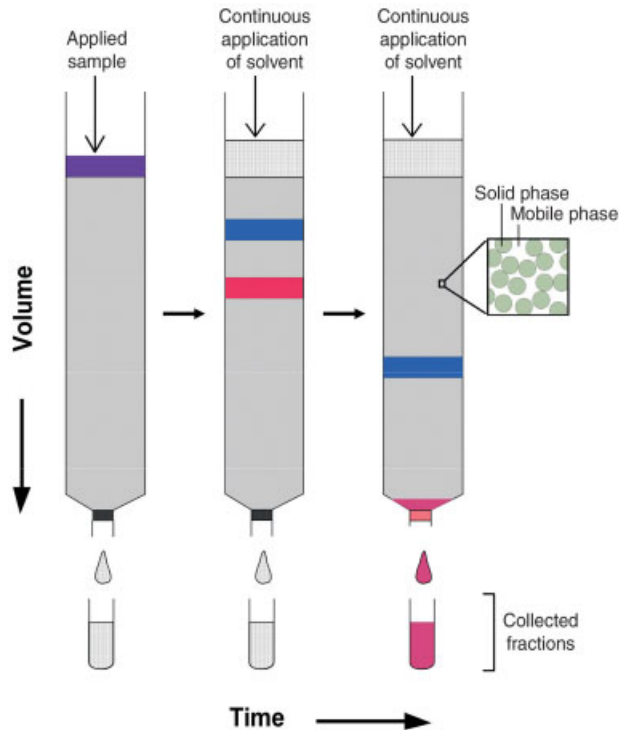
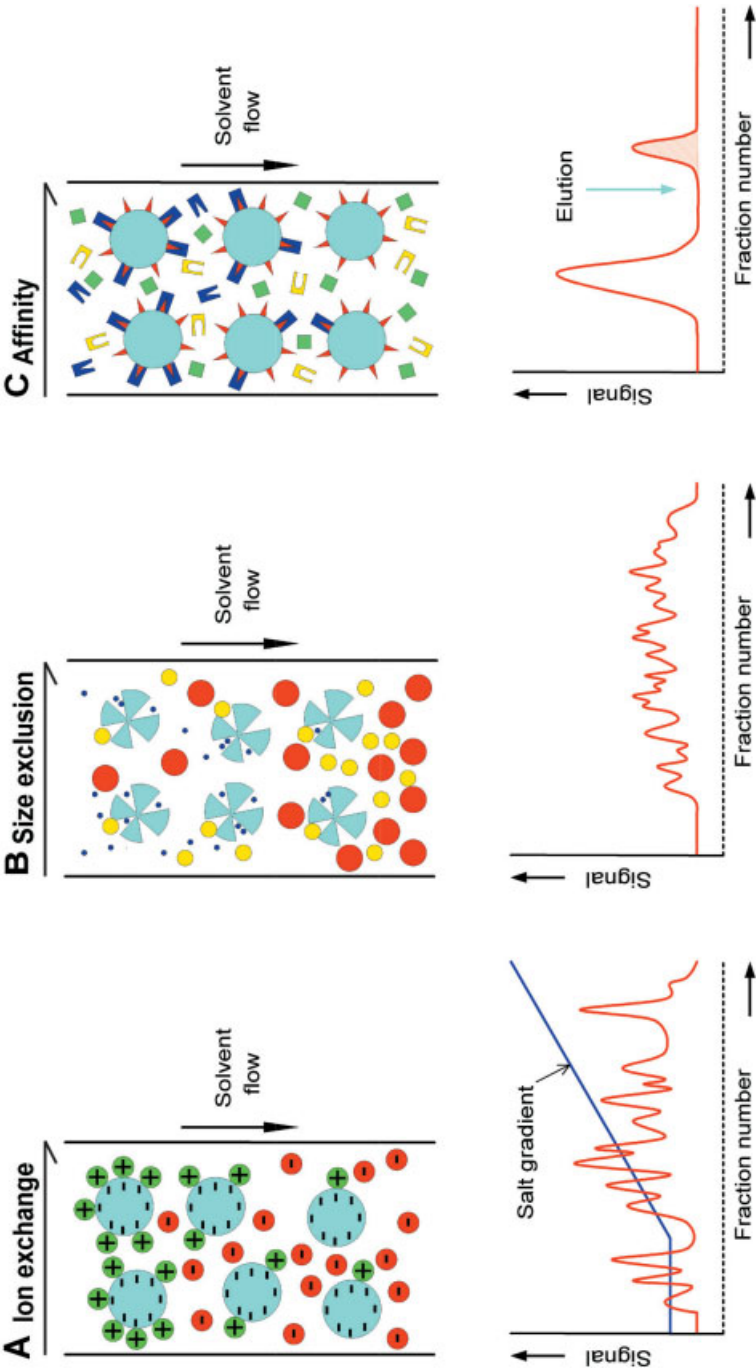


Figure 5.5 Column chromatography. A cylindrical glass, plastic, or metal column is filled with a permeable solid matrix (solid phase) that is immersed in solvent (mobile phase). The sample is applied to the top of the column.

Large amounts of solvent are pumped (or flow by gravity) through the column. Different molecule species are separated according to different migration rates through the column and are fractionally collected in different tubes.

The main advantages of HPLC as compared to previously used methods such as open-column chromatography, paper chromatography or thin-layer chromatography are velocity, reproducibility and high resolution.

The choice of solid and mobile phase determines the success of the separation effort. The solid phase in HPLC refers to the solid support contained within a column over which the mobile phase containing the sample flows. As the sample flows across the stationary phase, migration of the sample components is affected by non-covalent interactions of the compounds with the stationary phase. Samples that interact more strongly with the stationary phase than with the mobile phase will have a longer retention time and *vice versa*. The separation principles employed in the stationary phase include affinity chromatography, size exclusion chromatography and ion exchange chromatography. In addition, a polar solute may be used with a apolar stationary phase (so-called reverse phase chromatography), or an apolar solute may be used with a polar stationary phase (normal phase chromatography) (figure 5.6).



◀ **Figure 5.6 Different methods used in column chromatography.** **A)** ion exchange chromatography (IEC) means the reversible electrostatic binding of solubilised ions to ions of the opposite charge that are covalently bound to the solid matrix of the column. As an example, the principle of cation exchange is shown. While cations bind to the matrix anions do not (upper panel). The strength of the association between the cations and the matrix depends on the ion strength and the pH of the mobile phase. The lower panel shows the elution of bound polypeptides with a linear salt gradient. The increasing concentration of concurrent cations causes the elution of the more weakly bound polypeptides first while stronger bound polypeptides dissociate and elute later in the order of their increasing affinity to the matrix of the column. **B)** In size exclusion chromatography (SEC; also called gel filtration), the sample passes an inert matrix that is porous. Small molecules that enter the pores completely are

the most retarded and delayed. Depending on the pore size of the matrix, larger molecules are retarded to a lesser extent or not at all and migrate through the column faster (upper panel). The elution of a SEC column (lower panel) uses the same solvent that the sample was dissolved in. No further component is required. **C)** Affinity chromatography makes use of a highly specific interaction between a molecule covalently bound to the inert matrix and the molecule of interest in the applied sample (upper panel). The matrix-bound molecular species may for instance be an antibody or a natural binding partner of the protein that has to be isolated. The high affinity makes it possible to wash away undesired components efficiently after the specific binding has taken place. Subsequently, purified protein can be eluted from affinity chromatography columns using either an extreme pH value or highly concentrated salt solutions (lower panel).

5.2.3.1 Ion exchange chromatography (IEC)

Ion exchange chromatography (IEC) operates on the basis of selective exchange of ions in the sample with counterions in the stationary phase (figure 5.6.A). IEC is performed with columns containing charge-bearing functional groups attached to a polymer matrix. The functional ions are permanently bonded to the column and each has a counterion attached. The sample is retained by replacing the counterions of the stationary phase with the ionised sample. The sample is eluted from the column by changing the properties of the mobile phase so that the mobile phase will displace the sample ions from the stationary phase, *i.e.* by changing the ion strength or the pH. The use of gradient elution is favourable because of equivalent resolution as polyacrylamide gel electrophoresis and increased loading capacity when compared to size exclusion chromatography. Originally reserved for the analysis of proteins, IEC has become a useful tool in peptide analysis. The development of strong cation-exchange (SCX) columns allows the efficient retention and separation of peptides (Crimmins, 1994).

5.2.3.2 Size exclusion chromatography (SEC)

Size exclusion chromatography (SEC) operates by separating compounds on the basis of their size. The stationary phase consists of porous beads. The larger compounds are excluded from the interior of the beads and, thus, elute first. The small compounds enter the beads and elute according to their ability to exit through the bead pores (see figure 5.6.B). The column may be either silica or non-silica-based.

The use of SEC (as well as IEC) is well-suited for use with biologically active proteins, such as enzymes, hormones, and antibodies, since each protein has its own unique structure and these techniques may be performed under physiological conditions. Full recovery of activity after exposure to the chromatography may be achieved, and availability of SEC columns is diverse enough to allow fractionation from ten to one thousand kilodaltons. Extremely basic or hydrophobic proteins may not exhibit true SEC character since the columns tend to have slight hydrophobicity and an anionic character (Barth *et al.*, 1994).

5.2.3.3 Affinity chromatography

Affinity chromatography operates by using immobilised biochemicals that have a specific affinity to the compound(s) of interest (figure 5.6.C). Separation occurs as the mobile phase and sample pass over the stationary phase. The sample compound of interest is retained as the rest of the impurities and mobile phase pass through the column. The compound is then eluted by changing the ion strength, pH, temperature and other characteristics of the mobile phase. Affinity chromatography is therefore the method of choice to separate a specific protein or a group of proteins sharing a common affinity characteristic (Labrou & Clonis, 1994).

5.2.3.4 Reverse phase chromatography (RPC)

Reverse phase chromatography (RPC) operates on the basis of hydrophilicity and lipophilicity. The stationary phase consists of silica-based packings with n-alkyl chains covalently bound. For example, the octadecyl ligand that is often used is symbolised by C-18. The more hydrophobic the matrix on each ligand, the greater is the tendency of the column to retain hydrophobic moieties. Thus, hydrophilic compounds elute more quickly than hydrophobic compounds.

The use of RPC revolutionised peptide purification. Crude tissue extracts may be loaded directly onto the RPC system and mobilised by gradient elution although it is usual to perform fragmentation prior to RPC. A common mobile phase for RPC of peptides is a gradient of 0.1 percent of trifluoroacetic acid in water to 0.1 percent trifluoroacetic acid in an organic solvent such as acetonitrile (Regnier, 1983).

5.2.3.5 Normal phase chromatography

Normal phase chromatography operates on the basis of hydrophilicity and lipophilicity by using a polar stationary phase and a less polar mobile phase. Thus, hydrophobic compounds elute more quickly than hydrophilic compounds.

5.2.4

Capillary electrophoresis (CE) and capillary electrochromatography (CEC)

Like HPLC, capillary electrophoresis (CE) and capillary electrochromatography (CEC) are powerful tools for peptide and protein sample separation (Pretorius *et al.*, 1974; Wu *et al.*, 1997). The separation of compounds in CE relies on the same principles as in conventional electrophoresis techniques, but the separations take place within a silica capillary (column) with only 25 to 50 micrometres in diameter. The driving force of ion separation in CE is the applied voltage of up to 35 kilovolt. This high voltage causes partial deprotonation of the silica capillary wall, which ultimately results in the electroosmotic flow (EOF) of ions (Bartle & Myers, 2001). While CE separates sample compounds in buffer, in CEC the capillary column contains an additional stationary phase, that makes most conventional slab gel and HPLC techniques (*e.g.*, IEF, SDS-PAGE, affinity chromatography) accessible to the capillary system.

One advantage of CE and CEC as compared to HPLC is the necessity for only very small sample volumes of a few nanolitres. Another advantage is that the EOF in CE and CEC causes a nearly flat band profile while HPLC may show band broadening effects caused by transchannel diffusion in the broader column. In two-dimensional gel electrophoresis, proteins are first resolved, then digested with trypsin and the peptide fragments are analysed using mass spectrometry. HPLC and CE/CEC techniques have in common that the proteins are digested first and then separated afterwards. The advantage is that peptides are more soluble and easier to separate than the parent proteins, especially when there are hydrophobic and membrane proteins. The disadvantage is that tryptic digests may easily contain a million peptides to be analysed (Regnier *et al.*, 2001). No single chromatographic or electrophoretic method is capable of separating such a complex mixture.

Excellent overviews about the HPLC of proteins and the separation of biomolecules are given by Regnier (1983) and Turkova (1999), respectively.

5.2.5

Multi-dimensional separations

Each human cell contains at least 10,000 proteins at any given time (Celis & Gromov, 1999). The concentrations of these proteins vary with a dynamic range of approximately five-fold (Yates, 2001). Two-dimensional PAGE, the current method of choice to investigate whole proteomes, can resolve up to 10,000 spots with the disadvantage that low abundance proteins and hydrophobic proteins are poorly or not represented (see above). As no single chromatographic or electrophoretic method is likely to resolve such a complex mixture of a cell's or tissue's proteins, a multi-dimensional method which employs two or more orthogonal separation techniques (*i.e.* conventional slab gel electrophoresis, HPLC or CE) or separation methods with different mechanisms of separation (ion exchange, size exclusion, affinity chromatography, *etc.*) will significantly improve the chances of resolving such a complex

mixture of cell proteins into its individual components. All of the above mentioned separation techniques and mechanisms can be combined theoretically in multi-dimensional approaches, each with its own advantages. For example, the advantages of HPLC-combined CE multi-dimensional approaches are that they can be automated, are sensitive, reproducible, fast and quantitative. An overview of multi-dimensional approaches has been provided by Issaq *et al.* (2001).

5.2.6

New approaches

5.2.6.1 Isotope-coded affinity tag method (ICAT™)

The isotope-coded affinity tag method (ICAT™) is a new strategy for quantifying differential protein expression (figure 5.7). Two different protein mixtures are labelled, each with a different derivative of the ICAT™ reagent. The ICAT™ reagent consists of three structural elements: (i) Biotin as an affinity tag, which is used to isolate ICAT™-labelled peptides. (ii) A linker that can incorporate two stable isotopes, *e.g.*, hydrogen in light ICAT™ reagent (d0-ICAT) and deuterium in heavy reagent (d8-ICAT). (iii) A reactive group specific toward thiol groups, *i.e.*, to cysteine residues of proteins and peptides. The protein samples are differentially labelled by virtue of the differences that exist in their cysteine residues. The samples are then

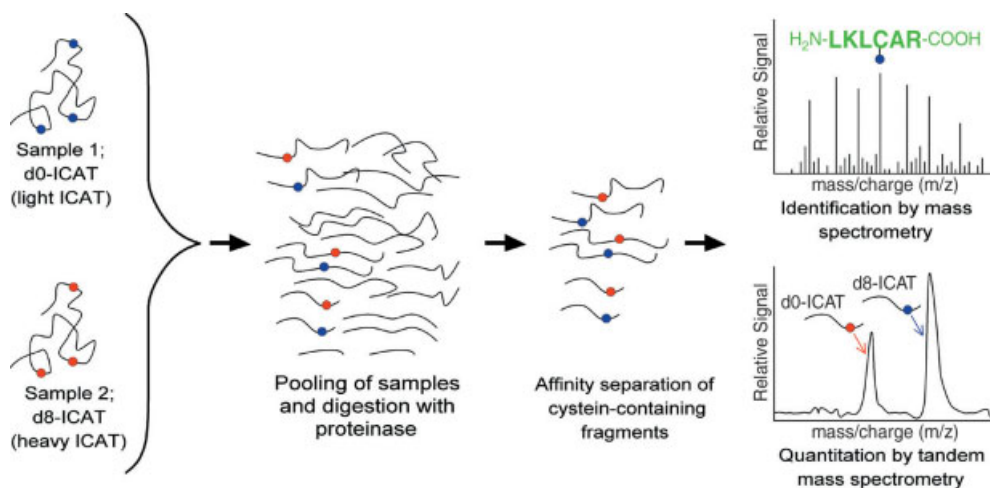


Figure 5.7 Principle of the isotope-coded affinity tag (ICAT™) method. Two samples are covalently labelled *via* their cysteine residues, one with the light hydrogen-containing ICAT reagent, and the other with the heavy deuterium-containing ICAT reagent. The two samples are pooled in a 1:1 ratio and polypeptides in the pool are subjected to an endoproteinase digest. The resulting fragments containing an ICAT™ label are separated from unlabelled fragments

by affinity chromatography. As the labelling with heavy ICAT™ reagent gives a higher molecular mass to a certain peptide in sample 2 as compared to the same peptide in sample 1, the relative amount of each peptide in the two samples can be determined by tandem mass spectrometry (MS/MS). The identification of the respective protein by mass spectrometry (MS) delivers the amino acid sequence.

combined and fragmented by means of trypsin digestion. The modified peptides are then isolated by affinity chromatography using immobilised avidin. The mixture is analysed by mass spectrometry delivering the ratio of the ion intensities for an ICAT™-labelled pair and therefore quantifies the relative abundance of its parent protein in the original samples. In addition, a tandem mass spectrum reveals the sequence of the peptide and unambiguously identifies the protein. This strategy results in the quantification and identification of all protein compounds in a mixture. Theoretically, ICAT™ is applicable to protein mixtures as complex as the entire proteome (Han *et al.*, 2001).

5.2.6.2 Multi-epitope ligand-Kartographie (MELK)

Although it is itself not a method to separate proteins, or peptides, the proprietary topological proteomics technology MELK (Multi-epitope ligand-Kartographie; Mel-Tec GmbH, Magdeburg, Germany) is mentioned here, because of its innovative

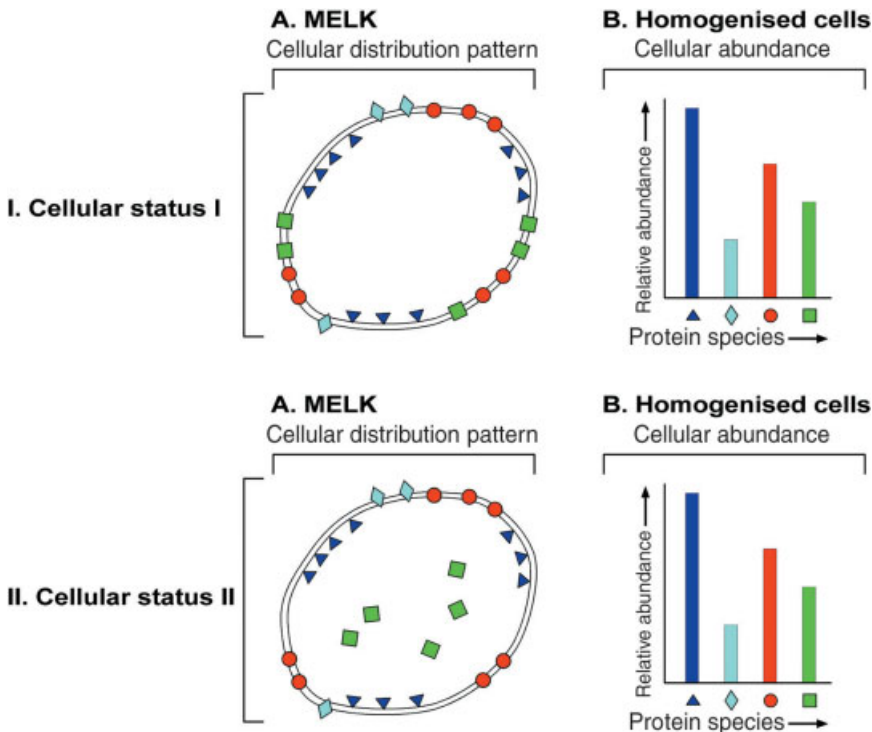


Figure 5.8 Multi-epitope ligand-Kartographie (MELK). In contrast to conventional methods used to investigate proteomes, MELK not only determines the cellular abundance of a selected protein species but also allows visualisation of its cellular distribution. The figure schematically shows that MELK allows the differential locali-

sation of four chosen proteins in the cellular status I and II of intact cells **A**), respectively, while conventional methods using homogenised cells **B**) only determine the (in this case) unchanged abundance of these proteins in status I and II of the cells.

character (figure 5.8). In a high-throughput process suitable for whole cell protein fingerprinting, MELK permits the deciphering of protein networks, their protein components and function in single intact cells delivering three-dimensional distribution patterns of marker proteins in cells to be compared. The technique is capable of detecting and tracking selected proteins independently of their biochemical properties, expression levels, or localisation. A fundamental advance of MELK is that proteomes of intact cells are analysed where context-depending protein information is preserved and expression profiles of selected proteins can be compared *in situ*. These topological information gets lost when cells are homogenised as it is usual for other applications in proteome analysis (Schubert, 2000).

5.2.7

Summary

For the last four decades, parallel developments in conventional slab gel electrophoresis and in HPLC/CE techniques have contributed enormously to the science of protein separation. Individual methods within these techniques are well suited for separating single proteins or groups of proteins. However, the separation of a whole proteome from a cell or tissue is an ambitious task, and, currently, there is no established method available to meet this challenge. In practice, three different approaches are actually used to deal with whole proteomes. The first strategy is to improve the most common conventional method of choice, *i.e.* 2D-PAGE. This method is fundamentally lacking in speed, sensitivity, reproducibility, and precision. The second strategy is to combine two or more of the established methods of separation successfully in so-called multi-dimensional approaches in order to achieve the desired effect. So far, this path has been used to probe many complex questions. A third strategy is to develop new techniques for comparing different proteomes. ICAT is and MELK seems to be a promising new tool for taking up the great challenge of proteomics.

5.3

Detection and identification of proteins by mass spectrometry

5.3.1

Introduction

Mass spectrometry (MS) is one of the most important tools in proteomic analyses today. It has moved out of the physicist's basement into the biological laboratory. Companies and technology platforms offering protein mass spectrometry continue to be founded and funded. In this section, the reader will be introduced to this technique stressing protein and peptide analysis, although mass spectrometry is of equal importance for the study of other biomolecules such as DNA or low-molecular weight metabolites. The focus of the following sections will be on what is understood as proteomic mass spectrometry: the analysis and identification of pro-

teins separated by gel electrophoresis, although the application of mass spectrometry in protein analysis is much wider and includes the measurement of intact proteins and their complexes. A distinction should be drawn between the huge demand for proteomic analyses of known proteins, which has led to the commercialisation of the method, and special applications from biomedical research such as *de novo* sequencing, the analysis of post-translational modifications, or cross-linking studies, where mass spectrometry is also of enormous value. The latter methods will be described only briefly and the reader is referred to the literature for further reading.

Since the 1960s, when derivatisation was still necessary in order to measure peptides with mass spectrometry, solutions were found to the problem of ionising huge molecules such as proteins and detecting them using mass spectrometry. Increasingly, mass spectrometry instrumentation has been adapted to demands of proteomic analyses and spectrometers have become benchtop size and much more user-friendly (table 5.2).

Table 5.2 History of mass spectrometry of biomolecules. For more detail, see the 50 year commemorative book review of the American Society for Mass Spectrometry from 2002 and the internet site 'A history of mass spectrometry' at <http://masspec.scripps.edu/hist.html>.

Year	Technical acquirement
1960s	Analysis of volatile compounds only.
1965	Analysis of permethylated peptide. At 1.4 kilodalton, largest natural compound analysed by mass spectrometry.
1974	Mass spectrometry of large underivatised peptides becomes possible.
1982	First complete protein spectrum obtained (insulin).
1983	Analysis of snake toxins of 13 kilodaltons. Analysis of proteins above 25 kilodaltons not possible. Sensitivity in low to mid-picomolar range.
After 1988	Analysis of almost any biomolecule possible.

5.3.2

How does mass spectrometry work?

Mass spectrometers measure molecular weight. To do this, molecules are first ionised and then separated according to their mass-to-charge ratio (m/z). Each atom has a characteristic mass and the symbol u (mass unit or unified atomic mass unit) represents by convention one twelfth of the mass of the most abundant naturally occurring stable isotope of carbon. The mass of a molecule is the sum of the atomic masses (in Dalton) of all the atoms of the elements composing it. The unit Dalton is widely used in biochemistry and is also accepted in the field of mass spectrometry. (Mills *et al.*, 1993).

5.3.2.1 Ionisation of molecules

Ionisation is the first step in mass spectrometric analysis of molecules. Although several other ionisation techniques are still in use, matrix-assisted laser-desorption/ionisation (MALDI) and electrospray ionisation (ESI) are the main techniques employed in proteomics. Other ionisation techniques are atmospheric pressure chemical ionisation, electron impact, chemical ionisation, fast-atom bombardment, and thermospray.

In MALDI, the sample is spotted together with a matrix onto a metal target and dried. The sample holder is then loaded into the mass spectrometer. Ionisation is achieved by directing a pulsed laser beam onto the sample. Most ions detected carry a single charge (see figure 5.9).

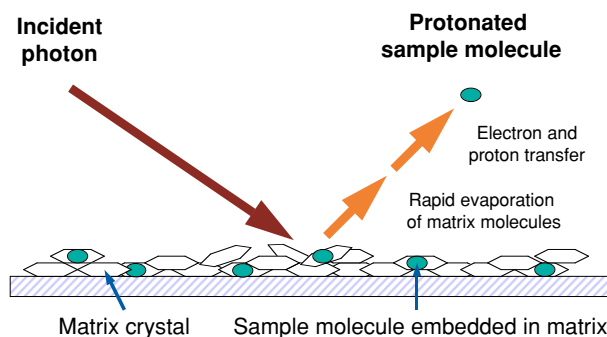


Figure 5.9 Principle of matrix-assisted laser-desorption/ionisation (MALDI). Biomolecules are co-crystallised with a matrix on a metal

surface. Irradiation with a laser beam causes ionisation, thus allowing detection and identification of the resulting ionised fragments.

The matrix is a low-molecular weight compound, that can absorb some of the energy from the laser light (for instance, ultraviolet light). This prevents decomposition of the sample and also facilitates vaporisation and ionisation of the analyte molecules (figure 5.10).

In electrospray ionisation, a liquid sample is exposed to a strong electric field. Ionisation is achieved for instance by applying a high voltage of up to six kilovolts directly to the solution *via* a metal wire placed inside a capillary. Highly charged droplets are formed, which are desolvated by heat or gas. In contrast to MALDI, ESI creates more ions which bear a multiple charge. Intact protein can take up as many as 50 protons so that their spectra show a distribution of charged state (figure 5.11).

5.3.2.2 Detection of ions

Ions are separated in electric and magnetic fields in a high vacuum. Modern mass spectrometers are based on a variety of technical designs and show differences in mass accuracy or resolution. The most important features for proteomics are to re-

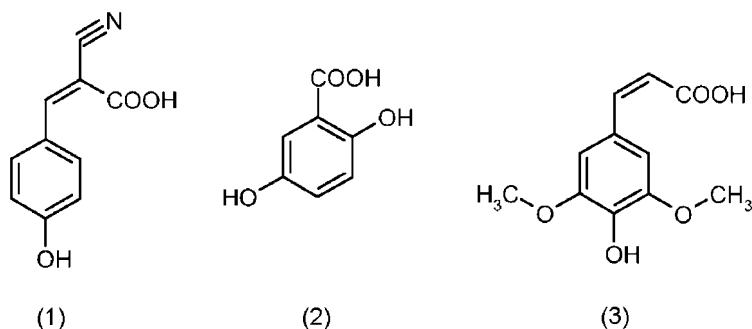


Figure 5.10 Matrices used for matrix-assisted laserdesorption/ionisation (MALDI) analysis in proteomics. α -Cyano-4-hydroxycinnamic acid (1) is mainly used for the analysis of peptides. Gentisic acid (2,5-dihydroxybenzoic acid) (2) shows advantages for some application such as the analysis of phosphopeptides and sinapinic acid (3,5-dimethoxy-4-hydroxycinnamic acid) (3) is often used for the analysis of proteins.

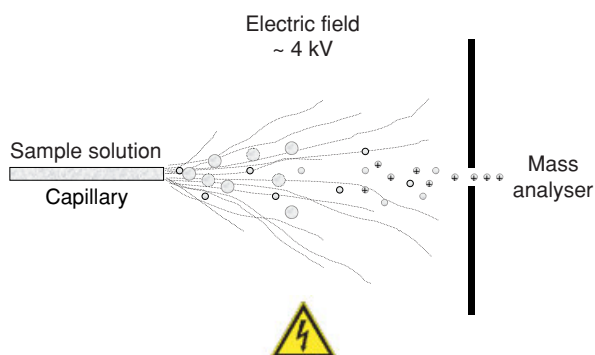


Figure 5.11 Electrospray ionisation (ESI). Ions are generated exposing the liquid sample to a high electric field.

solve the isotopes of peptides and to detect them with an error as small as possible (routinely better than 50 parts per million).

Ion separation can be described by the following fundamental equation

$$E_{kin} = z e V = \frac{1}{2} m v^2$$

where E_{kin} is the kinetic energy of the ions, V is the accelerating voltage, z is the number of elementary charges e ($1.60217733 \times 10^{-19}$ Coulomb), m is the molecular mass of the ions, and v is the velocity of the ions. This means that ions in electric and magnetic fields in a high vacuum are separated according to their mass-to-charge ration.

Technical solutions for ion detection mainly used in proteomics are ion traps, time-of-flight (TOF), and quadrupoles, whereas detection in other analytical fields is mainly performed by quadrupole systems, double-focusing sector, and Fourier transform-ion cyclotron resonance (FT-ICR).

5.3.2.3 Appearance of spectra

Modern instruments can routinely determine monoisotopic peptide masses at high accuracy. This is necessary to search databases effectively. For intact proteins, usually their average mass is measured, although instruments with very high resolving power such as Fourier transform-ion cyclotron resonance instruments have the capability to resolve individual peaks (figures 5.12 and 5.13).

Most elements found in biological compounds (^{12}C , ^1H , ^{14}N , ^{16}O) have low-abundant isotopes (^{13}C , ^2H , ^{15}N , ^{17}O , ^{18}O). It is important to realise that for masses larger than two kilodaltons the ^{12}C isotope is no longer the most abundant peak in an isotopic distribution of a biomolecule. Isotopes of other elements alter the isotopic distribution significantly (*i.e.* bromine and chlorine; see table 5.3).

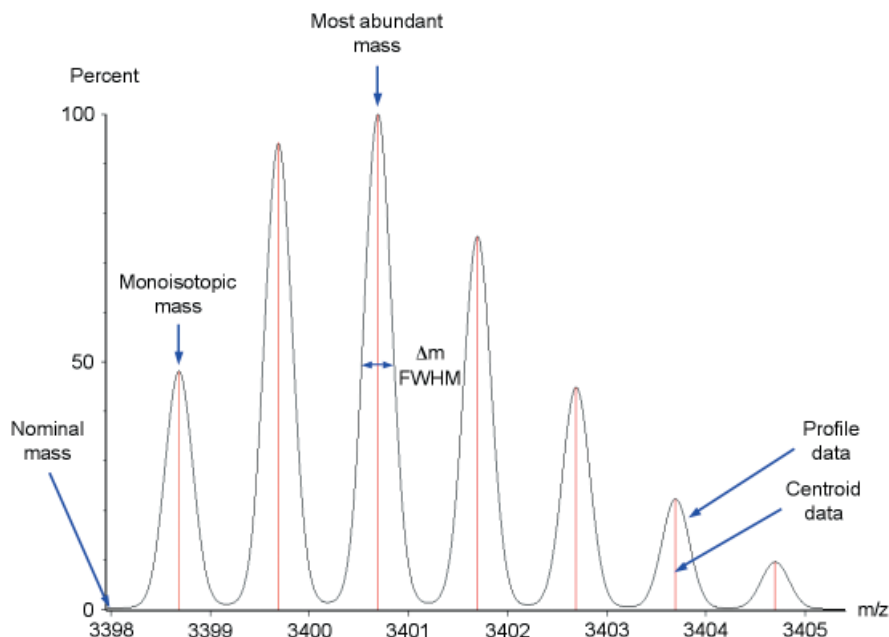


Figure 5.12 Resolving power of mass spectrometers: $m/\Delta m$. Instruments with high resolving power can separate the isotope peaks of molecules with high molecular weight. Shown is a spectrum of the protonated insulin B-chain with

a chemical composition of $\text{C}_{157}\text{H}_{233}\text{N}_{40}\text{O}_{41}\text{S}_2$. The following abbreviations are used in the figure: full width at half maximum, FWHM; mass, m ; charge, z).

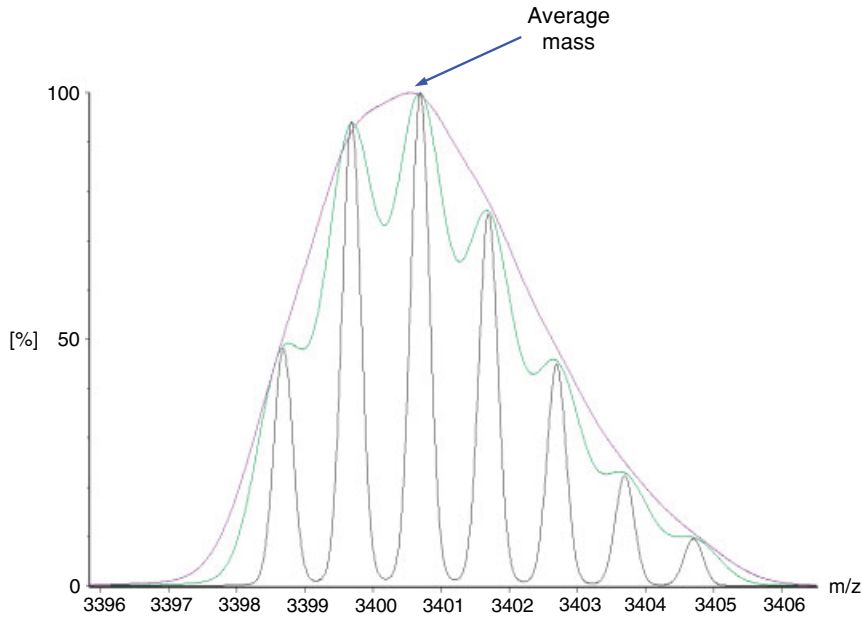


Figure 5.13 Average mass. For large molecules such as proteins the isotopes may not be resolved in mass spectrometric detection and only the average mass is measured.

Table 5.3 Atomic masses and abundances for biologically relevant isotopes (Rosman & Taylor, 1998).

<i>Isotope</i>	<i>Atomic weight</i>	<i>Natural abundance [%]</i>
^{12}C	12.0107	98.93
^{13}C	13.003354826	1.07
^1H	1.007825035	99.9885
^2H	2.014101779	0.0115
^{14}N	14.003074002	99.632
^{15}N	15.00010897	0.368
^{16}O	15.99491463	99.757
^{17}O	16.9991312	0.038
^{18}O	17.9991603	0.205
^{31}P	30.9737620	100
^{32}S	31.97207070	94.93
^{33}S	32.97145854	0.76
^{34}S	33.96786665	4.29
^{36}S	35.96708062	0.02

5.3.2.4 Modes of operation

Mass spectrometers can be operated in many ways in order to solve a specific problem. On one hand, the molecular weight of the analyte molecules can be determined. In this case, every compound present in the sample should in principle generate a separate peak. This is not possible, however, because factors such as ion suppression, adduct formation, and instrument settings restrict mass spectrometric output. For this reason, sample preparation and purification are very important and sample preparation has often to be adapted to the needs of mass spectrometry.

An important feature of some types of mass spectrometers is their capability to select and to fragment certain ions by collision in the gas phase. In this way, peptides can be sequenced. The product ions generate characteristic spectra, which can be searched against databases (figures 5.14 and 5.15, table 5.4).

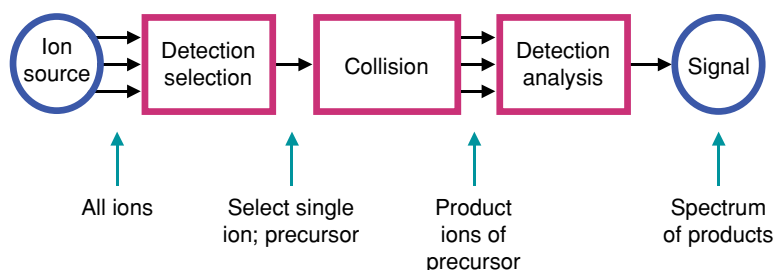


Figure 5.14 Tandem mass spectrometry (tandem MS) allows fragmentation and selected separation of biomolecules. Special spectro-

meter designs allow the selection of ions of interest and their fragmentation. In that way, the structure of the molecule can be determined.

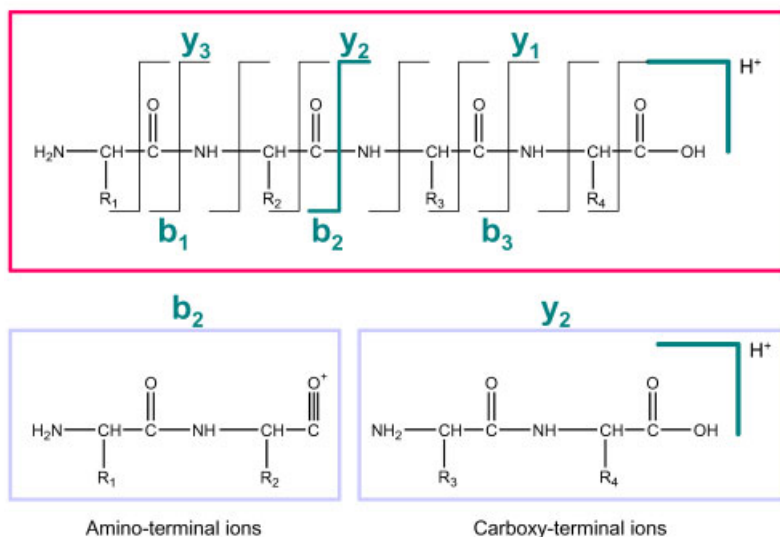


Figure 5.15 Peptide fragment ions. Collision-induced dissociation of molecules predominantly causes fragmentation at the amide

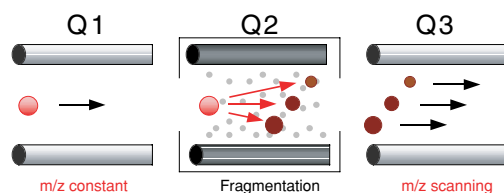
bonds of the polyamide backbone. Ion series b and y have been defined to describe cleavage from the carboxy or amino-termini, respectively.

Table 5.4 Amino acid residues and their mass.

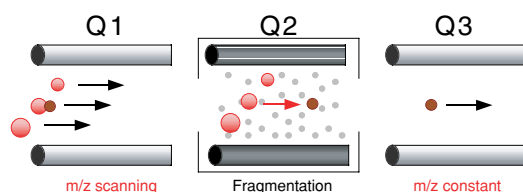
<i>Amino acid</i>	<i>Abbreviation</i>	<i>Symbol</i>	<i>Monoisotopic mass [g/mol]</i>	<i>Average mass [g/mol]</i>
Glycine	Gly	G	57.021	57.052
Alanine	Ala	A	71.037	71.079
Serine	Ser	S	87.032	87.078
Proline	Pro	P	97.053	97.117
Valine	Val	V	99.068	99.133
Threonine	Thr	T	101.048	101.105
Cysteine	Cys	C	103.009	103.145
Isoleucine	Ile	I	113.084	113.160
Leucine	Leu	L	113.084	113.160
Asparagine	Asn	N	114.043	114.104
Aspartic acid	Asp	D	115.027	115.089
Glutamine	Gln	Q	128.059	128.131
Lysine	Lys	K	128.095	128.174
Glutamic acid	Glu	E	129.043	129.116
Methionine	Met	M	131.040	131.199
Histidine	His	H	137.059	137.142
Phenylalanine	Phe	F	147.068	147.177
Arginine	Arg	R	156.101	156.188
Tyrosine	Tyr	Y	163.063	163.176
Tryptophan	Trp	W	186.079	186.213

Further options available in triple-quadrupole spectrometers are the neutral loss scan and the parent ion scan which are of importance for monitoring specific ions. This is important, for example, in phosphorylation analysis (figure 5.16).

Product ion scan



Precursor ion scan



Neutral loss scan

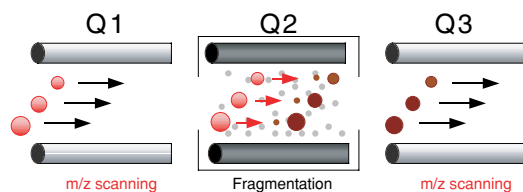


Figure 5.16 Modes of operation of a triple quadrupole instrument. Varying the direct current (DC) voltage or and a superimposed radio-

allows monitoring of product ions, precursor ions or neutral losses. See text for further details.

5.3.3

Proteomic mass spectrometry

The immediate goal in proteomics is the automated analysis and comparison of large sets of proteins. It is of advantage to digest the separated proteins of interest enzymatically, because (i) resulting peptides can be easily extracted and separated, (ii) a characteristic set of peptides is generated for each protein, and (iii) mass spectrometric determination of peptide masses (termed peptide mapping or peptide mass fingerprint) and sequence analysis allows protein identification.

The protease used predominantly is trypsin, but other cleavage methods may be more appropriate for specific analysis problems (for an overview for digestion methods, see table 5.5).

Table 5.5 Proteolytic agents.

	<i>Proteolytic agent</i>	<i>Cleavage site</i>
<i>Chemical</i>	BNPS skatol	Tryp...X
	Cyanogen bromide	Met...X
<i>Enzymatic</i>	Trypsin	Lys/Arg...X; X≠Pro
	Chymotrypsin	Phe/Tyr/Trp...X and other hydrophobic residues
	Endoproteinases	
	• Lys-C	Lys...X; X≠Pro
	• Glu-C	Glu...X
	• Asp-N	X...Asp
	Endopeptidase	
	• Asp-C	Asp...x
	Carboxypeptidase B	Carboxy-terminal residues, in particular Arg/Lys

Proteins may be enzymatically digested directly in the gel spot after SDS-based polyacrylamide gel electrophoresis. The resulting peptide is extracted after detection by silver or Coomassie staining that is generally sufficient for protein identification (figure 5.17). Silver staining of gels allows the detection of one to ten nanograms of protein, whereas Coomassie staining is less sensitive and allows the detection of 50 to 100 nanograms. Since it is possible to identify less than one picomole of a single protein using mass spectrometry, any bands detected by silver or Coomassie staining can be identified. Proteins may also be digested after electroblotting using a similar procedure.

The peptides are purified using solid-phase extraction and subjected to mass spectrometric analysis. In general, a MALDI peptide map is measured first. The list of monoisotopic peptide masses is used for a database search that will often result in the confident identification of the protein with high confidence. Because of the statistical basis of the search, the number of false-positive results can be quite high. Several factors may hinder protein identification: (i) modifications of the protein (see chapter 1), (ii) multiple proteins in a single spot, (iii) unusual contaminants, and (iv) difficulties in accessing database information due to properties of search algorithms or database entries.

It is therefore advisable to perform fragmentation analysis using electrospray ionisation tandem mass spectrometry (ESI-MS/MS). Sequence information gained for only one peptide may allow protein identification.

Detailed information on mass spectrometric analyses of proteins separated by polyacrylamide gel electrophoresis can be found in specialist literature: in-gel digest with spot excision is described by Rosenfeld *et al.* (1992) and Shevchenko *et al.* (1996), and digest of electro-blotted proteins with spot excision is referenced

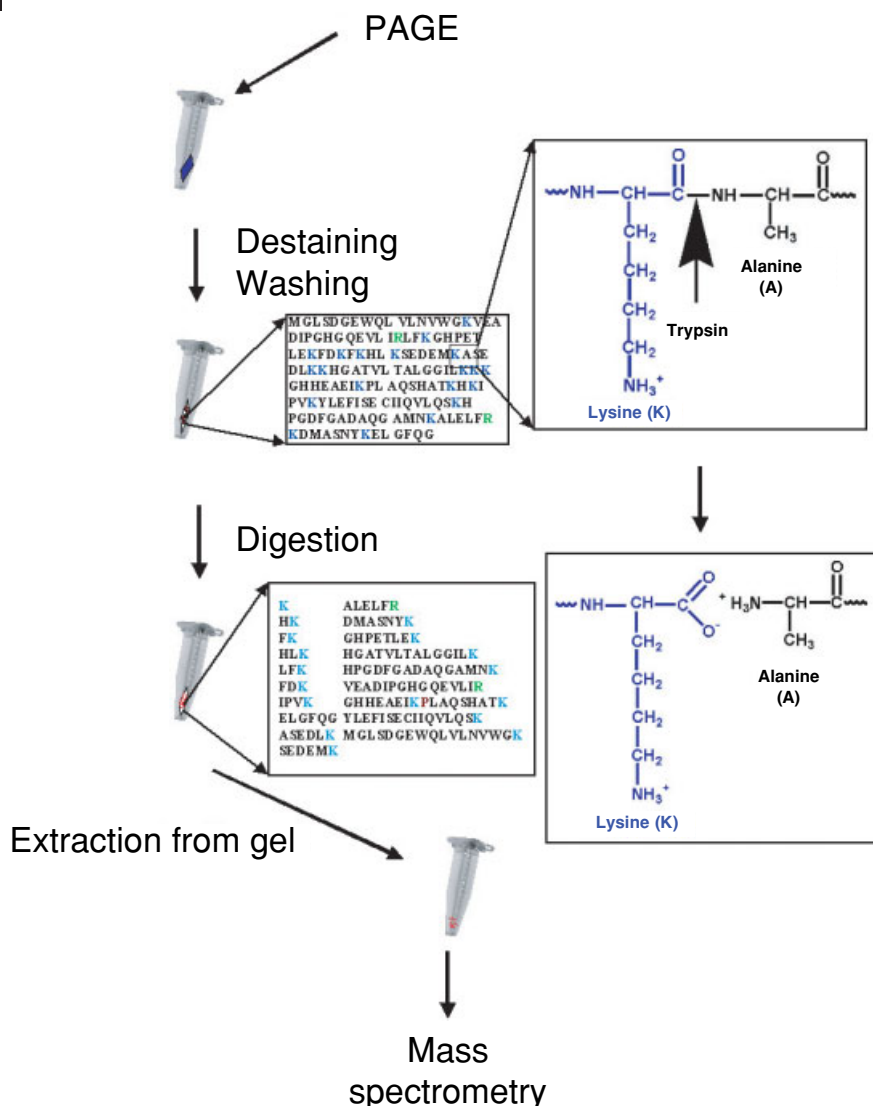


Figure 5.17 In-gel-digest of proteins separated by polyacrylamide gel electrophoresis (PAGE). Proteins can be digested enzymatically in the excised gel slice using, for example, trypsin. The resulting peptides are then extracted and purified for peptide mapping and sequencing.

by Fernandez *et al.* (1994) and Stults *et al.* (1996). The combination of in-gel and in-membrane digest without spot excision is described in detail by Binz *et al.* (1999).

Two-dimensional polyacrylamide gel electrophoresis is of limited use in the separation of very acidic or basic proteins, very large or small proteins, or membrane proteins. In such cases, the combination of high-performance liquid chromatogra-

phy and tandem mass spectrometry (HPLC-MS/MS) has been used. Using this approach, complex protein mixtures are digested and then automatically separated and fragmented by a high-performance liquid chromatography system that is coupled to a tandem mass spectrometer. Thus, a sequence of different chromatographic separation methods can be combined (*e.g.*, two-dimensional liquid chromatography, 2D-LC). Detailed information on mass spectrometric analyses of proteins separated by high-performance liquid chromatography can be found in Link *et al.* (1999), Peng & Gygi (2001) and Washburn *et al.* (2001).

Peptides are often separated with high-performance liquid chromatography using reversed phase material (silica particles covered with chemically-bonded hydrocarbon chains of varying length from four, C-4 column packing material, to eighteen carbons, C-18 column packing material). In HPLC-MS/MS experiments, both the ultraviolet absorption and the mass spectrometric signal (total ion current) are used for peptide detection. If their values reach preset values, they trigger the start of an tandem mass spectrometric experiment. In this way, more than 100 proteins have been identified in a single run (figure 5.18).

Peak lists generated with peptide mapping or tandem mass spectrometric fragmentation are the experimental data used to screen databases. Special search engines have been developed for this purpose that access the large public protein and genomic databases. The search programs differ in their algorithms and constraints.

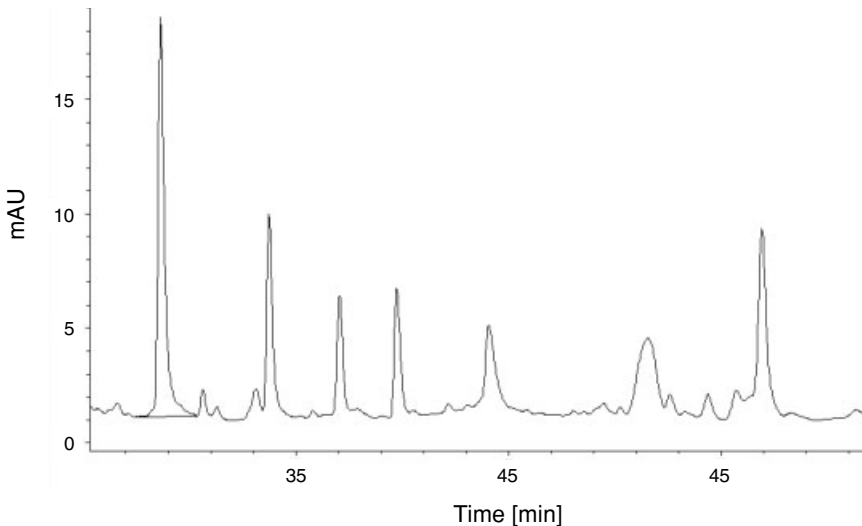


Figure 5.18 Ultraviolet chromatogram of a digested protein using nano-high-performance liquid chromatography (nano-HPLC). Ultraviolet signal at 214 nanometres of one picomole tryptic digest of myoglobin separated on re-

versed phase material. Online HPLC tandem mass spectrometry allows automatic fragmentation of many peptides. The following abbreviations are used: milli-absorption units, mAU; minutes, min.

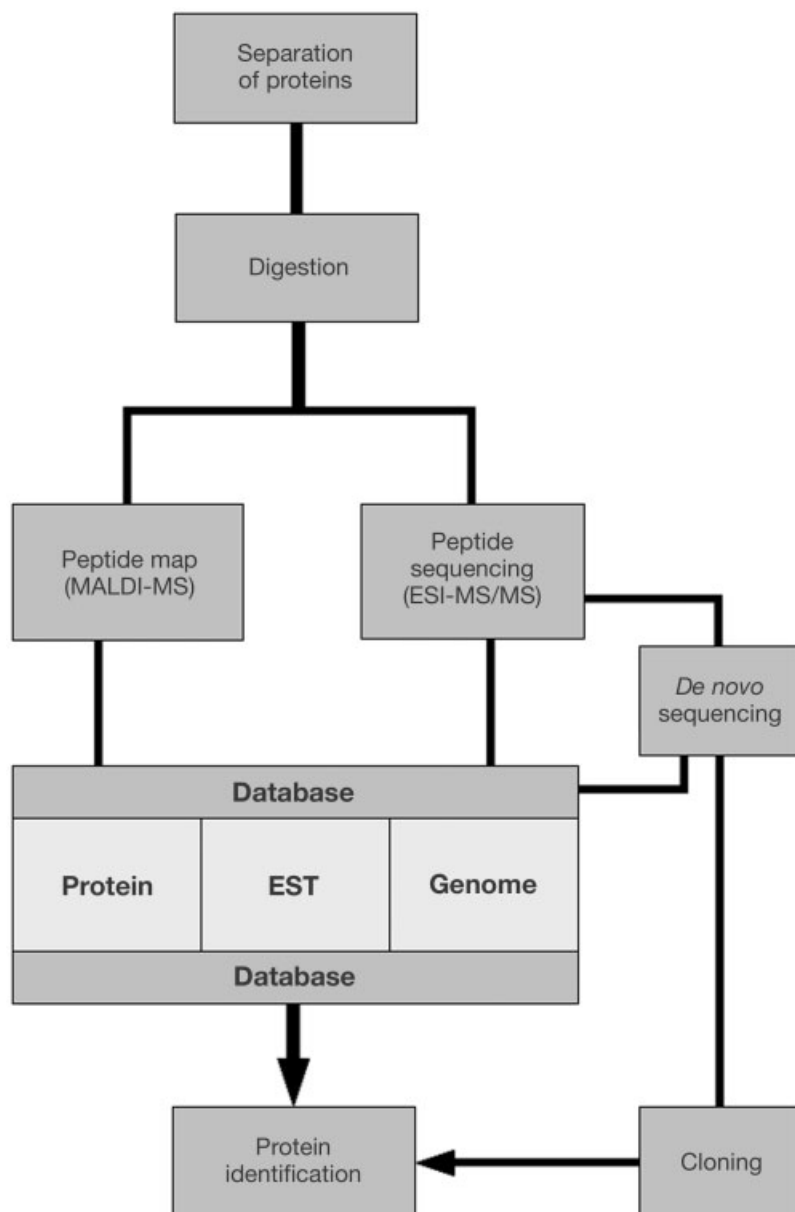


Figure 5.19 Flowchart of proteomic analyses using mass spectrometry. Routine protein identification is based on the digestion of electrophoretically separated proteins and the analysis of the peptides using mass spectrometry. Peptide masses and fragment data are used to search databases. Mass spectrometric data can provide sequence tags for unknown

proteins, thus enabling them to be isolated using classical cloning techniques. The following abbreviations are used: electrospray ionisation tandem mass spectrometry, ESI-MS/MS; expressed sequence tag, EST; matrix-assisted laser-desorption/ionisation mass spectrometry, MALDI-MS.

With this approach only known proteins can be found. Tandem mass spectrometric spectra of peptides from unknown proteins serve to find sequence tags, which can then help access the protein *via* classical cloning approaches (figure 5.19).

The following two sections describe some tools and databases which are useful for proteomic analyses. In this regard, the reader is also referred to chapter 7 that gives a broad overview of protein expression databases available in the internet.

5.3.3.1 Data analysis tools for mass spectroscopic experiments

Several software tools for mass spectrometric analyses are available *via* the internet. They allow to search databases with peptide mass fingerprinting or sequencing data. Short stretches of sequence (tags) can also be used as search input. Examples are:

Protein Prospector. This tool was developed by the San Francisco Mass Spectrometry Facility (University of California, California, USA) and which is publicly available at <http://prospector.ucsf.edu/> (Clauser *et al.*, 1999; Handley, 2002).

Mascot. This search engine is offered by Matrix Science Ltd. (London, United Kingdom) and is freely accessible at <http://www.matrixscience.com> (Handley, 2002; Perkins *et al.*, 1999).

PROWL. This tool is being developed in a collaborative fashion by ProteoMetrics, LLC, and the Rockefeller University (New York, New York, USA) and is provides at <http://prowl.rockefeller.edu/> (Fenyo *et al.*, 1996; Handley, 2002).

5.3.3.2 Proteomics servers and databases

Proteomic data are provided by several servers that are accessible *via* the internet. In the following paragraphs, some of the most popular servers are listed.

ExPASy (Expert Protein Analysis System). The ExPASy server of the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland) is dedicated to the analysis of protein sequences and structures as well as two-dimensional polyacrylamide gel electrophoresis. The server is available at <http://www.expasy.org/> and provides a broad range of useful proteomic tools and databases.

European Bioinformatics Institute. The European Bioinformatics Institute (EBI; Cambridge, United Kingdom) of the European Molecular Biology Laboratory (EMBL) is accessible at <http://www.ebi.ac.uk/> and manages databases of biological data including nucleic acid, protein sequences and biological structures as well as several tools of molecular biological analyses.

SWISS-PROT. SWISS-PROT at the Swiss Institute of Bioinformatics (SIB) provides a curated protein sequence database, with a high level of annotations including protein function, domain structure, post-translational modification and variants. It shows a minimal redundancy and high level of integration with other databases (see also chapter 7). The database is available at <http://www.expasy.org/sprot/>. The current SWISS-PROT release 40.13 (state: March, 2002) contains 106,734 entries.

TrEMBL. The computer-annotated supplement of SWISS-PROT, TrEMBL, contains all translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT and is available at <http://www.ebi.ac.uk/trembl/>. The current TrEMBL release 19.13 (state: March, 2002) contains 592,934 entries.

National Center for Biotechnology Information. The National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA) is a US American resource for molecular biology information and is available at <http://www.ncbi.nlm.nih.gov/>. NCBI creates public databases, conducts research in computational biology, develops software tools for analysing genome data, and disseminates biomedical information. GenBank – a very popular nucleic acid sequence database available at the National Center for Biotechnology Information – is an annotated collection of all publicly available DNA, RNA and protein sequences. The GenBank flat file release 128.0 (state February, 2002) contains 15,465,325 loci and 17,089,143,893 bases from 15,465,325 reported sequences.

Protein Information Resource. The Protein Information Resource (PIR; National Biomedical Research Foundation, Washington DC, USA) is available at <http://pir.georgetown.edu/>. The provided PIR-International Protein Sequence Database (PIR-PSD) is a comprehensive, non-redundant, expertly-annotated, fully classified and extensively cross-referenced protein sequence database in the public domain. Auxiliary databases provide integration of sequences, functional, and structural information to support genomics and proteomics research. The current release 72.00 (state: April, 2002) contains 283,174 entries. The Non-redundant REFERENCE Protein Database (PIR-NREF) provided by the Protein Information was designed to provide a timely and comprehensive collection of all protein sequence data, keeping pace with the genome sequencing projects and containing source attribution and minimal redundancy. The database contains all sequences stored in PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept, and PDB. Non-redundancy is achieved based on clustering by sequence identity and taxonomy at the species level. The NREF report provides source attribution with protein identification numbers and names from associated databases, in addition to protein sequence, taxonomy, and bibliography. The current beta-release (state: March, 2002) contains 888,230 entries.

5.3.4

An example: Identification of an isolated and unknown protein

Following one-dimensional polyacrylamide gel electrophoresis, a Coomassie-stained band of 18 kilodaltons was excised. A tryptic digest of the excised gel band was performed and the peptides were extracted and purified for mass spectrometry. MALDI-MS was performed with a reflectron-time-of-flight instrument with delayed extraction, nitrogen-ultraviolet laser at 337 nanometres and α -cyano-4-hydroxy cinnamic acid as a matrix. The experiment was run at an accuracy better than 50 parts per million with low femtomole sensitivity. The database search using the MALDI peptide map and all of the major search engines did not provide

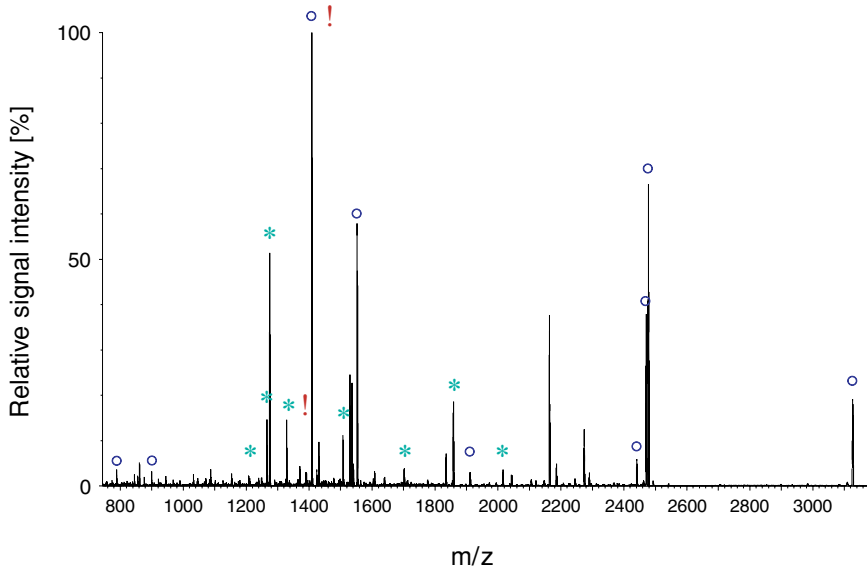
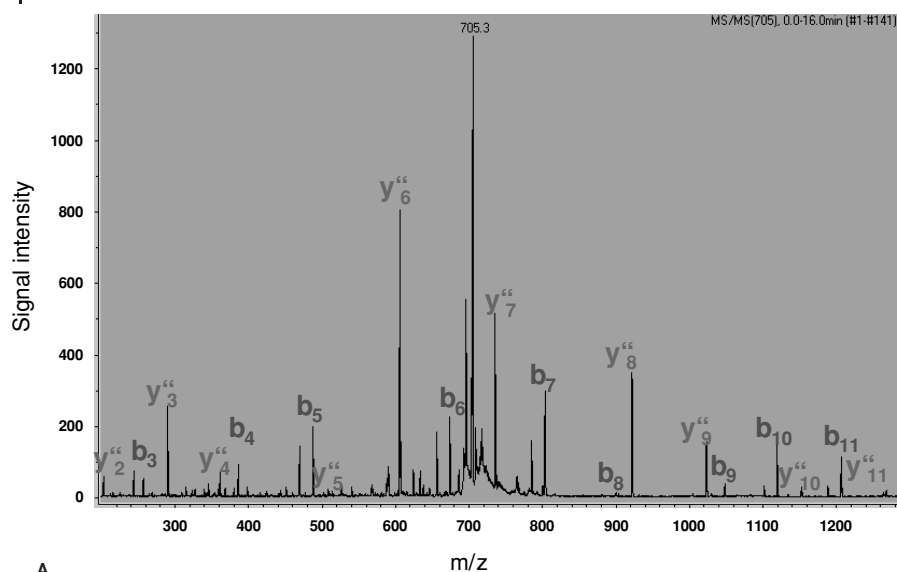


Figure 5.20 Peptide map of gel spot derived from matrix-assisted laser-desorption/ionisation (MALDI) analysis. This gel spot was originally thought to contain a pig lung protein. However, tandem mass spectrometry analysis revealed sequences from transthyretin precursor (prealbumin; marked with outlined circles) and the γ -chain of haemoglobin (marked with

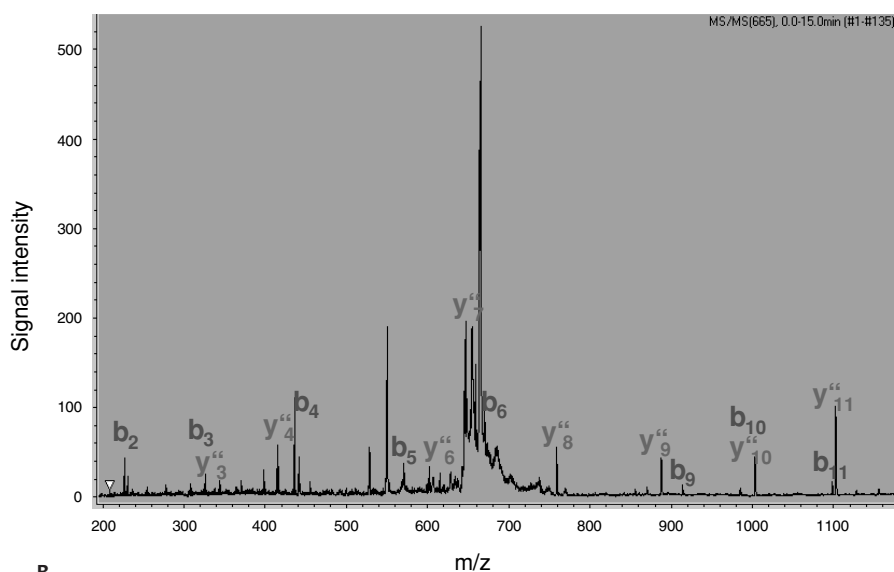
stars). Fragment spectra signals shown in figure 5.21 are indicated by exclamation marks. The authors acknowledge the collaboration of Prof. Dr. Hans-Joachim Galla (Institute of Biochemistry, University of Münster, Germany) in the course of which data shown were obtained. The following abbreviations are used: mass, m ; charge, z .

a conclusive answer with and without species and molecular weight constraints (figure 5.20).

The remaining sample digest was subjected to nanospray electrospray ionisation tandem mass spectrometry using an ion trap with a mass error of 0.4 Dalton. A search using the resulting tandem mass spectrometry spectra of the peptides led to the identification of albumin and haemoglobin peptide ions from *Bos taurus*. The MALDI map was then assigned accordingly (figure 5.21). Thus, only contaminant proteins had been isolated on the gel so far. The mass spectrometric analysis was therefore complicated by the fact that two proteins (the protein of interest and a contamination) were co-eluting in PAGE and that the species differed from the species expected. An additional difficulty was the fact that two proteins were co-eluting in the gel.



A



B

Figure 5.21 Ion trap fragmentation spectra of peptide digests. **A)** Tandem mass spectrometric analysis of transthyretin ion ($[AADETWEPFASGK]H^+$; m/z : 705.3). **B)** Tandem mass spectrometric analysis of haemo-

globin ion ($[VKVDEVGGEALGR]H^+$; m/z : 665.8). Sequences are assigned using b and y-ion series. Single signals are marked with the corresponding letters 'b' and 'y'. The following abbreviations are used: mass, m; charge, z.

5.3.5

Is quantification possible with mass spectrometry?

Although quantification is generally difficult in mass spectrometry, an approach has been proposed to identify and quantify proteins in a high-throughput environment. With isotope-coded affinity tags (ICAT™; Gygi *et al.*, 1999), differences in the relative protein abundance between two different samples can be compared using multi-dimensional liquid chromatography followed by mass spectrometry. ICAT™ has been described in detail in a previous section of this chapter. This recent technology is thought to enhance the detection of differential protein expression, in particular compared to the classical approach using two-dimensional polyacrylamide gel electrophoresis.

5.3.6

Analysis of protein modifications

The biological function of a protein is determined not only by its amino acid sequence but also by post-translational modifications, more than 200 of which have been described to date (see chapter 1). Among these, the major tasks for analysis concern phosphorylation, glycosylation and disulphide bridges. The analysis of modifications is based on the standard approach explained above with adaptations depending on the specific modification. In the analysis of disulphide bridges, it may be possible to obtain fragmentation spectra of the linked peptides *via* HPLC-MS/MS in order to assign the site, possibly utilising a comparison with the reduced and alkylated protein in which the disulphide link is absent.

5.3.6.1 Phosphorylation

Protein phosphorylation is a common and important modification of proteins, because nearly all aspects of cell life are influenced by reversible protein phosphorylation. Changes in the state of protein phosphorylation are regulated by the interplay of two types of counteracting enzyme activities: kinases, which catalyse the addition of phosphate from a nucleoside triphosphate donor to specific residues, mainly serine, threonine or tyrosine, while phosphatases catalyse the removal of specific phosphate groups.

Defining the sites of phosphorylation in a protein and the extent of phosphorylation at each specific site is an analytical challenge for a number of reasons. First, for many phosphoproteins, especially those involved in signalling, in cells their copy number is very low. Second, individual sites are often only partially phosphorylated and third, in mass spectrometry, phosphorylated peptides usually show lower response in positive ion mode due to the electronegativity of phosphate groups.

Solutions to improve mass spectrometric detection of protein phosphorylations involve the specific detection of characteristic marker ions for phosphopeptides and the specific enrichment of phosphopeptides using immobilised metal ion

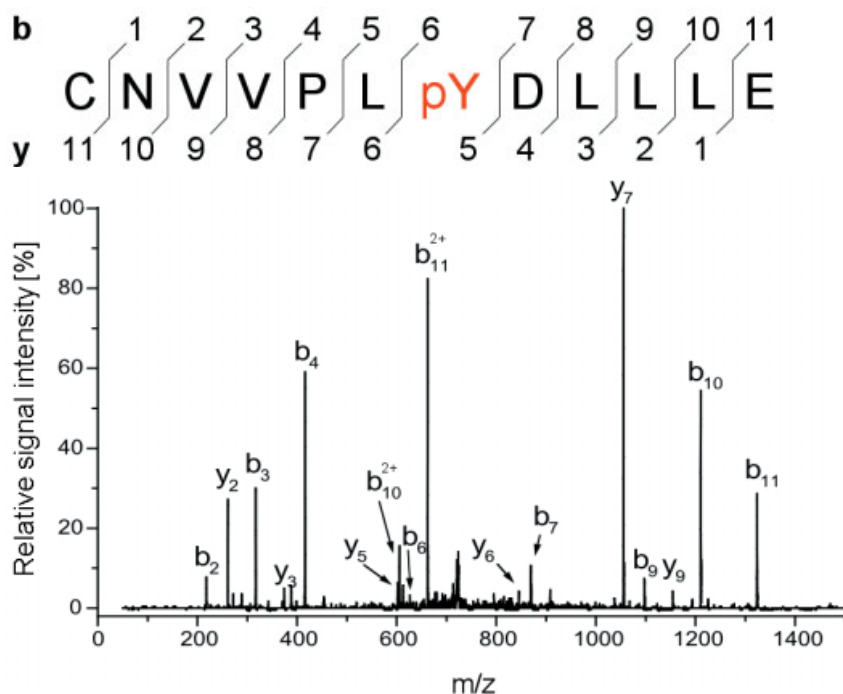


Figure 5.22 Tandem mass spectrometric analysis of phosphorylated anti-estrogen. The phosphopeptide was isolated and fragmented in an ion trap. Phosphorylation of tyrosine (pY)

can be identified using the y and b-ion series. See table 5.4 for the symbols used to abbreviate the names of amino acids. The following abbreviations are used: mass, m; charge, z.

affinity chromatography (IMAC). Once the phosphopeptide is isolated, tandem mass spectrometric analysis can determine the modified sites (figure 5.22).

For further information on protein phosphorylation, the interested reader is referred to Hunter (1995). Detailed information on the selective enrichment of phosphopeptides is described by Andersson & Porath (1986) and Posewitz & Tempst (1999), whereas the comparative MALDI map before and after treatment with phosphatase is described by Zhang *et al.* (1998). Further information on selective tandem mass spectrometers for phosphopeptide-specific ions is provided by Neubauer & Mann (1999) and Annan *et al.* (2001).

5.3.6.2 Glycosylation

Glycosylated proteins are very abundant among secreted and membrane-bound proteins. The biological role of the attached sugar chains (glycans) involves conformational stability, protection against degradation as well as essential molecular and cellular recognition. Glycans are branched structures consisting of different carbohydrate residues. They can either be attached to asparagines residues in the consensus sequence Asn-X-Ser/Thr/Cys (N-linked; where X can be any amino acid)

or to serine or threonine (O-linked). Each site may vary in the glycan structures attached, creating micro-heterogeneity. In addition, different sites may only be partially glycosylated. Depending on the information required (carbohydrate portion or protein part) different analysis strategies are available. They involve combinations of: (i) the release of N-glycans using peptide-N⁴-(N-acetyl- β -glucosaminyl) asparagines amidase F (PNGase F), (ii) the cleavage of O-glycans on reductive β -elimination, and (iii) the separation of the peptide and sugar fractions using reversed-phase high-performance liquid chromatography (RP-HPLC).

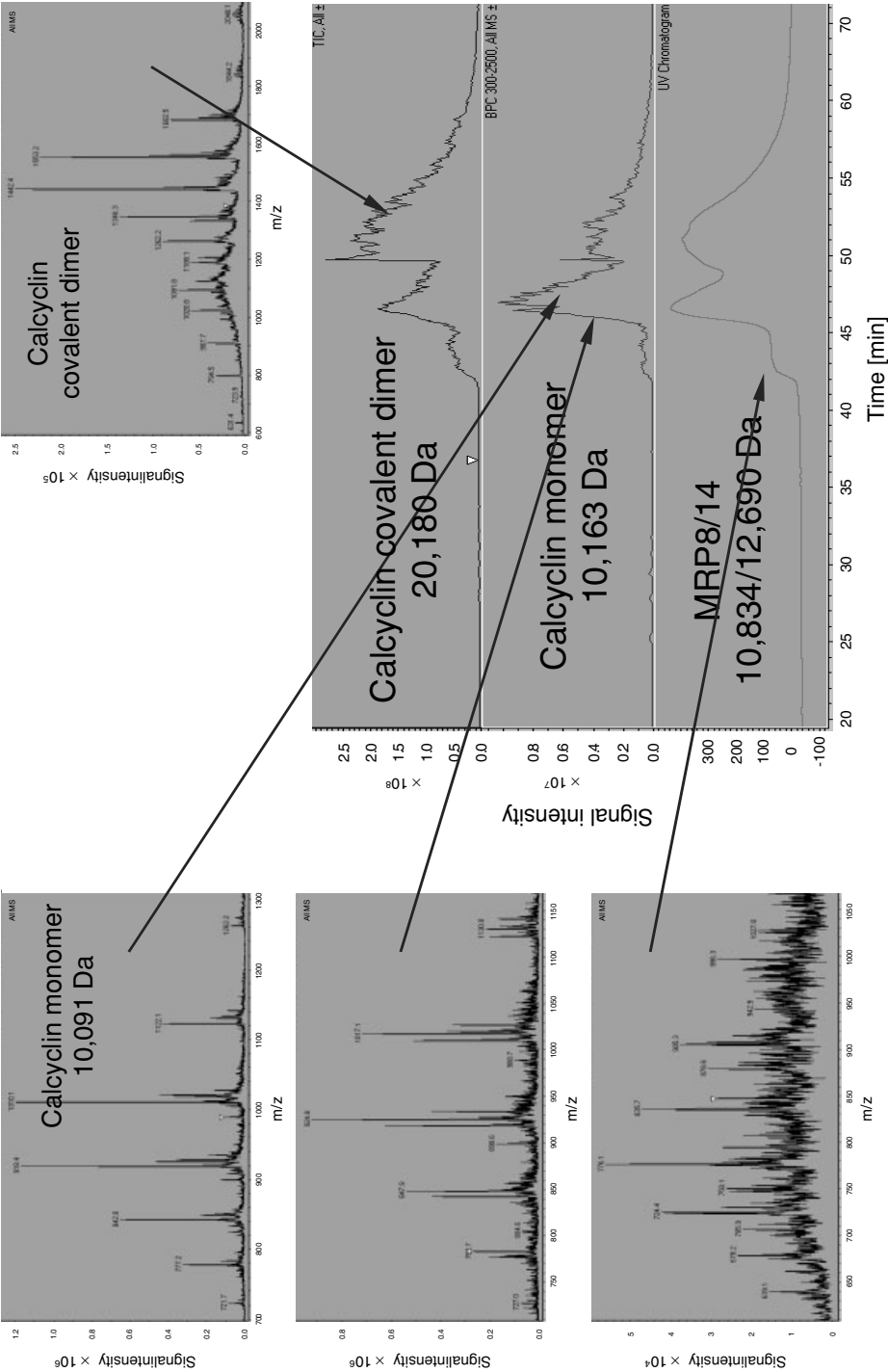
For further information on protein glycosylation, the interested reader is referred to Bill *et al.* (1998). Strategies for glycoprotein analysis are described in Rademaker & Thomas-Oates (1996).

5.3.7

Access to protein tertiary structure

In analogy to the detection of disulphide bridges, chemically cross-linked proteins can be analysed. Using a variety of linkers, amino acid distances can be defined in folded proteins. The potential of this simple approach for the analysis of the protein tertiary structure was recognised early on, and parallel advances in bioinformatics, mass spectrometry and databases of protein structure have fuelled strong renewed interest in the field. The main advantage of mass spectrometry is the low sample consumption in comparison to the traditional structure determination techniques of X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR). In addition, crystallisation and intense labelling are not required. There is no upper weight limit for the protein under study. The basic idea is the combination of 'easily' gained experimental data with molecular modelling. Structure generation will be driven by minimisation of three types of potential energy terms. (i) Experimental constraints of data obtained by mass spectrometric analyses such as long and short range constants, site-specific secondary structure elements, hydrogen exchange and charging patterns. (ii) Globally applicable covalent restraints for proteins such as bond length and angles, improper angles and van-der-Waals terms. (iii) Database potential; a database can contain information on any form of constraining knowledge including function, tendency of a given stretch of amino acids to form a specific secondary structure, phylogenetic information and so on. The method is still in its infancy. Templates for homologous structures available in three-dimensional databases are used for initial attempts.

Current information on high-throughput protein fold identification by using experimental constraints derived from intra-molecular cross-links and mass spectrometry are described in Young *et al.* (2000) and detailed three-dimensional macromolecular structure data of proteins can be found in the Protein Data Bank (PDB; Research Collaboration for Structural Bioinformatics, USA) at <http://www.rcsb.org/pdb/> (Berman *et al.*, 2000). The single international repository for processing and distribution contains data primarily determined experimentally.



- ◀ **Figure 5.23 High-performance liquid chromatography mass spectrometric (HPLC-MS) analysis of protein mixture from human granulocyte lysate.** The mixture was separated on C18-column material and measured *via* online coupling using an ion trap. The main components are calcyclin and its covalent dimer. The myeloid-related proteins 8 and 14 (MRP8/14) are contaminations that are left over from the affinity preparation. Protein spectra are characterised by a charge state envelope. The authors acknowledge the collaboration of Dr. Thomas Vogl (Institute of Experimental Dermatology, University of Münster, Germany) in the course of which data shown were obtained (König *et al.*, 2001). The following abbreviations are used in this figure: Dalton, Da; mass, m, minutes, min; charge, z.

5.3.8

Analysis of intact proteins and non-covalent protein complexes

In many cases it is advisable to measure the uncleaved protein, because this may provide information that is lost in digests. Based on the mass spectrum, one can determine if a protein is pure or partially or completely modified (figure 5.23). The number of attached groups can also be determined (usually the number of phosphorylated sites), if the resolution of the instrument is sufficient and the protein is not too large or inhomogeneous. Useful electrospray ionisation mass spectra can be measured for proteins of up to 70 kilodaltons in size. For higher mass proteins this appears to be difficult. The reason is not mass spectrometry, but the incapability to purify higher mass proteins in such a way that the charge state distribution can be resolved. Matrix-assisted laser-desorption/ionisation is more tolerant, but mass accuracy also suffers from that effect. Generally, the experiment is carried out under acidic conditions to support ionisation. By using volatile buffers such as ammonium bicarbonate, non-covalent protein complexes can also be measured. Although questions remain, correlations of the solution structure and the gas phase structure have been shown and mass spectrometry can be a valuable tool to study protein interaction (Hernandez & Robinson, 2001).

5.3.9

Affinity mass spectrometry

Interfacing bio-specific interaction analysis based on surface plasmon resonance and mass spectrometry allows the combination of information on binding events and the determination of the molecular weight of the interacting molecules. On-line and offline approaches have been demonstrated, but the use of bio-specifically-coated matrix-assisted laser-desorption/ionisation targets seems to be most promising with respect to protein chip development (see the following sections). Ultimately, the molecular weight of isolated bio-specific markers could be measured first followed by on-target enzymatic digestion directly on the same sample holder providing immediate protein identification (König *et al.*, 2002; Nedelkov & Nelson, 2000; Nelson *et al.*, 1999; Williams & Addona, 2000).

5.3.10

Conclusion and outlook

Mass spectrometry will continue to be of major importance in protein analysis. The technique complements established procedures in protein biochemistry and it opens up new possibilities to investigate the living organism. The analysis of proteins separated using polyacrylamide gel electrophoresis is already a routine operation in many laboratories. Special experimental set-ups allow the detection of post-translational protein modifications. Even intact proteins and their natural complexes with other proteins can be measured and bioaffinity mass spectrometry is well on the way facilitating interaction analysis. Mass spectrometry convinces by its non-ambiguous result – the molecular weight. Although sample preparation is sometimes a hindrance, ways will be found to increase the impact of the technique further.

5.4

Protein microarrays

5.4.1

Introduction

The high-throughput tools of the genomic era have yielded a wealth of genetic information that is of use in the diagnosis and treatment of diseases. However, while DNA-based technologies such as micro-capillary and miniature array platforms are most useful tools (Schena, 1996; Shalon *et al.* 1996), their ability to describe complex dynamic processes on the protein level is quite limited. One main reason for this is that there is often a poor correlation between the level of gene expression at the mRNA level, cellular protein abundance and protein activity (Gygi, 1999), a point that has also been made by several other contributors to this book. Furthermore, post-translational modifications of proteins, including phosphorylation, glycosylation, acetylation, lipidation and proteolysis, are often the key to function and affinity, but cannot be quantified by mRNA-based measurements. Finally, drug/target affinity studies that are an integral part of the drug discovery process are performed using the protein targets themselves rather than their encoding mRNA or DNA.

For these reasons, major efforts have been made in the past few years to identify and quantify the complete proteome of a cell or a tissue including all post-translational modifications at any given time (Lopez, 2000), and to make every component of this diverse and numerous population accessible for interaction studies with potential agonists, antagonists or drug candidates. At present, the identification and quantification of the protein complement of cells or tissues is mainly carried out using two-dimensional gel electrophoresis and mass spectrometry-based methods (Pandley & Mann, 2000). Affinity studies rely either on two-hybrid experiments (Bartel & Fields, 1995) or on protein affinity chromatography (Formosa & Alberts, 1984). Together, these approaches have been the workhorses of protein analysis for many years. However useful these methods have been, they suffer from a number

of disadvantages including limited analytical speed and throughput, modest sensitivity and limited flexibility. For this reason, much effort is now focused on the adaptation of highly sensitive techniques with a massive parallelisation potential such as DNA microarray technology in order to use them in protein assays and affinity studies (Abbott, 1999).

Early experiments employing protein microarrays date back to 1989 when Roger Ekins demonstrated that parallel microarray assays can deliver sensitive and selective results (Ekins, 1989). A few years before, fundamental work with a direct optical detection technology called surface plasmon resonance (SPR) had shown that protein interactions can be detected in real-time and without the use of labels on the sensorchip's surface (Liedberg *et al.*, 1983). Also in the 1980s, advances in mass spectroscopy made it possible to exactly determine the molecular masses of proteins from solid matrices – the matrix-assisted laser-desorption/ionisation (MALDI) spectrometer was born (Karas & Hillenkamp, 1988; Tanaka *et al.*, 1987 and 1988). In the following years, the latter two methods independently matured to become powerful analytical tools for protein analysis. The rapid progress in the field of miniature DNA arrays in the mid 1990s has spurred the development of – mostly fluorescence-based – biochip readers. Consequently, most of today's protein arrays employ fluorescence as their detection principle (MacBeath & Schreiber, 2000). However, unlike DNA which can be labelled without a significant alteration of its affinity, proteins are much more delicate, and, in some cases, their structure is greatly affected by chemical labels. Therefore, label-free methods such as the above-mentioned surface plasmon resonance, MALDI or a combination of both, biomolecular interaction analysis mass spectrometry, BIA-MS (Nelson *et al.*, 2000), have become increasingly popular in recent years. In addition to the further improvement of these detection platforms, one of the most challenging tasks today is the development of highly specific sensorchip surfaces, which are a prerequisite for sensitive and selective assays. The construction of large protein libraries still represents a major problem, too, for as yet there is no simple method of protein amplification, such as the PCR amplification of DNA, available.

5.4.2

Principles and basics, typical results and applications

Generally speaking, protein biochips can be considered as chromatography on a two-dimensional solid surface (a chip) combined with a sufficiently sensitive detection technique to analyse the binding of (bio)molecules from sample solutions (see figure 5.24 for the workflow of protein microarray operation). In most cases, the surface is coated with proteins or protein binding molecules immobilised in a regular pattern on surface-modified planar substrates. Piezoelectric or ring-pin spotters are used to fabricate such arrays. An alternative approach works with chip surfaces of different charge and hydrophobicity, each of which attracts proteins in a quasi-specific manner thanks to intermolecular attractive forces such as electrostatic interaction, hydrophobic interaction, co-ordinate covalent and hydrogen bonding (Merchant & Weinberger, 2000).

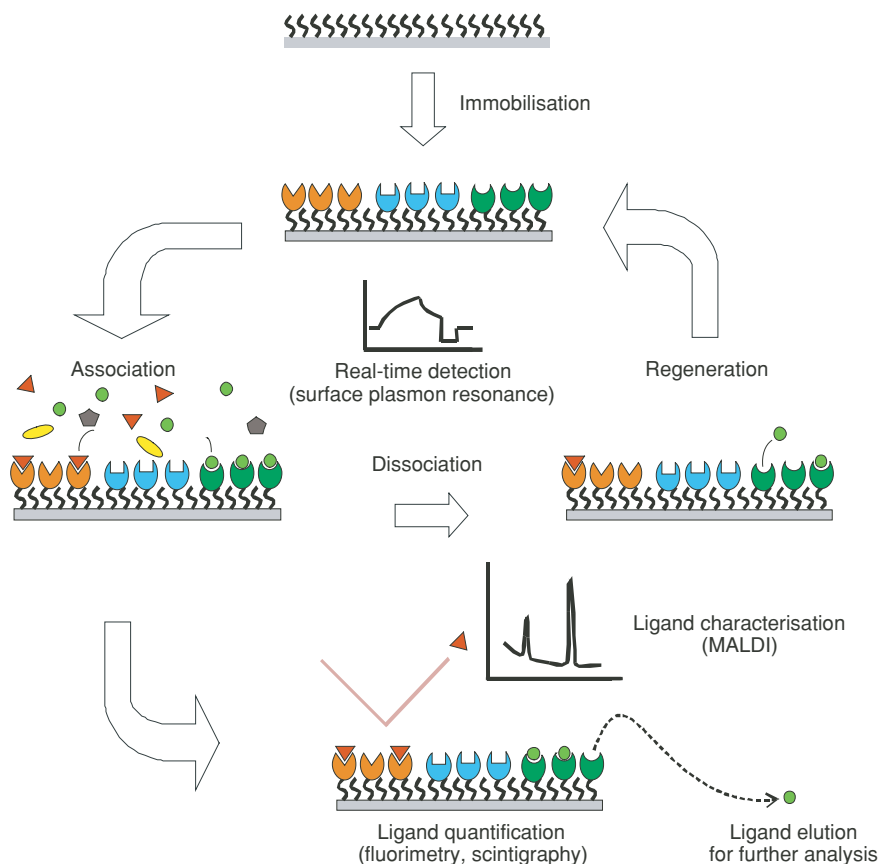


Figure 5.24 Workflow in protein biochip operation. After proteins have been immobilised to a solid surface, specific binding and dissociation of ligands can be monitored in real-time. In addition, it is possible to quantify or

characterise bound ligands in later steps. The following abbreviation is used in the figure: Matrix-assisted laser desorption/ionisation, MALDI.

Depending on the detection technique employed, the chips are made of glass, metal or plastic. The samples are either directly deposited upon the chip's active areas, or they are introduced in a continuous flow stream, or they are injected into a continuously mixed cuvette. The latter two methods require flow chambers or cuvette systems to ensure a sufficient convection that enhances the binding rate of high molecular weight analyte molecules. Discrimination between specific and non-specific binding is achieved either by washing sequences of increasing stringency, or, preferably, by the inherent bio-inertness of the surface coating which at best should suppress non-specific binding almost quantitatively. Alternatively, the sample solution can be pre-treated and purified by microfluidic devices before it reaches the protein chip surface (Figeys, 1999; Guetens *et al.*, 2000).

While DNA is stable enough to be spotted on almost every surface without losing its specific hybridisation properties, proteins tend to denature if the surfaces to which they are to be attached exhibit a high energy and/or charge density. Therefore, protein chips require a biocompatible coating prior to the immobilisation step. This coating must provide accessible covalent coupling sites, a structure-retaining immobilisation matrix, low non-specific binding and the option of quantitative regenerability after each binding cycle. The coating should also shield the chip substrate, thus enhancing the chemical resistance and providing a stable, homogeneous support for further modification. Finally, to allow a reliable interpretation of the experimental data, properties such as surface energy, charge and chemical functionality must be well-defined over the entire chip area.

Typically, the cleaned chip surface is first covered with a linker layer to which a bio-inert matrix is then covalently coupled. The matrix usually consists of hydrophilic compounds that contain functional groups for immobilisation of proteins. Frequently used matrix polymers are polysaccharides, poly L-lysine (Haab *et al.*, 2000), polyethyleneglycol or proteins, for example, avidin (Rowe *et al.*, 1999) or bovine serum albumin (MacBeath & Schreiber, 2000). Figure 5.25 shows an example

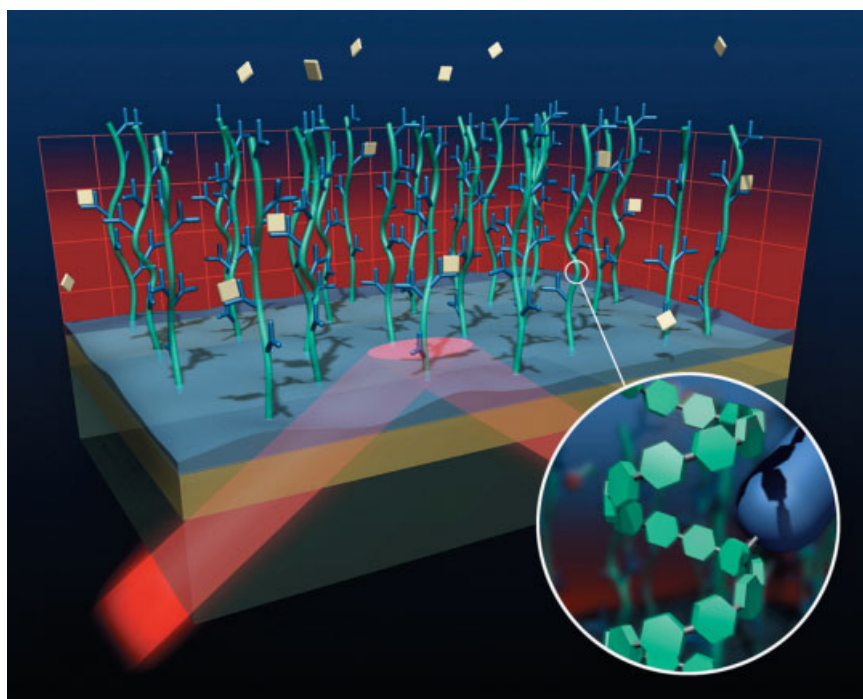


Figure 5.25 The molecular structure of a hydrogel-coated surface plasmon resonance protein biochip. A covalently-attached polysaccharide monolayer serves as an immobilisation matrix for receptor molecules and reduces the possibility of non-specific interactions with the

surface of the biochip. The binding of ligands is detected in real-time *via* modulations in the intensity of a reflected laser beam. Reprinted with kind permission from XanTec bioanalytics GmbH (Münster, Germany).

Table 5.6 The two main classes of protein microarrays and their applications.

Type of protein array	Immobilised element	Applications
Protein function	Proteins	Drug discovery
	Peptides	Antibody characterisation
	Glycosides	Organ and disease specific arrays
	Protein/nucleic acid conjugates	Autoimmune diagnostics
Protein detection	Antibodies	Second generation proteomics
	Single chain variable antibodies	Pharmacokinetics
	Protein and nucleic acid aptamers	Protein profiling
	Small molecules	Diagnostics
	Molecular imprinted materials	Environmental and food analysis

of a dextran-coated surface plasmon resonance sensor chip. The polysaccharide chains are covalently attached to a linker layer in a brush-like architecture, serve as an immobilisation matrix for ligands, and stabilise the chip surface against non-specific binding.

Depending on their biochemical functionality, protein chips and microarrays have been divided into protein function and protein detecting systems (Kodadek, 2001; also see table 5.6). Protein function arrays consist of native proteins which can be obtained from cDNAs by *in vitro* transcription/translation or some other technique and which are spotted onto a suitable surface. The principle of a typical procedure is shown in figure 5.26. Protein function arrays may be used to probe the function or binding properties of native proteins. An interesting application for such arrays is the mapping of all proteins of a specific organ onto one array in order to study the interactions of small molecules with each arrayed protein.

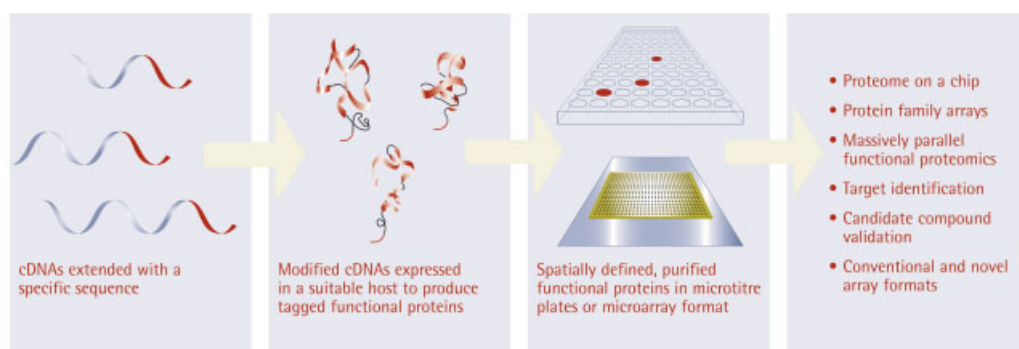


Figure 5.26 Process for the production of protein libraries. Expression of modified cDNAs yields tagged proteins which can be readily immobilised to chip surfaces or microtiter wells. The figure is reprinted with kind permission of Sense Proteomic Ltd. (Babraham, Cambridge, United Kingdom).

This type of application is relevant to a problem frequently encountered in the pharmaceutical industry, namely, whether a given drug candidate binds tightly to proteins other than its intended target (Lecitra & Liu, 1996). Another example is illustrated in figure 5.27. DNA-binding proteins of the same type showing different kinds of mutations in their peptide sequence are arrayed on a solid surface. Afterwards, the array is incubated with labelled DNA molecules that are able to bind to the arrayed proteins. The mutations of the proteins influence their affinity of binding the DNA molecules. Using this procedure, proteins can be identified that bind with a high affinity to a certain DNA molecule.

In protein-detecting arrays, ligands that bind with high affinity and specificity to each protein are arrayed on the appropriate surface. Today, such ligands are usually antibodies or single chain variable fragments. The production of tens of thousands of characterised antibodies is a time-consuming and expensive process. For this reason, new methods for the generation of single-chain Fv (scFv) antibody libraries are currently being developed (de Wildt *et al.*, 2000). Other approaches, for example, molecular imprinted materials, might become a promising alternative in the future. The ready manufacture of such artificial antibodies, which are much more stable than their natural counterparts may become feasible in the future. However, the affinity of this kind of ligands as they currently exist is too small to make them a serious alternative to antibodies at the present time.

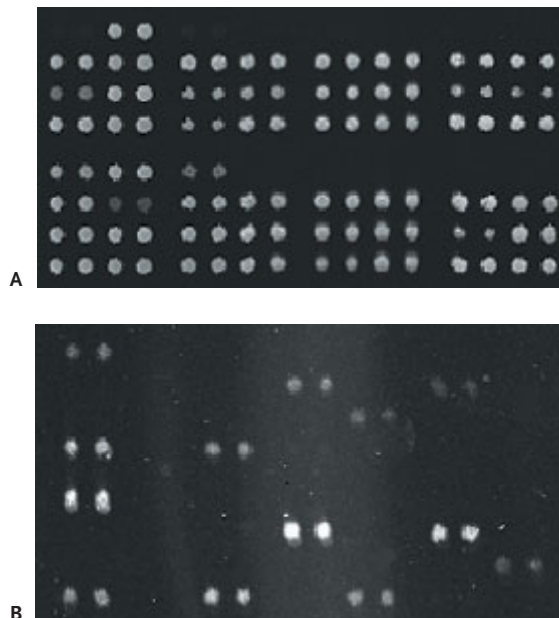


Figure 5.27 Differential DNA binding to single nucleotide polymorphisms (SNP)-encoded mutants in an array format. Protein arrays showing **A)** the abundance of attached protein detected by labelled specific antibody and **B)** differential DNA binding to the proteins due to SNP-encoded mutations in some members of the panel. The figure is reprinted with kind permission of Sense Proteomic Ltd. (Babraham, Cambridge, United Kingdom).

Arrays of the protein-detecting type will have wide applications in the field of proteomics and pharmacokinetics, as they allow for a protein profiling, *i.e.* they give quantitative information about the relative abundance of any given protein in an organism at a given time. Such arrays can also be used in diagnostics and for food and environmental analysis.

In addition to the biospecificity of the surface and the immobilised ligand, the sensitivity of protein chips critically depends on the detection method (see table 5.7 for an overview). Compared with DNA oligonucleotides, the molecular weight span of the analytes to be detected is much larger and can range from several hundred Dalton in applications such as drug discovery to some hundred kilodalton. As a consequence, the detection principles developed for DNA arrays may not be suitable for analytes other than DNA. Moreover, in many cases it is not only the concentration of the ligand that is of interest, but also the binding constants or even the molecular weight.

The most commonly used detection method today is fluorescence (MacBeath & Schreiber, 2000). Many variants exist, including fluorescence lifetime, polarisation or fluorescence resonance energy transfer (FRET). An advantage of fluorescence-based methods is the high sensitivity which allows even single molecule detection. Arrays are normally constructed on glass when using fluorescence detection, since glass auto-fluorescence is low. The array readers can detect up to five wavelengths and use confocal optics to enhance the signal-to-noise ratio. A higher sensitivity can be achieved with fluorescence planar waveguide (FPWG) detection (Neuschäfer *et al.*, 1996). This technique relies upon fluorescence excitation in the evanescent field of light which propagates in a planar waveguide. The advantage of this detection method is that only surface-bound fluorophores are excited, and not any labels or other fluorescent moieties found in the free solution.

Despite the high sensitivity, there are several problems associated with fluorescence detection. First, due to the chemical heterogeneity of proteins, some proteins label far more than others, which makes it necessary to obtain calibration curves for each arrayed protein prior to measurement. Secondly, as the chemical labelling process converts positively charged lysine side chains to amides, and as the labels themselves are relatively hydrophobic, the surface characteristics of the protein are greatly changed. In the worst case, the labelled proteins are denaturated and may no longer bind to the immobilised ligand. This effect becomes more problematic with analytes of low molecular weight. An extreme example is provided by the low molecular weight compounds analysed in drug discovery. It is clear that labelling with a one kilodalton fluorophore may drastically alter the affinity profiles of such small organic molecules.

More traditional protein quantification methods such as enzyme-linked immunosorbent assay (ELISA) detect coloured products of enzymatic reactions. Enzymatic detection is usually sensitive due to the signal amplification provided by enzyme activity. However, the bulky enzyme may alter the diffusion properties of the labelled molecules. Enzyme-linked detection has been used to detect proteins on both filter (Büssow *et al.*, 1998) and glass arrays (Arenkov *et al.*, 2000).

Table 5.7 Suitable detection techniques for biomolecular interaction analysis.

<i>Information</i>	<i>Transducer classification</i>	<i>Examples</i>
Quantitative	Label-free detection	Micro-refractometry
		Surface plasmon resonance (SPR)
		Grating coupler
		Resonant mirror
		Mach-Zehnder interferometers
	Reflectometry	Brewster angle spectrometry
		Ellipsometry
		Reflectometric interferometry
	Atomic force microscopy	–
	Micro-gravimetry	Quartz microbalance
	Detection with labels	Fluorimetry
		Fluorescence immunoassay (FIA)
		Fluorescence lifetime assays
		Fluorescence polarisation (FPIA)
		Fluorescence resonance energy transfer (FRET)
		Fluorescence planar waveguide (FPWG)
		Luminometry
		Luciferase ‘glow’ assays
	Densitometry	Particle-based assays
	Colourimetry	Enzyme-linked immunosorbent assay (ELISA)
	Scintigraphy	Radio immunoassay (RIA)
	Amperometry	–
Qualitative	Mass spectrometry	Matrix-assisted laser-desorption/ionisation (MALDI)
		Surface-enhanced laser-desorption/ionisation (SELDI)
	Evanescant field spectrometry	Attenuated total reflection (ATR) couplers

Protein arrays have also been detected using incorporation of radioactive isotopes (Ge, 2000). Isotopic labelling is sensitive, because the emitted radioactivity can be integrated over time and the background is low. Moreover, isotopic labels do not alter the labelled protein's structure or activity. Disadvantages of isotopic labelling are high cost, lengthy preparations and the extra precautions required when working with radioactivity.

A limitation of both isotopically labelled and enzyme-linked reactions is the fact that only the end point of the binding reaction is quantified. It is impossible to measure the complete binding curve in order to obtain kinetic data. Additionally, the amount of capture material in each spot must be tightly controlled to enable quantitative comparison between spots.

The best way to avoid problems with labels, of course, would be to use a detection platform that requires no sample labelling whatsoever. Several technologies have been developed to allow the direct detection of binding events at the sensor surface without the use of labels. In addition to electromechanical or electro-acoustic approaches such as the quartz microbalance (Rickert *et al.*, 1996) or surface acoustic wave (SAW) sensors (Tom-Moy *et al.*, 1995), direct optical biosensors are being increasingly used for the thermodynamic and kinetic characterisation of affinity reactions on chip surfaces. There are two groups of methods. The most frequently-used detection principle measures the small refractive index changes at the sensor surface that occur when biomolecules with a refractive index of approximately 1.5 replace water of which the refractive index is 1.33. This method is called micro-refractometry. Examples for micro-refractometric sensors are surface plasmon resonance (see figure 5.28 for the principle of this method), the grating

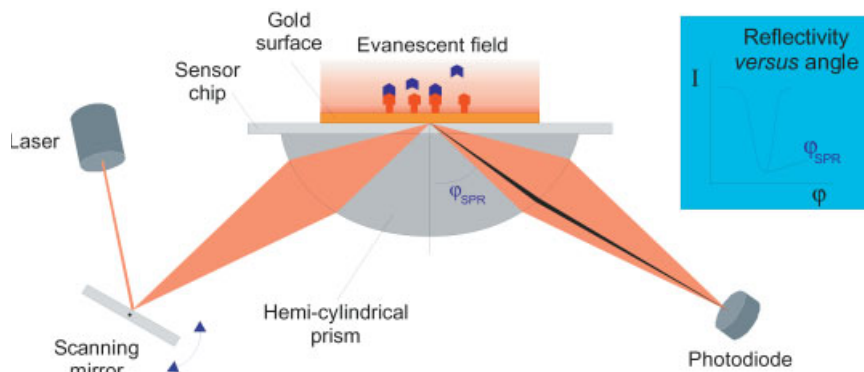


Figure 5.28 Surface plasmon resonance bio-sensor set-up. Polarised laser light is shone at different angles onto a gold-coated sensor chip. At a critical angle φ_{SPR} , which depends on the refractive index close to the sensor chip surface, a resonance effect occurs and a sharp decrease in the intensity of the reflected light is measured. As the refractive index changes upon

binding of biomolecules, biomolecular interactions can be monitored in real-time *via* the shift of the angle φ . The following abbreviation is used in the figure: Surface plasmon resonance, SPR. The figure is reprinted with kind permission of XanTec bioanalytics GmbH (Münster, Germany).

coupler (Lukosz & Tiefenthaler, 1988), the resonant mirror (Cush *et al.*, 1993) and the Mach-Zehnder interferometer (MZI) (Brosinger *et al.*, 1997). The sensitivity of these devices allows the detection of differences in refractive index of as little as 10^{-6} , corresponding to mass changes of a few picograms per square millimetre. It is thus possible to quantify proteins in femtomolar concentrations even in matrices with a high non-specific background (Schneider *et al.*, 2000).

The second approach, reflectometry, is based on changes of the optical properties of the chip surface itself which leads to spectral or polarisation shifts of light reflected at the chip surface. This method includes the Brewster angle spectrometry, ellipsometry and reflectometric interferometry (Gräfe *et al.*, 1998).

Surface plasmon resonance biosensors have been commercially available since 1990 and have now become part of established methodology. In addition to their label-free operation, a major advantage of this technology lies in the possibility of real-time data acquisition. The binding curves obtained in this fashion can be kinetically evaluated and allow the rapid assessment of kinetic data such as the association and dissociation rate constants k_a and k_d respectively (see figure 5.29).

While the direct optical techniques described above are well suited to obtaining quantitative information on the concentration or the kinetic characteristics of a specific biomolecule, they do not provide qualitative data such as the molecular weight or amino acid sequence of detected analytes. Over the past few years, this gap has been filled by the use of matrix-assisted laser-desorption/ionisation mass spectrometry instrumentation which, under certain conditions, is able to desorb surface-bound proteins. In addition to directly modified mass spectrometry, enhanced and simplified procedures for sample extraction and clean-up facilitated effective on-probe investigation of biomolecules. For example, an approach termed surface-enhanced laser-desorption/ionisation (SELDI) (Merchant & Weinberger,

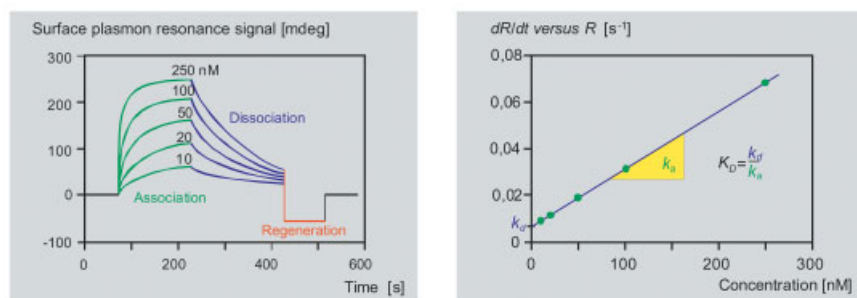


Figure 5.29 Assessment of kinetic data derived from binding curves obtained by surface plasmon resonance biosensors. Important kinetic data, such as the association rate constant k_a and the dissociation rate constant k_d can be obtained from surface plasmon resonance sensograms by first calculating the dissociation constants K_D of different binding curves and then plotting these constants *versus* the corre-

sponding ligand concentration. The association rate constant k_a is then calculated from the slope of the resulting straight line and its y-intercept. The following abbreviations are used in the figure: millidegrees, mdeg; seconds, s; nanomolar, nM; rate constant, R; time, t. The figure is reprinted with kind permission of XanTec bioanalytics GmbH (Münster, Germany).

2000) makes it possible to perform analyte-specific sample cleanup directly on a MALDI target. Another promising technique, biomolecular interaction analysis mass spectrometry (BIA/MS), combines mass spectrometry with realtime SPR detection (Nelson *et al.*, 1999).

Both approaches allow a molecular mass determination of bound analyte molecules with an accuracy of some 100 parts per million. If an additional enzymatic proteolysis step is performed, the amino acid sequence of bound proteins can also be determined.

5.4.3

Discussion

In recent years, protein biochips and protein arrays have become a valuable addition to the portfolio of modern bioanalytical tools. Although the technology is still in its infancy, many types of experiments have already been reported, such as the measurement of protein/protein, protein/DNA, protein/RNA and protein/small molecule interactions. The main advantage of current protein array technologies over conventional analytical devices is the possibility of screening thousands of proteins against one sample and of obtaining a wide variety of information which would be extremely tedious or even impossible to gain using other techniques.

However, the method has weaknesses which are either inherent or awaiting a technical solution. A fundamental problem is that reactions that occur in solutions are examined on the surface of the protein array. Mass transfer limitations during analyte binding (Schuck & Minton, 1996) and steric hindrance on the chip surface can complicate the kinetic analysis of biomolecular interactions. The latter is likely to become a serious problem when the analytes under investigation form larger complexes with other proteins, a phenomenon frequently observed in living systems (Alberts, 1998). What is more, the immobilisation process includes at least some chemical modification, attachment to bulky binding domains, or even denaturing on the chip surface, and may therefore alter the binding characteristics of the immobilised ligand. The problems associated with protein labelling can be avoided by the use of label-free detection technologies, which will probably gradually replace conventional detection methods.

Furthermore, although the experimental methodology has greatly advanced, the challenge remains to generate a significant number of capture reagents for the arrays. It is certainly a daunting task to express and purify thousands of different proteins, and some proteins will inevitably prove refractory to biochemical manipulation. Nonetheless, the effort will be worthwhile.

In the future, multi-analyte assays will be capable of simultaneously measuring the expression of very large numbers of genes and proteins. This is likely to reveal important information on the causation, progression, and response to treatment of disease (Cahill, 2001). Protein function arrays will have a considerable impact on drug discovery, and it is possible that in the coming ten to twenty years protein arrays will become a mainstay of medical diagnostics.

5.5

Protein expression analysis with immunoassays

5.5.1

Introduction

Post-transcriptional regulation of gene expression is frequent in higher organisms. It is therefore not possible to accurately predict protein expression levels from mRNA transcript data. This is why we need to be able to quantify protein expression. For decades, immunoassays have been a mainstay of protein expression analysis. For proteins at low concentrations in complex biological matrices such as blood or cells, immunoassays are the method of choice because of their sensitivity, low cost, speed and technical ease of use. Furthermore, many convenient and reliable commercial kits and well-documented test protocols are available. The following sections summarise the most important immunoassay techniques that have been developed for quantification of proteins in biological and medical applications. Detailed descriptions of immunoassays and their automation are available in references (Bock, 2000; Guardigli, 2002; Wild, 2001).

5.5.2

Immunoassay reagents (binding agents)

Antibodies are complex Y-shaped glycoproteins that exhibit specific binding capabilities for surface structures (epitopes) on an antigen ('antibody-generating') molecule, usually a protein (figure 5.30). Epitopes typically consist of five to ten amino acids and/or sugar molecules. The specific binding of the antigen binding site (paratope) of the antibody to the epitope is achieved by the exquisite complementarity of their surfaces. The epitope and the paratope match each other like a key matches its lock. This unique property of antibodies is the reason for their usefulness in immunoassays, because only the antigen of interest is recognised by the antigen-binding site of the antibody. For this reason, immunoassays do not require many of the purification or pre-treatment steps that are often needed with other analytical methods such as mass spectrometry or high-performance liquid chromatography. Although non-covalent, the binding of the antigen to the antibody is strong because it results from the sum of a large number of weak interactions such as hydrogen bonds, ionic bonds, van der Waals and hydrophobic interactions. Furthermore, the binding is reversible and does not alter permanently either the antigen or the antibody. The affinity of an antibody for a particular antigen refers to the strength with which the paratope of the antibody and the epitope of the monovalent antibody interact. Immunoassays and therapeutic applications generally require antibodies of high affinity.

Polyclonal antibodies are produced by injection of the antigen into an experimental animal. For an effective immune response, the antigen should have a molecular weight of greater than five kilodalton. If the antigen produces a poor immune response, it may be necessary to link it to a high molecular weight carrier

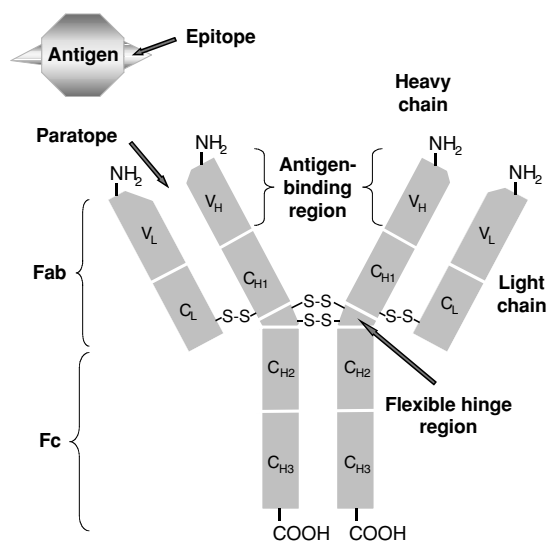


Figure 5.30 The structure of immunoglobulin G (IgG), virtually the only isotype used in immunoassays. Immunoglobulin G is a glycoprotein, consisting of two identical heavy chains with three constant domains (C_{H1} , C_{H2} , and C_{H3}) and one variable domain (V_H) paired with two identical light chains. The light chains consists of one constant domain (C_L) and one

variable domain (V_L). The variable domains (V_H and V_L) are important for antigen binding. The chains are linked by a variable number of disulphide bonds, forming a flexible Y-shape. The molecule can be cleaved chemically into an Fab and an Fc fragment. This is done in order to remove the Fc fragment which may interfere in certain immunoassays.

such as keyhole limpet haemocyanin or bovine serum albumin. The antibodies produced by this technique usually retain the ability to bind to the uncoupled antigen. The main disadvantage of polyclonal antibodies is that they consist of an undefined mixture of different antibodies with different specificities, many of which are irrelevant for antigen recognition but are a potential source of interference. Further purification steps using solid-phase antigen on an affinity column such as cyanogens bromide-activated Sepharose 4B may therefore be required before such polyclonal antibodies can be used in immunoassays. The simplest way to determine the purity of an antibody solution is to separate an aliquot on a sodium dodecyl sulphate polyacrylamide gel and to examine the components using Coomassie blue or silver staining (Harlow & Lane, 1998).

Monoclonal antibodies are produced in cell culture by the descendants of a single clone of plasma cells that is immortalised by cell fusion using the hybridoma method described by Köhler & Milstein in 1975. Deservedly, this work was rewarded by a Nobel prize in 1984. Monoclonal antibodies have uniform molecular structure, high specificity, and reproducible properties from batch to batch. For these reasons, it is preferable to use monoclonal antibodies in immunoassays if at all possible. Nevertheless, even monoclonal antibodies present particular problems. It may be technically difficult to select suitable clones and to purify the

monoclonal antibodies from other cell proteins. A relatively long time of at least four months is required to generate the antibody and it may not be easy to scale production up to commercial levels. Further disadvantages are the high level of technology required and difficulty in producing monoclonal antibodies from species other than rodents (Shepherd & Dean, 2000).

Recently, great advances in the construction of antibodies have been achieved using the techniques of molecular biology. Single chain fragments (Fab) of antibodies that bind to a particular antigen, or, in terms of immunoassays, analyte, may be isolated from a recombinant phage or ribosome antibody library by passing the library over a column containing immobilised analyte. The selection of suitable clones can be simplified by display technologies, and selection conditions can be adjusted to improve affinity, specificity or stability of the antibody. Different fragments of the variable region of the antibody can be combined to produce multivalent antibodies that are capable of recognising adjacent epitopes on a large antigen. In addition, mutations may be introduced by recombination in order to generate antibodies with new specific features. Hybrid proteins containing an antigen-binding site and, for example, a specific enzyme activity can also be constructed. The expression of antibodies in mammalian cell culture and in cell-free systems is possible. Computer-assisted molecular modelling may also be employed to increase the specificity and affinity of the antibody (Kontermann & Dübel, 2001).

Ligand-binding assays represent potential alternatives to immunoassays. In these assays, binding proteins such as receptors or transporters are used instead of antibodies for the recognition of antigens. Receptor assays possess some theoretical advantages over immunoassays. For example, an antibody may have difficulty in discriminating between proteins with similar structure but different activities, such as protein isoforms that arise as a consequence of post-translational modification. Receptor assays may be better at measuring such proteins since the signal they record reflects the biological activity rather than the concentration of the analyte (Soldin, 1996).

Other types of binding reagents are aptamers, polynucleotides with a high and specific affinity for a particular analyte. It has been suggested that molecularly imprinted polymers might provide an alternative to antibodies for use as recognition elements. Synthetic ligand binders have the advantage of a completely defined structure, clear binding affinity to an antigen, a high physical and chemical stability, and low production costs. However, due to the satisfactory performance of antibodies in existing assays and the availability of a wide range of robust antibody-based protocols, use of synthetic binders has been restricted to special applications where, for example, no antibody or natural receptor is available (Jayasena, 1999).

5.5.3

Design options of immunoassays**5.5.3.1 Competitive immunoassays**

Immunoassays are divided into competitive and non-competitive assays (figure 5.31). In competitive immunoassays (figure 5.31.A), the analyte competes with a labelled antigen for a limited amount of antibody molecules. The presence of the antigen in the sample inhibits binding of the labelled antigen. Thus, the amount of signal produced by the labelled antibody is inversely proportional to the amount of antigen in the sample. A standard curve with increasing amounts of known concentration of unlabelled antigen is used to determine the quantity of antigen in the sample. The detection limit of competitive immunoassay is improved by the use of highly-specific labels and high-affinity antibodies. The number of immobilised antibodies must not greatly exceed the smallest amount of antigen to be measured. If the antigen concentration is very low, the binding reaction is slow and long incubation times are required. The advantage of the competitive assay is that specificity derives from labelled antigen, not from the antibody. For this reason, purified antibody is not required. The serum of an immunised animal such as a rabbit can be used directly as a binding reagent. This is why this test format is generally used as a first approach for measurement of newly discovered proteins and for quantifying small molecules such as hormones and drugs in almost all commercial immunoassay kits. A number of variations of this competitive principle have been established. For example, a limited amount of labelled antibodies may be used to bind to an antigen fixed on a solid phase, or to the antigen in the sample. After incubation and a washing step the signal from the solid phase is measured (Ekins, 1997).

5.5.3.2 Immunometric immunoassays

In non-competitive or immunometric immunoassays ('sandwich-assays') (figure 5.31.B), two specific antibodies are present in excess. Each antibody detects a different epitope on the antigen. The antigen is first allowed to react with the first antibody ('capture antibody'), which is firmly attached to a solid surface. A second antibody ('signal antibody') is then allowed to react with the antigen. After a washing step to remove unbound signal antibodies, the activity of the solid phase is determined. In order to quantify formation of the sandwich complex, the second antibody is usually labelled with an enzyme that converts a substrate to produce a coloured dye, which is quantified by optical methods. Standard curves are generated using solutions with defined concentrations of antigen. In contrast to competitive assays, sandwich assays do not depend on antibody affinity. Their requirement for two antibodies makes them more specific than single-site assays. The advantages of sandwich assays are high sensitivity and specificity, a wide measuring range, low background signals, greater robustness and shorter incubation times than competitive tests. Due to the fact that one immunoreactant is in excess, non-competitive assays have much shorter incubation times than their competitive counter-

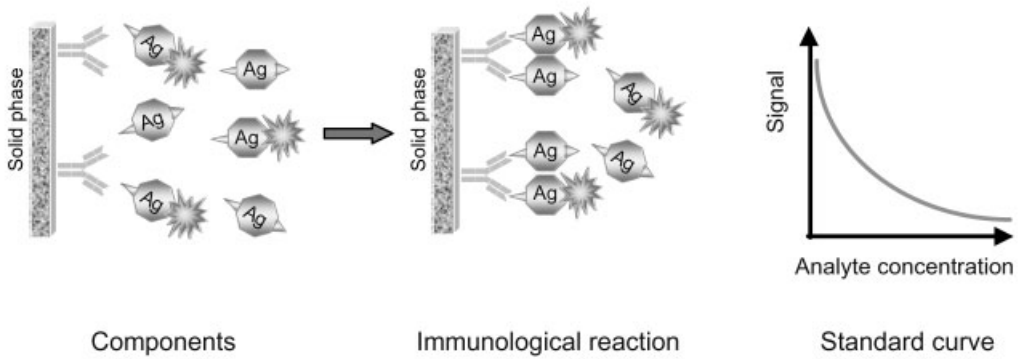
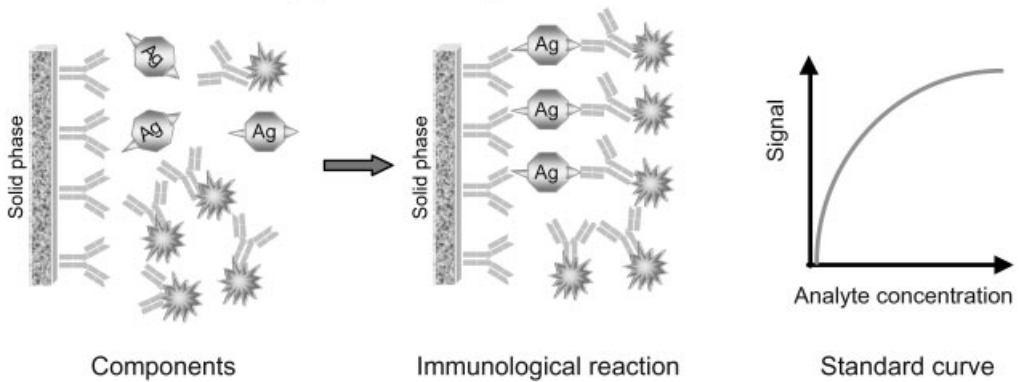
A Competitive assay**B** Immunometric assay (sandwich assay)

Figure 5.31 Principles of different kinds of immunoassays. A) Competitive immunoassay. A labelled antigen (Ag) competes with the antigen of the sample for a limited number of antibody sites. The amount of labelled antigen binding to the immobilised antibodies is indirectly proportional to the concentration of the

analyte in the sample. **B) Immunometric assay.** The first or capture antibodies are immobilised on a solid surface. A second labelled antibody binds to another epitope of the analyte to generate the 'sandwich complex'. The signal response in this assay design is proportional to the concentration of analyte in the sample.

parts. Furthermore, the high concentration of antibodies gives sandwich assays a long shelf-life with prolonged stability of the calibration curve (Miles & Hales, 1968).

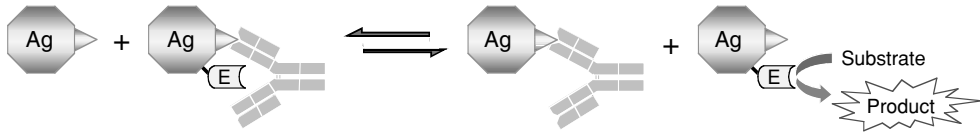
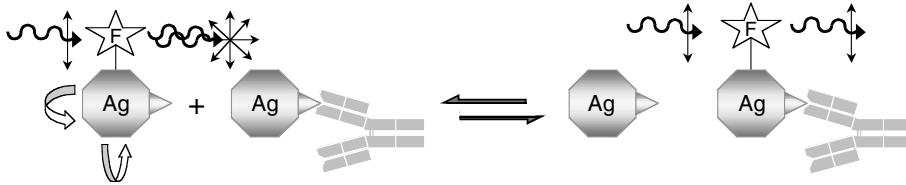
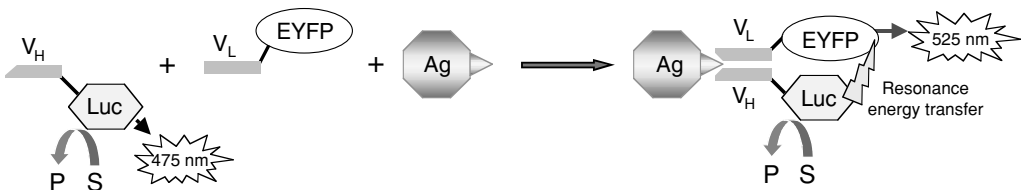
Both competitive and immunometric assays require separation of the antigen/antibody complex from the free antigen or labelled antibody. Both also have the drawback of a long dwell time due to several cycles of consecutive binding and washing steps, which makes automation difficult. Separation of the phases can be simplified by attaching the antibodies to a solid phase, such as a microtiter plate, coated tubes, paramagnetic or synthetic particles. Standard immunoassays are usually performed in a 96-well microtiter plate with immobilised antibody or

antigen. In automated systems, microparticles are often used in cuvette-based analysers because they can be readily separated from the solution by filtration or, in the case of paramagnetic beads, by attractive force (see figure 5.34.B). The antibody is fixed covalently or *via* a streptavidin/biotin system to the solid phase. The basis of this system is the chemical attachment of the vitamin biotin *via* its carboxy terminus to the antibody. The biotinylated antibody usually binds selectively the target protein in the assay. The protein streptavidin is usually conjugated to the solid phase and is often preferred to avidin because it is not glycosylated, which helps to decrease non-specific binding. This solid phase can be used as a general binding system for a biotinylated antibody. The interaction between biotin and avidin or streptavidin is the strongest protein/ligand interaction yet known in nature (Diamondis & Christopoulos, 1991).

5.5.3.3 Homogenous and heterogeneous immunoassays

Immunoassays can be classified into homogeneous assays, which do not require separation, and heterogeneous assays, which require separation steps (see above). Heterogeneous assays are usually more sensitive, less prone to interference and applicable to high molecular-weight molecules such as proteins (see figures 5.33 and 5.34). Homogenous immunoassays have the advantage of being rapid, easy to perform, and amenable to automation. In homogenous immunoassays, one immunoreactant is coupled with an enzyme as in enzyme-multiplied immunoassay technique, or EMIT, with an enzyme fragment, as in the cloned enzyme donor immunoassay, or CEDIA, with a substrate as in substrate-labelled-fluorescent immunoassay, or SLFIA, with an inhibitor as in the enzyme-modulator mediated immunoassay, or EMMIA, with a cofactor as in the apoenzyme-reactivation immunoassay, or ARIS, with liposomes as in the liposome immune lysis assay, or LILA, or with a fluorescence molecule as in the fluorescence polarisation immunoassay, or FPIA. The first commercial available homogenous enzyme immunoassay was the competitive EMIT (figure 5.32.A). If the limited reagent antibody binds to an antigen/enzyme conjugate, the conformation or the steric accessibility of the catalytic centre is changed, leading to an inhibition of enzymatic activity. Another example for a widely used homogeneous technique is the competitive FPIA (figure 5.32.B). FPIA is based on the increase of the fluorescence polarisation of a small fluorescein-labelled antigen when bound by a specific antibody. The method is suitable for the measurement of small antigens, because a large increment of the analytical signal is achieved upon binding to a high molecular weight antibody. Homogenous immunoassays are predominantly used for low molecular weight analytes such as drugs or hormones.

A homogenous open-sandwich bioluminescent-immunoassay (OS-BLIA) that combines the advantages of homogenous and heterogeneous assays has been described recently (figure 5.32.C). An antigen induces an association between an antibody heavy chain fragment that is coupled to luciferase and an antibody light chain fragment that is conjugated to a yellow fluorescent protein. This coupling leads to bioluminescence resonance energy transfer between the two chains.

A Enzyme-multiplied immunoassay technique (EMIT)**B Fluorescence polarisation immunoassay (FPIA)****C Open sandwich bioluminescent immunoassay (OS-BLIA)****Figure 5.32 Homogenous Immunoassays.**

A) Enzyme-multiplied immunoassay technique (EMIT). EMIT is a competitive immunoassay based on the ability of an antibody to inhibit the enzyme activity of an enzyme/antigen (E/Ag) conjugate. The antigen of the samples competes with the conjugate for the limited number of antibody. In the presence of high antigen concentration in the sample, less antibody binds to the enzyme-antigen conjugate, which results in an increase of a coloured product produced by the enzyme. **B) Fluorescence polarisation immunoassay (FPIA).** FPIA is a competitive assay. The antigen of the samples competes with a fluorophore/antigen (F/Ag) conjugate for the limited antibody. In the assay, polarised light, which travels in a single plane, is focused on the sample containing antigen and conjugate. When the polarised light strikes the fluorescent molecule, the emitted fluorescence light is no longer polarised, because the

molecule is tumbling rapidly in addition to the Brownian molecule motion. If an antibody is bound to the fluorescence-labelled antigen, the large complex rotates more slower and emits polarised fluorescence light that passes through the polarisation filter. In the presence of high antigen concentration, less antibody binds to the fluorophore/antigen conjugate, which results in a decrease of polarised fluorescent light. **C) Open sandwich bioluminescent immunoassay (OS-BLIA).** OS-BLIA is a non-competitive assay based on the antigen-dependent association of an antibody light-chain fragment (V_L) with a heavy chain fragment (V_H). The association can be monitored by resonance energy transfer from *Renilla* luciferase (Luc) to the enhanced yellow fluorescent protein (EYFP), accompanied by a wavelength shift of the emitted light. Without the antigen, the antibody chains remain monomeric.

The result is a shift in the wavelength of emitted light. This principle was also described as an open-sandwich fluorescent immunoassay, which employs antibody fragments labelled with fluorescence molecules or green fluorescent protein variants (Arai *et al.*, 2001).

5.5.4

Types of label

Immunoassays measure the primary product of an immunological reaction, the formation of an antigen/antibody complex. The binding of an antigen to an antibody produces few physicochemical changes and therefore labelling or other techniques must be used to detect the reaction.

When antigen and antibody are mixed in nearly equal amounts, a macroscopic aggregate is generated. This aggregate can be visualised in a gel by radial immunodiffusion. Modifications of the gel methods for performance in solution are nephelometry and turbidimetry. These detect the aggregate by means of light scattering or changes in light transmission. In order to increase the agglutination signal, the antibody may be conjugated to small particles such as latex (latex agglutination assay or LAA). Particle-enhanced immunoassays such as the LAA are used widely in the clinical laboratory (von Schulthess *et al.*, 1976).

An interesting means of directly detecting antigen/antibody binding is by use of immunosensors, devices that consist of a surface coated with a capture antibody linked to a physico-chemical transducer (Stefan *et al.*, 2000). When the antigen binds to the antibody, changes in an electrochemical or even acoustic property, or refractive index lead to generation of an electrical signal. However, the development of immunosensors with the reliability and ruggedness of ion-sensitive electrodes still faces some problems. New approaches for regeneration of immunosensors without addition of elution reagents or the removing of non-specific-binding proteins have recently been described (Keay & McNeil, 1998). Commercially available immunosensor systems that operate by means of surface plasmon resonance such as the Biacore device (Biacore International AB, Uppsala, Sweden) are technically demanding optical bench-sized devices.

Usually, however, the binding of an antibody is detected by an auxiliary reaction, in which one immunoreactant is labelled with a substance which can be detected by optical, electrochemical, or mass sensitive measurements such as surface acoustic wave or microbalance. A great number of different labels, including isotopes, enzymes, fluorophores, fluorescence quenchers, chemiluminescence compounds, latex particles, liposomes, and redox compounds have been used. An overview of the various detection methods is given in table 5.8. For details and practical help, the reader is referred to the recent book by Kessler (2000).

In 1959, Rosalyn Yalow and Solomon A. Berson developed the first ever immunoassay, which was used to measure insulin levels. For this work, Rosalyn Yalow was awarded the Nobel prize for Medicine in 1977. This radioimmunoassay used insulin labelled with the γ -emitter ^{131}I iodine. The quantity can be determined by measurement of the radioactive decay. Most radioimmunoassays now use either

Table 5.8 Examples of labels for immunoassays.

<i>Immunoassay</i>	<i>Abbreviation</i>	<i>Substrates for detection</i>
Enzyme-linked immunoassay	ELISA	Alkaline phosphatase; horseradish peroxidase; β -galactosidase; urease; glucose-oxidase; luciferase; glucose-6-phosphate dehydrogenase
Fluorescence immunoassay	FIA	Fluorescein; coumarin derivatives; Eu^{3+} , Tb^{3+} and Sm^{3+} -chelate or cryptate; umbelliferone derivatives; Texas Red; rhodamine; phycoerythrin
Chemiluminescence immunoassay	LIA	Acridinium ester; luminol derivatives
Electrochemiluminescence immunoassay	ECLIA	Ruthenium(II)-tris(bipyridyl) complex
Bioluminescence immunoassay	–	Luciferin
Nephelometry turbidimetry	–	Latex-particles; inorganic colloidal particles
Radioimmunoassay	RIA	^{125}I iodine; tritium
Electrochemical immunoassay	–	Ferrocene
Ligand-binding assay	–	Metabolites; hormones; transport, binding or receptor proteins

^{125}I iodine or tritium, which offer higher sensitivity with less background interference. Radioactive iodine has the advantages of being cheap and capable of being readily attached to proteins. For these reasons, radioactive labels are still widely used, especially for the development of immunoassays for new proteins. However, the usefulness of radioimmunoassays is limited by the relatively short shelf life of radioisotopes, the potential health hazards, difficulties of automation and problems associated with radioactive contamination and waste disposal. For these reasons, much effort is being devoted to the development of non-radioactive immunoassays with comparable sensitivity to radioimmunoassays. The use of non-isotopic labels has facilitated the development of robust automation and assay reagents with longer shelf life.

5.5.4.1 Enzyme-linked immunosorbent assay (ELISA)

The enzyme-linked immunosorbent assay, or ELISA has become the workhorse of immunoanalysis. In the enzyme-linked immunosorbent assay, the antibody is conjugated to an enzyme to enable detection of binding to the antigen. The term enzyme-linked immunosorbent assay is often used interchangeably with enzyme immunoassay, enzyme-immunoassay and immunoenzymometric assay. The most frequently used enzymes are alkaline phosphatase from calf intestine, β -galactosi-

dase from *Escherichia coli* and horseradish peroxidase. Urease, glucose oxidase, and microperoxidase are used to a lesser extent. The enzyme is joined to the antibody by means of glutaraldehyde. The bound antibody is detected by adding a substrate to the reaction, which is converted by the enzyme to a coloured product. This test design has the advantage that even weak signals can be detected by means of enzymatic amplification. The substrates used are *p*-nitrophenyl phosphate for alkaline phosphatase, chlorophenol red- β -D-galactopyranoside for β -galactosidase and 2,2-azino-bis-[3-ethyl-benzthiazoline] sulphate for horseradish peroxidase. Depending on the substrate, measurement may be based on fluorescence, luminescence, or electro-physical detection. The detectable species may be a direct product of the enzyme reaction (direct substrate), or a detectable product formed by one or more additional reactions (indirect substrate) (figures 5.33.A and 5.33.B). Increase in sensitivity and acceleration of the detection reaction may be obtained by using an enzymatic cascade (figure 5.33.C). The enzyme may be conjugated either directly to the antibody or to biotin molecules, which are then linked by means of streptavidin to a biotin-labelled antibody. The latter system also amplifies the signal strength of the assay by increasing the amount of enzyme linked to each antibody (figure 5.33.D) (Crowther, 2001).

Proteins secreted by groups of cells may be analysed in biological fluids or cell culture supernatants. However, it may also be of interest to investigate the products secreted by individual cells. In the 'capture ELISPOT' technique, cells are incubated in microwell-plates coated with a capture antibody directed against the cell product of interest. After washing, an enzyme-linked antibody is added to form a sandwich complex. Quantification is based on cell numbers. An enzyme-linked immunosorbent assay with whole cells attached to microtiter plates is also available for the detection of tumour antigen and may be used to quantify protein expression in cells transfected with plasmids or infected with recombinant viruses (Erdile *et al.*, 2001).

The capture enzyme-linked immunosorbent assay uses microwell-plates coated with an antibody or protein to bind the antigen of interest. An innovative variant of this assay uses glutathione cross-linked to casein as a capture protein to bind recombinant protein antigens fused to amino-terminal glutathione S-transferase (figure 5.34.A) (Sehr *et al.*, 2001).

Albert H. Coons and co-workers introduced fluorochromes for use as antibody labels in 1941 (Coons *et al.*, 1941). Fluorescence is a type of photoluminescence in which fluorochrome molecules are excited by light of one wavelength to emit light of a longer wavelength. Fluorescence is much more sensitive than photometry. Antibodies conjugated to alkaline phosphatase, which cleaves 4-umbelliferyl phosphate to yield a fluorescent product are often used in fluorescence-immunoassays.

Hartmut R. Schroeder and colleagues described the first chemiluminescence immunoassay in 1976 (Schroeder *et al.*, 1976). Chemiluminescence refers to the conversion of chemical energy into light and is the most sensitive optical signal available for enzyme-labels. The analytical sensitivity is often better than that of radioactivity. Measurements are also faster. No external light source is required. All light reaching the detector originates from the chemical reaction with none of the interference from scattered excitation light that besets fluorescence detection. Chemilu-

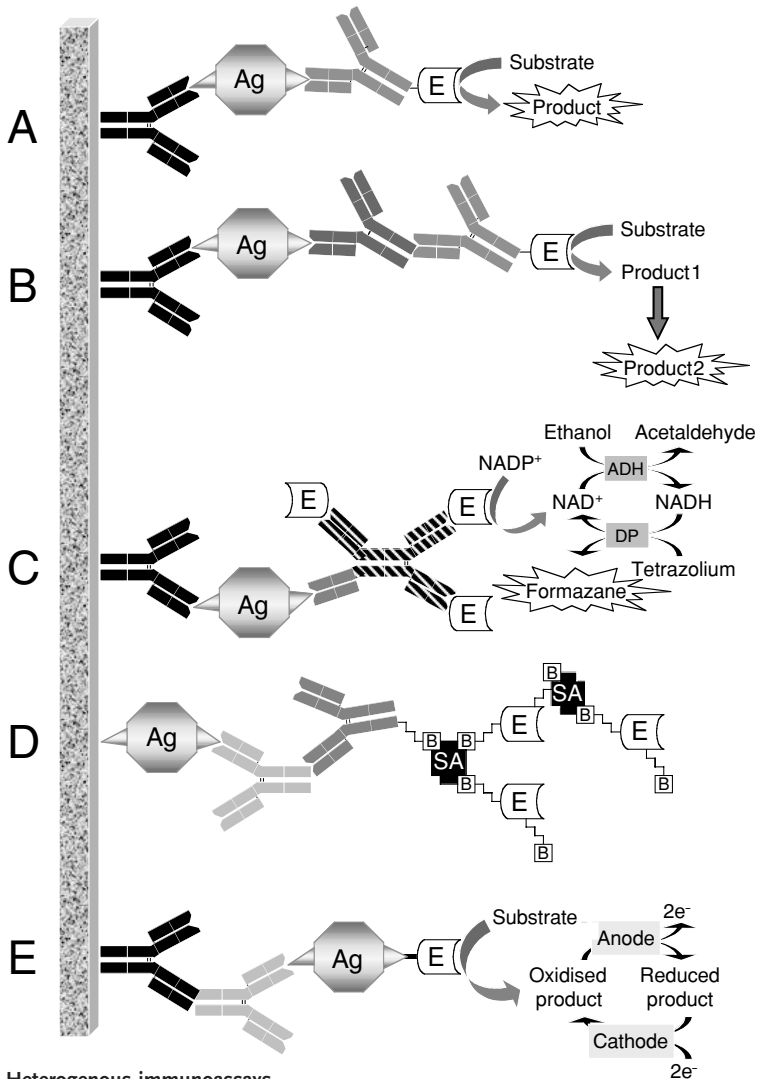


Figure 5.33 Heterogeneous immunoassays.

A) Direct immunometric-enzyme immunoassay with direct signal generation. A substrate is acted on by an enzyme to produce a coloured, fluorescent or luminescent end-product that can be readily measured. **B)** Indirect immunometric-enzyme immunoassay with a coupled detection reaction. **C)** Direct immunometric-enzyme immunoassay with a chimeric detection antibody. A fusion protein is created containing an antigen-binding site linked to an anti-enzyme antibody. Increase in sensitivity and speed of detection is achieved by means of an enzyme enzymatic cascade consisting of alco-

hol dehydrogenase (ADH), diaphorase (DP), and the oxidised (NAD^+) or reduced (NADH) form of nicotinamide-dinucleotide. **D)** Direct immunometric-enzyme immunoassay for the detection of antibody with defined specificity. Signal amplification may be achieved using biotinylated (B) enzymes linked to streptavidin (SA) molecules. **E)** Direct capture enzyme immunoassay with an immobilised capture antibody can be used for the detection of specific antibody. An antigen-enzyme complex in combination with an electrochemical detection system is utilised.

minescence-based immunoassays have led to a dramatic reduction of detection limits, resulting in the addition of the zeptomole (10^{-21} mol) and even the yoctomole (10^{-24} mol) to the vocabulary of the clinical chemist (Kricka, 1993).

In 1,2-dioxetane chemiluminescent systems, enzymatic cleavage of a phosphate or β -galactoside derivative forms an unstable intermediate, which decomposes to emit light with a relatively long decay time. The chemiluminescence signal can be amplified by coupling it to fluorophores that are secondarily activated by the light of the luminescence reaction.

Bioluminescence, a phenomenon found in many simple life forms, is based on a two-step reaction. First, D-luciferin is released from D-luciferin derivatives by enzymatic hydrolysis. Then, the enzyme luciferase catalyses the oxidation of D-luciferin to oxyluciferin. The latter reaction is accompanied by the emission of light. Recently, the use of other recombinant photo-proteins such as aequorin and obelin for direct labelling of antibodies has also been described. Microtiter plate readers are now available for use with chemiluminescent or bioluminescent-based immunoassays (Roda *et al.*, 2000).

In electrochemical immunoassays, the antibody/antigen complex is attached directly to an electrode. Electrochemical immunoassays have been developed in competitive and non-competitive formats with redox compounds or enzymes serving as labels. For example, alkaline phosphatase may be used to cleave *p*-aminophenyl phosphate to the electrochemically active compound *p*-aminophenol, which is then specifically oxidised and quantified at an electrode (figure 33.E). The development of simple handheld devices for bedside analysis in the hospital or for patient self-testing is a fast-developing technology in this area (Warsinke *et al.*, 2000).

5.5.4.2 Direct label of antibody

It is also possible to label antibodies directly using highly sensitive detection systems that do not need an enzyme in order to achieve the same detection limits as enzymatically amplified systems. For certain applications, enzymes may suffer from disadvantages, including dependence of their catalytic rate on temperature, pH, and ionic strength. In addition, enzymes may be unstable at high temperature and are subject to chemical degradation.

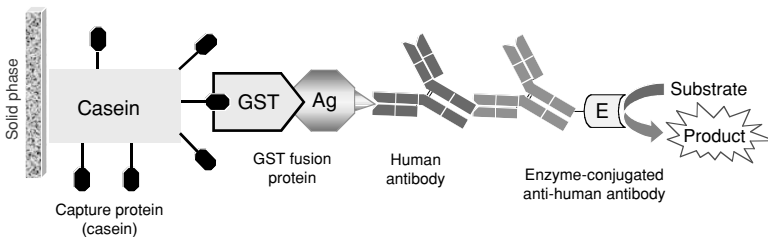
Commonly used direct fluorescence labels include fluorescein, the coumarins, resorufin, oxazine, derivatives of cyanine and rhodamine, and boradiazaindacene (BODIPY[®]; dipyrrometheneboron difluoride) fluorophores. The sensitivity of these compounds is exploited in fluorescence immunoassays, flow cytometry, fluorescence-activated cell sorting, and fluorescence microscopy. The detection and identification by laser technology of a single labelled antibody has been published. In the liposome immune lysis assay, the fluorescence label and other molecular components of the assay are transported in liposomes that are lysed once the antibody binds to the antigen (Yu *et al.*, 1987).

One of the main problems of the fluorescence immunoassay is the difficulty in separating the fluorescence emitted by the sample from interference produced by the excitation light source and light scattering. In addition, background fluores-

cence may be produced from substances in the sample. One way to circumvent these problems is by means of time-resolved fluorescence detection using rare-earth chelates or cryptates such as europium (Eu^{3+}) or terbium (Tb^{3+}) cations. These elements exhibit the property of delayed fluorescence. After the excitation light source is extinguished and background fluorescence has faded, the long-lived metal fluorescence persists and can be measured (Dickson *et al.*, 1995).

Acridinium ester derivatives are also widely used as direct chemiluminescent labels; the light-producing reaction in this case is triggered by the addition of hydrogen peroxide. Luminol derivatives can also be coupled to antibodies to be used as

A Capture enzyme-linked immunosorbent assay (ELISA)



B Electrochemiluminescence immunoassay (ECLIA)

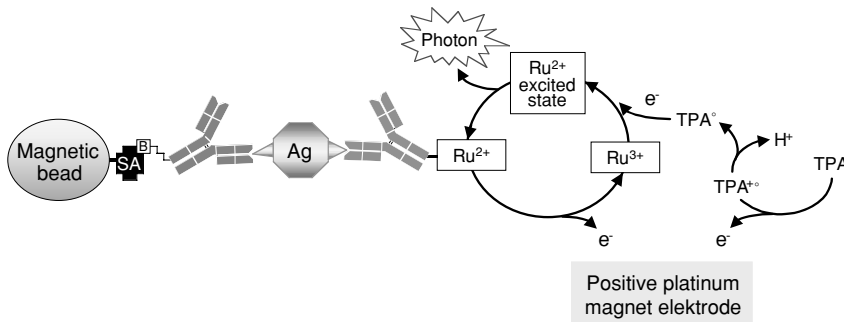


Figure 5.34 Principle of enzyme-linked immunosorbent assay (capture ELISA) and electrochemiluminescence immunoassay (ECLIA). **A)** Capture enzyme-linked immunosorbent assay (ELISA). A solid phase coated with glutathione casein is used to capture a recombinant antigen via amino-terminal glutathione S-transferase (GST). Bound antibodies from human serum are detected by a secondary polyclonal anti-human antibody conjugated to an enzyme (E). **B)** Electrochemiluminescence immunoassay (ECLIA). Suspensible paramagnetic particles (e.g., iron oxide, chromium dioxide) are

used as solid-phase for immunometric assays. They are coated with streptavidin (SA) in order to bind a biotinylated antibody specific for the test antigen. A secondary antibody is coupled to a ruthenium(II)-tris(bipyridyl) complex. Ruthenium ions and tripropylamine (TPA) are oxidised at a platinum electrode. After release of a proton, the tripropylamine radical reacts with Ru^{3+} to generate excited Ru^{2+} , which decays with the emission of a photon and regenerated a Ru^{2+} in the basic state. This reaction can go through multiple light-generating cycles.

direct labels; signal generation here requires the addition of peroxidase and hydrogen peroxide (Seitz, 1984).

Other direct labels are ruthenium (Ru^{2+}) cation complexes that produce electro-generated chemiluminescence (figure 5.34.B). In the electrochemiluminescence immunoassay, ruthenium(II)-tris(bipyridyl) complexes ($[\text{Ru}(2,2'\text{-bipyridine})_3]^{2+}$ or $[\text{Ru}(\text{phenanthroline})_3]^{2+}$) labels are oxidised at the electrode. At the same time, an electron donor generated at the same electrode reduces the ruthenium complex to $\text{Ru}[(2,2'\text{-bipyridine})_2]^{2+}$, hereby emitting light. This reaction can be repeated multiple times (Blackburn *et al.*, 1991).

In the immuno-polymerase chain reaction, a nucleic acid is employed as label. Following binding of the antibody to the analyte, the nucleic acid is amplified by the polymerase chain reaction and detected using a complementary probe (Case *et al.*, 1999).

5.5.5

Interferences

Figure 5.35 lists some of the most important factors that may interfere with immunoassays and cause false results. In a sandwich assay, due to what is known as the high-dose hook effect, the signal can be paradoxically low at very high antigen concentrations. This is caused by the antigen binding only to the second or labelling antibody, but not to the first or capture antibody, which means that the sandwich complex fails to form. This problem is solved by diluting the sample or by adding a smaller volume of sample to the assay (Wolf & Brem, 1991).

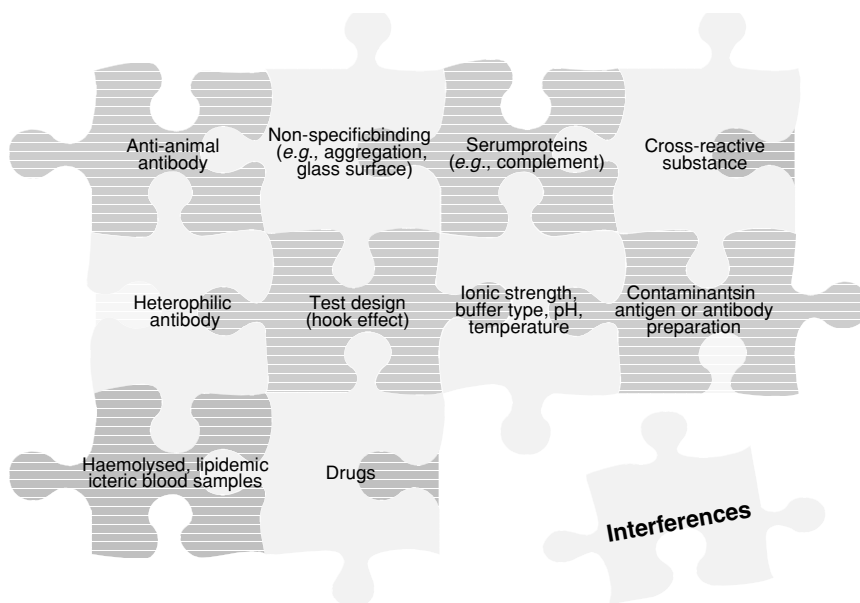


Figure 5.35 Factors interfering with immunoassays which may lead to alterations in the apparent concentration of analyte.

In order to ensure precise measurements, temperature, incubation time, intensity of washing, and volume of reagent and sample must be carefully controlled. The accuracy of the assay must also be regularly checked using quality control material. Any changes in the test conditions require a fully recheck of the quality control parameters sensitivity, precision, linearity and specificity. In practise, this is best achieved by automation, which allows one to hold all parameters constant during the period of measurement. Immunoassays have been adapted for use with highly sophisticated automated analysers. These allow the precise analysis of a great number of samples within a short time. Such auto-analysers can utilise both homogenous and heterogeneous assays. A wide range of standard procedures and reagents are available. International reference preparations for proteins and external quality proficiency testing are important for standardisation and comparability of commercial and in-house immunoassays (Plebani & Carraro, 1997).

Cross-reactivity is the bugbear of immunoassays. The antibody often binds substances similar to the analyte. Substances chemically similar to the analyte may therefore be added to the sample in order to define the extent of cross-reactivity.

Applying the assay to a different biological matrix, for example, measurement in plasma instead of serum, may also cause discrepant results. Some analytes exhibit non-specific binding to proteins such as albumin, to plastic or to glass. Other proteins tend to self-aggregate, particular at high concentrations. The biological material in which the measurement is being performed may interfere with the detection system. Examples are interference due to the colour of haemoglobin, the fluorescence of bilirubin, or the turbidity of lipids. Chemical compounds present such as drugs or components of cell culture medium and other exogenously added substances may also interfere with the assay. To resolve the latter problem, special dyes with distinct spectral properties may be used, or the sample may be processed before the measurement is performed (Wild, 2001).

Most of the monoclonal antibodies used in immunoassays are of mouse origin. For this reason, human anti-mouse antibodies (figure 5.36) present in a patient sample may interfere with a sandwich immunoassay by bridging between the capture antibody and the detection antibody, producing a false-positive result. False-negative or low results may also be observed if human anti-mouse antibodies react with only one of the antibodies or prevent the binding of antigen. Anti-animal antibodies are not rare and are estimated to be present in between ten and forty per cent of human sera. They are thought to result from the increasing use of diagnostic and pharmaceutical substances of animal origin, from the keeping of animals, and from the ingestion of food contaminated with animal antigens. Interference from human anti-mouse antibodies can be reduced by the use of antibody-fragments that lack the Fc domain, by use of humanised or chimaeric antibodies, or by the chemical modification of antibodies by polyethyleneglycol or other substances. Blocking reagents such as non-immune mouse serum may be added to the assay in order to prevent interference by anti-animal antibodies. The sample may be also subjected to immunoextraction using murine antibody or protein G before the assay is performed. Several blocking reagents and kits for the measurement of human anti-mouse antibodies are commercially available (Klee, 2000).

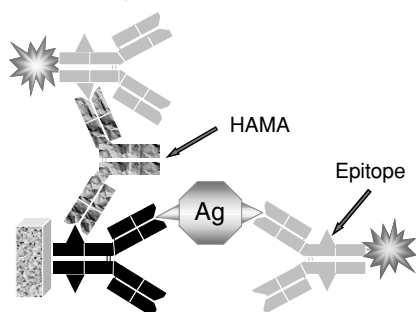
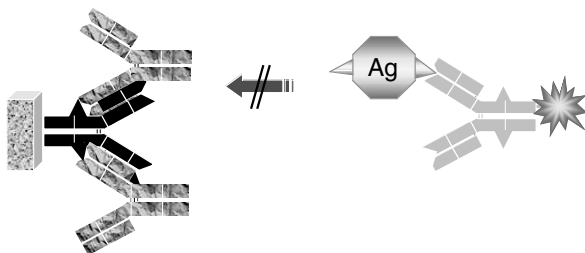
A False-high results**B False-low results**

Figure 5.36 Mechanisms of interferences in immunoassays. **A)** Falsely high test values. Human anti-mouse antibodies (HAMA) form a bridge between the capture and the signal antibody by binding to epitopes on the Fc part

of each antibody. **B)** Falsely low test values. Human anti-mouse antibodies bind to the capture antibody, thereby blocking the paratope.

Human anti-mouse antibodies are often confused with heterophile antibodies, a class of antibodies with low specificity and low antigen affinity. Heterophile antibodies are often seen in patients with chronic infections such as hepatitis or tuberculosis, and in patients suffering from autoimmune diseases such as *lupus erythematosus*. Interference by heterophile antibodies is eliminated using the same strategies outlined above for human anti-mouse antibodies (Kricka, 1999).

5.5.6

Conclusions

Despite many developments in the analysis of protein expression, immunoassays remain the cornerstone of measuring proteins in biological matrices. Following the groundbreaking introduction of this technique by Rosalyn Yalow and Solomon A. Berson some 40 years ago (Yalow & Berson, 1959), technical improvements have increased specificity and sensitivity to the point where it is now possible in certain circumstances to assay for single molecules. At the same time, the cost of analyses

has fallen dramatically and fully-automated high-throughput methods have been developed for most common applications. Great ingenuity has been applied to the development of particular test constellations for particular experimental circumstances, which has led to an explosion of sometimes confusing acronyms. Nevertheless, the fundamental principles of immunoassays remain the same in all cases. In the post-genomic era, the analysis of proteins in biological matrices can be expected to increase even further in importance.

5.6

Flow cytometry and protein expression analysis

5.6.1

Introduction

Flow cytometry is a powerful method for the rapid analysis, screening, and characterisation of cells. In recent years, cytometry has played a prominent role in the study of gene expression at the protein level. The rise in popularity and prevalence of this type of analysis is due, in large part, to the development and increased availability of fluorochrome-conjugated antibodies and reporter genes such as green fluorescent protein (GFP). These fluorescent compounds, when used separately or collectively, make it possible to label and detect, with high levels of specificity and sensitivity, the expression of one or more specific gene products for individual cells. Here we illustrate, by using fluorescent markers, how flow cytometry can be used to analyse two different types of gene products, intracellular and surface-bound proteins. Applications that have benefited from this technology include screening cells for successful gene transfection or transformation, determining expression patterns of cell surface markers, studying enzyme kinetics, monitoring proteins produced during viral infection, or expression profiling of gene products or other molecules involved in important cellular processes such as apoptosis or cell cycle regulation. This section is intended to serve as a brief overview, and as such, cannot adequately encompass the elegant and sophisticated technology of flow cytometry and cell sorting; however, many excellent reviews and books have been written on the subject (Diamond & DeMaggio, 2000; Shapiro, 1994).

5.6.2

Principles and basics

Reporter systems such as GFP work by using recombinant DNA technology to clone the reporter gene into a vector. This vector, such as a virus, is then incorporated into a eukaryotic cell by a process called transfection. When successful, the cells own biochemical machinery activates expression of the introduced gene to produce a fluorescing protein that can be detected and analysed. Alternatively, cells may be prepared and stained using fluorescently tagged primary, and possibly secondary, antibodies directed against specific protein species. After the antibodies

bind to the designated protein target(s), the fluorescent signals can be used to detect or quantify the gene product.

Large numbers of individual cells containing one or more fluorescent markers can be rapidly analysed using a flow cytometer to identify subpopulations with desired gene expression characteristics. For this process, the cells are generally suspended in a fluid medium that is passed through the flow cytometer. The liquid sample is positioned coaxially within a larger carrier stream and is illuminated with a light source, typically using either an arc lamp or laser(s). Carried by the stream, individual cells pass sequentially through these illumination points where information is obtained on each cell. This includes the real-time detection of reflected and refracted light and laser-stimulated emission of the fluorescent markers.

Refracted light from the cells is detected by a photodiode coupled with a notch or bandpass filter specific for the wavelength of the light source. The waveform of the forward scatter refraction is loosely associated with the size of a cell or particle, while reflected light (or side scatter) is associated with granularity or other surface characteristics. Emitted fluorescence is collected orthogonally by detectors or photomultiplier tubes using various combinations of filters and dichroic mirrors designed to provide the best optical resolution of the fluorescence signals elicited from each cell (figure 5.37). Because the filters are configured to allow only certain wavelengths of light to reach a given detector, each photomultiplier tube becomes specific for detecting fluorescent compounds with compatible emission spectra such as fluorescein isothiocyanate (FITC), 4',6-diamidino-2-phenylindole (DAPI), GFP, or Texas Red. Photomultiplier tubes convert this reflected or fluorescent light into an electrical signal that describes the intensity of light from each cell. These signals, each describing a light scatter or fluorescent parameter, are then subjected to analogue to digital conversion. The result is a digital description

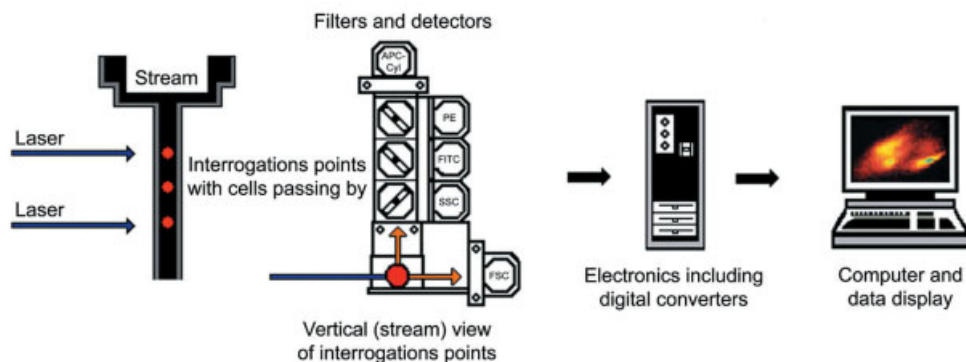


Figure 5.37 A detailed view of cells flowing past the interrogation points within the stream. Light and fluorescent signals are passed through various filters to detectors that convert the information to electrical signals. These signals are then digitalised, and the resulting data

is displayed and analysed using a computer. The following abbreviations are used: allophycocyanin (APC), fluorescein isothiocyanate (FITC), forward scatter (FSC), phycoerythrin (PE) and side scatter (SSC).

that is assigned to a single event that contains an intensity value for each fluorescent or light scatter 'colour.' Thus, the display and interpretation of the data becomes paramount. Data may be visualised by a number of methods including a histogram, a dot plot, a contour plot, or some innovative combination of these using any combination of light scatter or fluorescent parameters. Statistics describing the entire population or sub-populations may be calculated; further, sub-populations may be subjected to individual display or analysis by gating.

Analysers and cell sorters are commercially available with the capacity to process upwards of 50,000 events per second and collect signals for up to nine fluorescent channels and two light scatter parameters, including forward and side scatter. However, it is the ability to rapidly detect, quantify and analyse one or more gene products of cells using fluorescent labels that has allowed this system to become a powerful tool in the understanding of gene expression and function. The expression patterns of multiple proteins such as cytokines, cell surface markers, receptor proteins, enzymes, or oncogene products can be simultaneously monitored and studied within phenotypically defined cells, including those involved in complex cellular processes such as apoptosis, oncogenesis, cell division, or differentiation. Cell sorters provide the added benefit of allowing one to isolate and collect cells containing interesting expression patterns for further study.

The following examples describe a multiparametric approach to gene product analysis involving (i) a transfected reporter gene system and (ii) the use of fluorescently labelled antibodies targeting native surface antigen expression. Both examples reflect procedures that are in common usage and generally easy for the novice cytometrist to accomplish.

Data for both examples was collected on a MoFlo[®] cytometer and high speed cell sorter from Cytomation, Inc., (Fort Collins, Colorado, United States of America). All configuration notes apply specifically to this instrument but may be extrapolated to other types of cytometers.

5.6.3

Results

5.6.3.1 Example 1: Analysis of dual fluorescent protein transfection targets

In this first example, eukaryotic cells were transfected using two different viral constructs, the first containing a strong constitutive promoter linked to the gene encoding GFP, and the second containing a strong constitutive promoter linked to the gene encoding cyan fluorescent protein, CFP. Successful transfection of the individual cell is marked by transcription and translation of the fluorescent gene products.

Cytometer configuration. The cells were sequentially illuminated with 488 and 457 nanometres laser lines. An argon laser was operated at 488 nanometres in light regulation mode at 100 megawatt and provided illumination to the forward scatter photodiode that is used to trigger data acquisition. A second argon laser provided 457 nanometres illumination at 100 megawatt in light regulation mode and was focused on the stream independently of the first laser. GFP fluorescence was

collected using a 570/40 bandpass filter, while CFP was detected using a 485/22 bandpass filter.

Data visualisation. Data is represented using a false-colour dot plot. This type of display uses colours to represent increasing numbers of events that occupy the same x and y coordinates. The relevant parameters being used for analysis are indicated with labels along the histogram axes. Data for some parameters was collected using linear amplification, while others used logarithmic amplification, as indicated by the appropriate scale. Linear amplification is typically used for nearly homogeneous populations that possess small variance and is generally used for light scatter collection of cultured populations. Log amplification is more appropriate for those populations having a large difference between the smallest value and the largest and is typically used for fluorescence collection. Regions have been created within each plot for both statistical and gating purposes. For example, based on the primary light scatter parameters, the gate using region *R1* allows only those cells of interest to be included for fluorescent analysis (see figure 5.38); all outliers are excluded. This gating discrimination for a region ensures that only members of the homogenous population are displayed on the fluorescence dot plots for fluorescent gene product analysis.

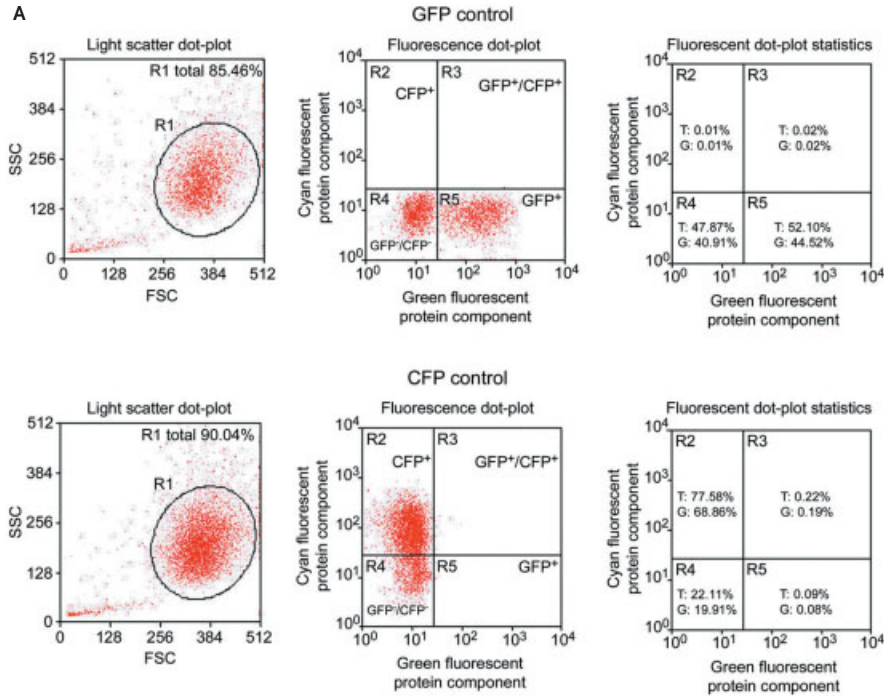
Fluorescence compensation. All fluorescence dot plots have been compensated to clearly discriminate and resolve the spectral overlap from individual GFP and CFP signals. In the case of GFP, the fluorescence intensity was modified by subtracting 6.03 percent of the intensity of the CFP on a per event basis. For CFP, the compensation was 12.99 percent of the CFP signal, again on a per channel basis. Thus, the GFP-transfected control does not appear in the CFP-positive quadrant and *vice versa*.

Figure 5.38 summarises the results of this experiment. For all four data sets, light scatter gates are shown. These plots demonstrate that the light-scatter homogeneity of the populations remain relatively unaffected by the transfection status. It is important to note that small cell fragments and other debris are not included in the fluorescence analysis. All elliptical regions are positioned identically, and the percent of total events constrained by this region is annotated within each histogram. Statistically significant variance between one or more of the percentages

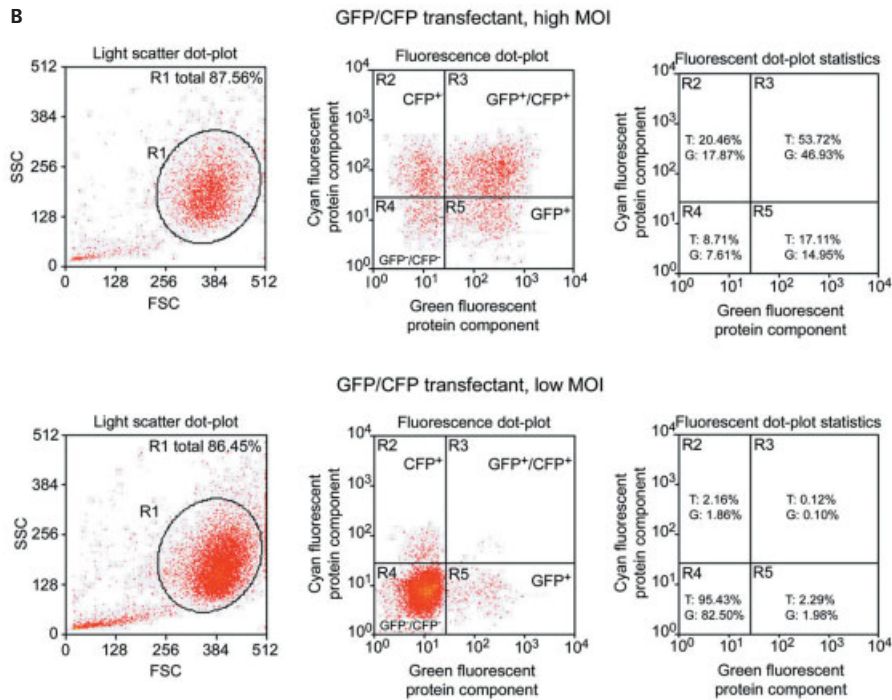
Figure 5.38 Analysis of dual fluorescent protein transfection targets (green fluorescent protein and cyan fluorescent protein). Four representative data sets are illustrated (left and middle). Diagrammatic plots with quadrant statistics for the fluorescence dot plots are on the right. **A) Control cells. Upper row:** Control light scatter and fluorescence dot-plots depict the transfection of cells with a single gene construct encoding green fluorescent protein (GFP). **Lower row:** The control light scatter and fluorescence dot plots depict the transfection of cells with a single gene construct encoding cyan fluorescent protein (CFP). Both controls are

infected at a high ratio of virus numbers to cells. **B) Transfected cells. Upper row:** The dual transfected, high multiplicity of infection (MOI) light scatter and fluorescent dot plots show a population double-positive GFP and CFP cells. **Lower row:** The final data set, the dual transfected, low multiplicity of infection light scatter and fluorescent dot plots depict a paucity of double-positive cells (FSC: forward scatter; G: gated events; SSC: side scatter; T: total events). Data described in this figure was contributed by Dr. Richard Cross of St. Jude's Children's Research Hospital (Memphis, Tennessee, USA).

A



B



within region *R1* may indicate a problem during sample preparation. This was not the case for this example, but under such circumstances, the investigator should usually re-analyse the suspect samples. The right-hand columns of the fluorescence dot-plots illustrate the differences in gene product expression based on the transfected gene construct. Quadrants are identically positioned within all four dot plots, and each quadrant is annotated with the percent of total events, *T*, and the percent of the gated events, *G*. The gated percent does not include those events near the origin of the corresponding light scatter histograms. In both, the GFP and CFP control fluorescence dot plots, the individual signals of the expressed proteins are compensated correctly and, thus, do not contribute to the other's fluorescence. The two controls were transfected at a high multiplicity of infection, thus ensuring each cell received at least one copy of the gene construct. The dual GFP and CFP transfection target was infected at both a high and low multiplicity of infection. Among the four cell populations, the differences between the resulting gene expressions are strikingly clear, demonstrating that the detection of one or more gene products by cytometry is an effective and rapid means for assessing the prevalence of transfection within a nearly homogeneous population.

5.6.3.2 Example 2: Analysis of surface antigen expression

In this second example, flow cytometry is used to characterise murine dendritic cells based on variable expression of cell surface markers such as cluster of differentiation 4 and 8 (CD4 and CD8). Previous studies involving these cells have identified three sub-populations based on unique expression patterns of surface-bound proteins. The large dendritic cells typically lack or exhibit low expression of both CD4 and CD8. The small dendritic cells, on the other hand, can be divided into two groups, those expressing CD4 in the absence of CD8 and those expressing CD8 without CD4.

The sample was prepared from lymph nodes, and the resulting cell suspension was analysed using fluorescent monoclonal antibodies: FITC-conjugated anti-major histocompatibility (MHC) class II, phycoerythrin-conjugated anti-CD11c, phycoerythrin/Cy7-conjugated anti-CD4 and peridnin chlorophyll protein-conjugated anti-CD8.

The light scatter histogram depicts a single population composed primarily of lymphocytes. Note the relative lack of macrophage/monocyte and granulocyte populations (see figure 5.39). The contaminating smaller cells are excluded from analysis by exclusion or Boolean 'NOT' gating of all events outside of region *R1* from the light scatter dot plot.

The top middle dot plot shows the dendritic cell distribution of MHC class II and CD11c staining. Here, two populations are discerned. The first is constrained by region *R4* and contains the large dendritic cell population. The second, within region *R3*, contains the small dendritic cell population. Since the small dendritic cell population is not clearly resolved in the MHC class II/CD11c dot plot, a contour plot of this data was used to provide a clear representation of this population. The biological distinction between large dendritic cells and small dendritic cells

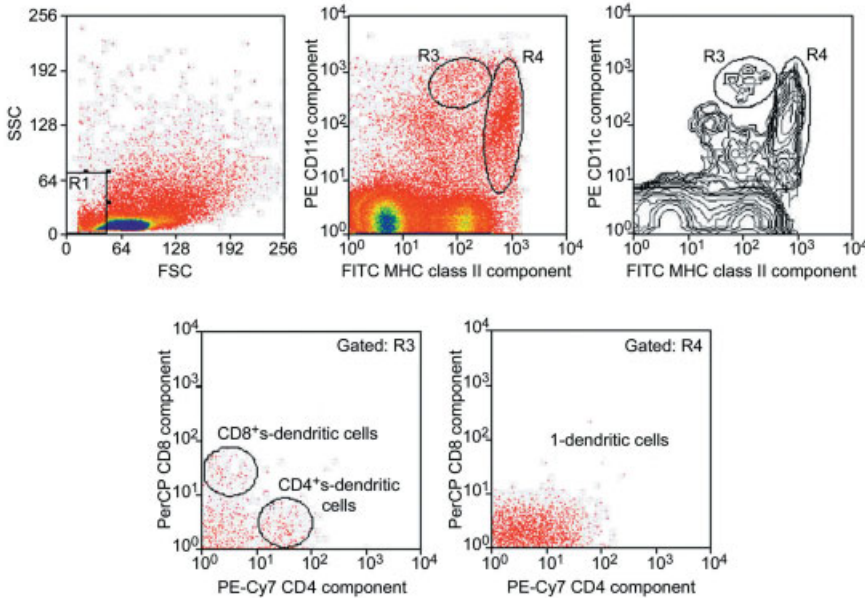


Figure 5.39 Analysis of surface CD4 and CD8 antigen expression. Murine lymph node dendritic cells stained with FITC-conjugated anti-major histocompatibility (MHC) class II, phycoerythrin-conjugated (PE) anti-CD11c, phycoerythrin/Cy7-conjugated (PE-Cy7) anti-CD4 and peridnin chlorophyll protein-conjugated (PerCP) anti-CD8. The light scatter histogram depicts a single population composed primarily of lymphocytes. Region *R1* was used to exclude smaller cells and debris from analysis. The top middle dot plot shows the dendritic cell distribution of MHC class II and CD11c staining. Two populations, large (l) dendritic cells and small (s) dendritic cells, are defined by regions *R4* and *R3*, respectively. Because the small dendritic cell population is not clearly resolved in the MHC class II and CD11c dot plot, this

same data is displayed as a contour plot that clearly shows the small dendritic cell population. The phycoerythrin/Cy7-conjugated anti-CD4 and peridnin chlorophyll protein-conjugated anti-CD8 analysis of the large dendritic cell population, bottom right and gated from region *R4*, shows a single well-defined grouping of $CD4^{\text{low}}/CD8^{\text{negative}}$ cells. This same phycoerythrin/Cy7-conjugated anti-CD4 and peridnin chlorophyll protein-conjugated anti-CD8 analysis of the small dendritic cell population, bottom left and gated from region *R3* shows a broad distribution of CD4 and CD8 staining. The cells expressing $CD4^{\text{intermediate}}/CD8^{\text{low}}$ and $CD4^{\text{low}}/CD8^{\text{intermediate}}$ antigen profiles represent two distinct functional dendritic cell types, each with a distinct lymphoid origin and function (FSC: forward scatter; SSC: side scatter).

arises from their capacity to take up and process antigens differently and in their maturation and turnover dynamics.

Phycoerythrin/Cy7-conjugated anti-CD4 and peridnin chlorophyll protein-conjugated anti-CD8 analysis of the large dendritic cell population gated from region *R4* shows a single well-defined grouping of $CD4^{\text{low}}/CD8^{\text{negative}}$ cells. This same phycoerythrin/Cy7-conjugated anti-CD4 and peridnin chlorophyll protein-conjugated anti-CD8 analysis of the small dendritic cell population gated from region *R3* shows a broad distribution of CD4 and CD8 staining which may be resolved

into two functionally distinct populations based on the expression profiles of CD4 and CD8. The cells expressing CD4^{intermediate}/CD8^{low} and CD4^{low}/CD8^{intermediate} antigen profiles represent two distinct functional dendritic cell types, each with a distinct lymphoid origin and function.

This example demonstrates the ability to identify sub-populations based on variable protein expression patterns. This type of complex expression profiling can be used to address functional studies, expression dynamics, and other studies of biological relevance.

5.6.4

Discussion

The examples provided here demonstrate the utility of flow cytometry for analysing gene expression. However, while providing a great deal of information, flow cytometry requires a number of experimental considerations, including a comprehension of the unique emission spectra of the various fluorescent compounds used to label the cells. This is necessary in order to properly configure the filters for detecting the discrete fluorescent signals, and to minimise the spectral overlap or 'crosstalk' among fluorochromes. This last issue is beyond the scope of this chapter, but represents one of the foremost obstacles to complex, multiple colour experiments. Generally, the more complex an analysis scheme, the more the researcher depends on the care and experience of a well-trained cytometrist.

Although a variety of techniques are available to study gene expression, regulation, and function, flow cytometry is differentiated by its ability to rapidly obtain large quantities of information about cells expressing a particular gene. Flow cytometry has the added advantage of being able to detect and analyse multiple gene products simultaneously. This is important to allow examination of the activities of gene cascades involved in complex biological processes such as differentiation, apoptosis, or organogenesis.

5.7

Summary

As has been commented by many of the authors of this chapter, a main result to emerge from the sequencing of the human genome is the realisation that a surprising amount of diversity in gene expression is generated at the transcriptional and post-transcriptional levels. Not only is there a very large degree of variant splicing, producing multiple transcripts from a single mRNA, but each transcript may be subject to cleavage and to a wide array of covalent modifications. It is thus abundantly clear that proper analysis of gene expression requires not only examination of intermediate message but also of end-product.

The analysis of proteins is in many ways a more difficult nut to crack than the examination of nucleic acids. Nucleic acids are fairly homogenous from a chemical point of view, but this cannot be said for proteins. Indeed, it is this very diversity in

the chemical properties of proteins which so suits them to their task as the main structural and functional building blocks of living organisms.

However, the diversity of proteins has been matched by the ingenuity of researchers in developing tools to unlock their secrets. Advances in mass spectrometry now allow detection, identification and quantification of minute amounts of even very large and recalcitrant proteins. Many of these techniques may be adapted for high-throughput applications. The use of the exquisite specificity of antibodies has also been exploited in a dizzying variety of immunoassays and flow cytometric applications.

Despite these technical advances, a bottleneck still remains in the analysis of proteins, namely their physical separation into isolated entities. At the time of writing, the mainstay of protein separation remains the two-dimensional polyacrylamide gel, a technique that was introduced over thirty years ago. For this reason, much interest is focused on the use of two-dimensional arrays to examine protein expression without the need for extensive preparatory steps. While not routine, this is an approach that holds much promise providing the technical difficulties in assembling such diverse molecules on a single surface can be overcome.

In the future, we may expect this trend towards analysis of proteins to increase in exponential fashion. Analysis of nucleic acid expression will not be supplanted, but will be complemented by this development. Ultimately, we may realise the vision of obtaining a complete picture of gene expression in a living cell.

5.8

References

Recommended readings

a) Introductory readings

Godovac-Zimmermann, J., and Brown, L. R. Perspectives for mass spectrometry and functional proteomics. *Mass Spectrometry Reviews* **2001**; 20(1):1-57.

Liebler, D. C. Introduction to proteomics: tools for the new biology. Humana Press, Totowa, New Jersey, USA **2001**.

Suhai, S. Genomics and proteomics: functional and computational aspects. Plenum Publishers Corporation, New York, New York, USA **2000**.

Structural Genomics supplement. *Nature Structural Biology* **2000**; 7(1):927-994.

Nature insight contributions on functional genomics. *Nature* **2000**; 405(6788):819-868.

b) Mass spectrometry

Beynon, R., and Hubbard, S. Mass spectrometry. Bios Scientific Publishers, Oxford, United Kingdom **2002**.

Hoffmann, E. D., and Stroobant, V. Mass spectrometry: principles and applications. John Wiley & Sons, Chichester, United Kingdom **2001**.

Burlingame, A. L., Carr, S. A., and Baldwin, M. A. Mass spectrometry in biology and medicine. Burlingame, A. L., Carr, S. A., and Baldwin, M. A., editors. Humana Press, Totowa, New Jersey, USA **2000**.

Dass, C. Principles and practice of biological mass spectrometry. Wiley-Interscience series on mass spectrometry. Wiley Interscience, New York, New York, USA **2000**.

c) Proteomics

James, P. Proteome research: mass spectrometry. Principles and practice. Springer-Verlag, Berlin, Germany **2001**.

Snyder, A. P. Interpreting protein mass spectra. A comprehensive resource. Oxford University Press, Oxford, United Kingdom **2000**.

Kinter, M., and Sherman, N. A. Protein sequencing and identification using tandem

- mass spectrometry. Wiley Interscience, New York, New York, USA **2000**.
- Chapman, J. R. Mass spectrometry of proteins and peptides. Chapman, J. R., editor. Humana Press, Totowa, New Jersey, USA **2000**.
- Rabilloud, T. Proteome research: two-dimensional gel electrophoresis and identification methods. Rabilloud, T., editor. Springer-Verlag, Berlin, Germany **1999**.
- Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. Proteome research: new frontiers in functional genomics. Wilkins, M. R., Williams, K. L., Appel, R. D., and Hochstrasser, D. F., editors. Springer-Verlag, Berlin, Germany **1997**.
- Internet resources**
- A history of mass spectrometry (Scripps Center for Mass Spectrometry, Beckman Center for Chemical Sciences, La Jolla, California, USA); accessible at <http://masspec.scripps.edu/hist.html>.
- European Bioinformatics Institute (EBI; Cambridge, United Kingdom); accessible at <http://www.ebi.ac.uk/>.
- ExPASy (Expert Protein Analysis System) Molecular Biology Server (Swiss Institute of Bioinformatics, SIB; Geneva, Switzerland); available at <http://www.expasy.org/>.
- Mascot (Matrix Science Ltd., London, United Kingdom); accessible at <http://www.matrix-science.com>.
- National Center for Biotechnology Information (NCBI; National Library of Medicine, Bethesda, Maryland, USA); accessible at <http://www.ncbi.nlm.nih.gov/>.
- Non-redundant REference Protein Database (PIR-NREF; Protein Information Resource, PIR; National Biomedical Research Foundation, Washington DC, Columbia, USA); available at <http://pir.georgetown.edu/>.
- PIR-International Protein Sequence Database (PIR-PSD; Protein Information Resource, PIR; National Biomedical Research Foundation); available at <http://pir.georgetown.edu/>.
- Protein Data Bank (PDB; Research Collaboration for Structural Bioinformatics, USA); available at <http://www.rcsb.org/pdb/>.
- ProteinProspector (San Francisco Mass Spectrometry Facility, University of California, California, USA); available at <http://prospector.ucsf.edu/>.
- Protein Information Resource (PIR; National Biomedical Research Foundation, Washington DC, Columbia, USA); available at <http://pir.georgetown.edu/>.
- PROWL (ProteoMetrics, LLC, and Rockefeller University, New York, New York, USA); available at <http://prowl.rockefeller.edu/>.
- SWISS-PROT (Swiss Institute of Bioinformatics, SIB; Geneva, Switzerland); available at <http://www.expasy.org/sprot/>.
- TrEMBL at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/trembl/>.
- WORLD-2DPAGE (Swiss Institute of Bioinformatics); available at <http://www.expasy.org/ch2d/2d-index.html>.
- Cited publications**
- Abbott, A. A post-genomic challenge: learning to read patterns of protein synthesis. *Nature* **1999**; 402(6763):715-720.
- Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **1998**; 92(3): 291-294.
- Andersson, L., and Porath, J. Isolation of phosphoproteins by immobilized metal (Fe^{3+}) affinity chromatography. *Analytical Biochemistry* **1986**; 154(1): 250-254.
- Annan, R. S., Huddleston, M. J., Verma, R., Deshaies, R. J., and Carr, S. A. A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Analytical Chemical* **2001**; 73(3):393-404.
- Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R., Golaz, O., Pasquali, C., and Hochstrasser, D. F. Federated 2-DE database: a simple means of publishing 2-DE data. *Electrophoresis* **1996**; 17(3):540-546.
- Arai, R., Nakagawa, H., Tsumoto, K., Mahoney, W., Kumagai, I., Ueda, H., and Nagamune, T. Demonstration of a homogeneous noncompetitive immunoassay based on bioluminescence resonance energy transfer. *Analytical Biochemistry* **2001**; 289(1):77-81.
- Arenkov, P., Kukhtin, A., Gemmell, A., Voloshchuk, S., Chupeeva, V., and Mirzabekov, A. Protein microchips: use for immunoassay and enzymatic reactions. *Analytical Biochemistry* **2000**; 278(2):123-131.
- Bartel, P. S., and Fields, S. Analyzing protein-protein interactions using the two-hybrid system. *Methods in Enzymology* **1995**; 254: 241-263.

- Barth, H. G., Boyes, B. E., and Jackson, C. Size exclusion chromatography. *Analytical Chemistry* **1994**; 66(12):595R-620R.
- Bartle, K. D., and Myers, P. Theory of capillary electrochromatography. *Journal of Chromatography A* **2001**; 916(1-2):3-23.
- Berggren, K. N., Chernokalskaya, E., Lopez, M. F., Beechem, J. M., and Patton, W. F. Comparison of three different fluorescent visualization strategies for detecting Escherichia coli ATP synthase subunits after sodium dodecyl sulfate-polyacrylamide gel electrophoresis. *Proteomics* **2001**; 1(1):54-65.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**; 28(1):235-242.
- Bill, R. M., Revers, L., and Wilson, I. B. H. Protein glycosylation. Bill, R. M., Revers, L., and Wilson, I. B. H., editors. Kluwer Academic Publishers, Boston, New York, USA **1998**.
- Binz, P.-A., Müller, M., Walther, D., Bienvenut, W. V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M. R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmime, A., Bairoch, A., Sanchez, J. C., Appel, R. D., and Hochstrasser, D. F. A molecular scanner to automate proteomic research and to display proteome images. *Analytical Chemistry* **1999**; 71(21):4981-4988.
- Bjellquist, B., Ek, K., Reghetti, P. G., Gianazza, E., and Gorg, A. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *Journal of Biochemical and Biophysical Methods* **1982**; 6(4):317-339.
- Blackburn, G. F., Shah, H. P., Kenten, J. H., Leland, J., Kamin, R. A., Link, J., Peterman, J., Powell, M. J., Shah, A., Talley, D. B., Tyagi, S. K., Wilkins, E., Wu, T. G., and Massey, R. J. Electrochemiluminescence detection for development of immunoassays and DNA-probe assays for clinical diagnostics. *Clinical Chemistry* **1991**; 37(9):1534-1539.
- Bock, J. L. The new era of automated immunoassay. *American Journal of Clinical Pathology* **2000**; 113(5):628-646.
- Brosinger, F., Freimuth, H., Lacher, M., Ehrfeld, W., Gedig, E., Katerkamp, A., Spener, F., and Cammann, K. A label free affinity sensor with compensation of unspecific protein interactions by a highly sensitive integrated optical Mach-Zehnder interferometer on silicon. *Sensors and Actuators B* **1997**; 44:350-355.
- Büssow, K., Cahill, D., Nietfeld, W., Bancroft, D., Scherzinger, E., Lehrach, H., and Walter, G. A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Research* **1998**; 26(21):5007-5008.
- Cahill, D. J. Protein and antibody arrays and their medical applications. *Journal of Immunological Methods* **2001**; 250(1-2):81-91.
- Case, M. C., Burt, A. D., Hughes, J., Palmer, J. M., Collier, J. D., Bassendine, M. F., Yeaman, S. J., Hughes, M. A., and Major, G. N. Enhanced ultrasensitive detection of structurally diverse antigens using a single immunoprecipitation assay protocol. *Journal of Immunological Methods* **1999**; 223(1):93-106.
- Celis, J. E., and Gromov, P. 2D protein electrophoresis: can it be perfected? *Current Opinions in Biotechnology* **1999**; 10(1):16-21.
- Clauser, K. R., Baker, P. R., and Burlingame, A. L. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry* **1999**; 71(14):2871-2882.
- Coons, A. H., Creech, H. H., and Jones, R. N. Immunological properties of an antibody containing a fluorescent group. *Proceedings of the Society for Experimental Biology and Medicine* **1941**; 47:200-202.
- Crimmins, D. L. Applications of strong cation-exchange (SCX)-HPLC in synthetic peptide analysis. *Methods in Molecular Biology* **1994**; 36:53-64.
- Crowther, J. R. The ELISA guidebook. Humana Press, Totowa, New Jersey, USA **2001**.
- Cush, R., Cronin, J. M., Goddard, N. J., Maule, C. H., Molloy, J., and Steward, J. The resonant mirror: A novel optical biosensor for direct sensing of biomolecular interactions. Part I: Principle of operation and associated instrumentation. *Biosensors and Bioelectronics* **1993**; 8(7-8):347-354.
- De Wildt, R., Mundy, C. R., Gorick, B. D., and Tomlinson, I. M. Antibody arrays for high throughput screening of antibody-antigen interactions. *Nature Biotechnology* **2000**; 18(9):989-994.
- Diamond, R. A., and DeMaggio, S. In living color. Protocols in flow cytometry and cell

- sorting. Diamond, R. A., and DeMaggio, S., editors. Springer-Verlag, Berlin, Germany **2000**.
- Diamandis, E. P., and Christopoulos, T. K. The biotin-(strept)avidin system: principles and applications in biotechnology. *Clinical Chemistry* **1991**; 37(5):625-636.
- Dickson, E. F., Pollak, A., and Diamandis, E. P. Ultrasensitive bioanalytical assays using time-resolved fluorescence detection. *Pharmacology and Therapeutics* **1995**; 66(2):207-235.
- Ekins, R. P. Immunoassay design and optimization. In: Principles and practice of immunoassay. Price, C. P., and Newman, D. J., editors. 2nd edition. Stockton Press, New York, New York, USA **1997**:173-207.
- Ekins, R., Chu, F., and Micallef, J. High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays. *Journal of Bioluminescence and Chemiluminescence* **1989**; 4(1):59-78.
- Erdile, L. F., Smith, D., and Berd, D. Whole cell ELISA for detection of tumor antigen expression in tumor samples. *Journal of Immunological Methods* **2001**; 258(1-2):47-53.
- Fenyo, D., Zhang, W., Chait, B. T., and Beavis, R. C. Internet-based analytical chemistry resources. A model project. *Analytical Chemistry* **1996**; 68(23):721A-726A.
- Fernandez, J., Andrews, L., and Mische, S. M. An improved procedure for enzymatic digestion of polyvinylidene difluoride-bound proteins for internal sequence analysis. *Analytical Biochemistry* **1994**; 218(1):112-117.
- Figeys, D. Array and lab on a chip technology for protein characterization. *Current Opinion in Molecular Therapeutics* **1999**; 6(1):685-694.
- Formosa, T., and Alberts, B. M. The use of affinity chromatography to study proteins involved in bacteriophage T4 genetic recombination. *Cold Spring Harbor Symposia on Quantitative Biology* **1984**; 49:363-370.
- Ge, H. UPA, an universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein ligand interactions. *Nucleic Acid Research* **2000**; 28(2):E3.
- Gräfe, D., Elender, G., Grau, W., Dobschal, H.-J., and Berthel, G. Anordnung zum Nachweis biomolekularer Reaktionen und Wechselwirkungen. *DE19828547* **1998**.
- Guardigli, M. Immunoassay: Methods and protocols. Humana Press, Totowa, New Jersey, USA **2002**.
- Guertens, G., van Cauwenberghe, K., de Boeck, G., Maes, R., Tjaden, U. R., van der Greef, J., Highley, M., van Oosterom, A. T., and de Bruijn, E. A. Nanotechnology in bio/clinical analysis. *Journal of Chromatography. B, Biomedical Sciences and Applications* **2000**; 739(1):139-150.
- Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology* **1999**; 19(3):1720-1730.
- Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **1999**; 17(10):994-999.
- Haab, B. B., Dunham, M. J., and Brown, P. O. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology* **2000**; 1(6):RESEARCH0004.1-RESEARCH0004.13.
- Han, D. K., Eng, J., Zhou, H., and Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology* **2001**; 19(10):946-951.
- Handley, J. Software for MS protein identification. *Analytical Chemistry* **2002**; 74(5):159A-162A.
- Harlow, E., and Lane, D. Using antibodies: A laboratory manual. Harlow, E., and Lane, D., editors. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA **1999**.
- Hernandez, H., and Robinson, C. V. Dynamic protein complexes: Insights from mass spectrometry. *Journal of Biological Chemistry* **2001**; 276(50):46685-46688.
- Hoogland, C., Sanchez, J. C., Walther, D., Baujard, V., Baujard, O., Tonella, L., Hochstrasser, D. F., and Appel, R. D. Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* **1999**; 20(18):3568-3571.
- Hoving, S., Voshol, H., and van Oostrum, J. Towards high performance two-dimensional gel electrophoresis using ultrazoom gels. *Electrophoresis* **2000**; 21(13):2617-2621.
- Hunter, T. Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell* **1995**; 80(2):225-236.

- Issaq, H. J. The role of separation science in proteomics research. *Electrophoresis* **2001**; 22(17):3629-3638.
- Jayasena, S. D. Aptamers: an emerging class of molecules that rival antibodies in diagnostics. *Clinical Chemistry* **1999**; 45(9):1628-1650.
- Karas, M., and Hillenkamp, F. Laser desorption/ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* **1988**; 60(20):2299-2301.
- Keay, R. W., and McNeil, C. J. Separation-free electrochemical immunosensor for rapid determination of atrazine. *Biosensors and Bioelectronics* **1998**; 13(9):963-970.
- Kessler, C. Nonradioactive analysis of biomolecules. Kessler, C., editor. Springer-Verlag, Berlin, Germany **2000**.
- Klee, G. G. Human anti-moose antibodies. *Archives of Pathology and Laboratory Medicine* **2000**; 124(6):921-923.
- König, S., Grote, J., and Gedig, E. High-capacity dextran hydrogels for protein mass spectrometry. *International Biotechnology Laboratory* **2002**; 20(1):10.
- Kodadek, T. Protein microarrays: prospects and problems. *Chemistry and Biology* **2001**; 8(2):105-115.
- König, S., Zeller, M., Peter-Katalinic, J., Roth, J., Sorg, C., and Vogl, T. Use of non-specific cleavage products for protein sequence analysis as shown on calyculin isolated from human granulocytes. *Journal of the American Society for Mass Spectrometry* **2001**; 12(11):1180-1185.
- Kontermann, R., and Dübel, S. Antibody engineering. Kontermann, R., and Dübel, S., editors. Springer-Verlag, Berlin, Germany **2001**.
- Kricka, L. J. Ultrasensitive immunoassay techniques. *Clinical Biochemistry* **1993**; 26(5):325-331.
- Kricka, L. J. Human anti-animal antibody interferences in immunological assays. *Clinical Chemistry* **1999**; 45(7):942-956.
- Labrou, N., and Clonis, Y. D. The affinity technology in downstream processing. *Journal of Biotechnology* **1994**; 36(2):95-119.
- Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**; 227(259):680-685.
- Lecitra, E. J., and Liu, J. O. A three-hybrid system for detecting small ligand-protein interactions. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(23):12817-12821.
- Liedberg, B., Nylander, C., and Lundström, I. Surface plasmon resonance for gas detection and biosensing. *Sensors and Actuators* **1983**; 4:299-304.
- Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., 3rd. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology* **1999**; 17(7):676-682.
- Lopez, M. F. Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoresis* **2000**; 21(6):1082-1093.
- Lukosz, W., and Tiefenthaler, K. Sensitivity of integrated optical grating and prism couplers as (bio)chemical sensors. *Sensors and Actuators* **1988**; 15:273-281.
- MacBeath, G., and Schreiber, S. Printing proteins as microarrays for high-throughput function determination. *Science* **2000**; 289(5485):1760-1763.
- Merchant, M., and Weinberger, S. R. Recent advancements in surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. *Electrophoresis* **2000**; 21(6):1164-1177.
- Miles, L. E., and Hales, C. N. Labelled antibodies and immunological assay systems. *Nature* **1968**; 219(150):186-189.
- Mills, I., Cvitaš, T., Homann, K., Kallay, N., and Kuchitsu, K. Quantities, units and symbols in physical chemistry. 2nd edition. International Union of Pure and Applied Chemistry Division. Blackwell Science, Oxford, United Kingdom **1993**.
- Molloy, M. P. Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Analytical Biochemistry* **2000**; 280(1):1-10.
- Nedelkov, D., and Nelson, R. W. Exploring the limit of detection in biomolecular interaction analysis mass spectrometry (BIA/MS): detection of attomole amounts of native proteins present in complex biological mixtures. *Analytica Chimica Acta* **2000**; 423(1):1-7.
- Nelson, R. W., Jarvik, J. W., Taillon, B. E., and Tubbs, K. A. BIA/MS of epitope-tagged peptides directly from *E. coli* lysate: Multiplex detection and protein identification at low-femtomole to subfemtomole levels. *Analytical Chemistry* **1999**, 71(14):2858-2865.

- Nelson, R. W., Nedelkov, D., and Tubbs, K. A. Biomolecular interaction analysis mass spectrometry. *Analytical Chemistry* **2000**; 72(11): 404A-411A.
- Neubauer, G., and Mann, M. Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry: potentials and limitations. *Analytical Chemistry* **1999**; 71(1):235-242.
- Neuschafer, D., Budach, W., Bar, E., Pawlak, M., and Duveneck, G. Planar waveguides as efficient transducers for bioaffinity sensors. *Proceedings of SPIE – International Society for Optical Engineering* **1996**; 2836(12):221-234.
- O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry* **1975**; 250(10):4007-4021.
- Pandley, A., and Mann, M. Proteomics to study genes and genomes. *Nature* **2000**; 405(6788):837-846.
- Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**; 20(18):3551-3567.
- Peng, J., and Gygi, S. P. Proteomics: the move to mixtures. *Journal of Mass Spectrometry* **2001**; 36(10):1083-1091.
- Plebani, M., and Carraro, P. Mistakes in a stat laboratory: types and frequency. *Clinical Chemistry* **1997**; 43(8 Part 1):1348-1351.
- Posewitz, M. C., and Tempst, P. Immobilized Gallium(III) affinity chromatography of phosphopeptides. *Analytical Chemistry* **1999**; 71(14):2883-2892.
- Pretorius, V., Hopkins, B. J., and Schieke, J. D. Electro-osmosis – A new concept for high-speed liquid chromatography. *Journal of Chromatography* **1974**; 99(1):23-30.
- Quadroni, M., and James P. Proteomics and automation. *Electrophoresis* **1999**; 20(4-5): 664-677.
- Rademaker, G. J., and Thomas-Oates, J. Analysis of glycoproteins and glycopeptides using fast-atom bombardment. In: *Methods in Molecular Biology*. Chapman, J. R., editor. Humana Press, Totowa, New Jersey **1996**; 61.
- Regnier, F. E. High-performance liquid chromatography of proteins. *Methods in Enzymology* **1983**; 91:137-190.
- Rickert, J., Weiss, T., Kraas, W., Jung, G., and Göpel, W. A new affinity biosensor: Self assembled thiols as selective monolayer coatings of quartz crystal microbalances. *Biosensors and Bioelectronics* **1996**; 11(6-7):591-598.
- Rilbe, H. Rapid isoelectric focusing in density gradient columns. *Annals of the New York Academy of Sciences* **1973**; 209:80-93.
- Robertson, E. F., Dannelly, H. K., Mallov, P. J., and Reeves, H. C. Rapid isoelectric focusing in a vertical polyacrylamide minigel system. *Analytical Biochemistry* **1987**; 167(2): 290-294.
- Roda, A., Pasini, P., Guardigli, M., Baraldini, M., Musiani, M., and Mirasoli, M. Bio- and chemiluminescence in bioanalysis. *Fresenius Journal of Analytical Chemistry* **2000**; 366(6-7):752-759.
- Rosenfeld, J., Capdevielle, J., Guillemot, J. C., and Ferrara, P. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical Biochemistry* **1992**; 203(1):173-179.
- Rosman, K. J. R., and Taylor, P. D. P. (Commission on Atomic Weights and Isotopic Abundances). Isotopic compositions of the elements. *Pure and Applied Chemistry* **1998**; 70(1):217-236.
- Rowe, C. A., Scruggs, S. B., Feldstein, M. J., Golden, J. P., and Ligler, F. S. An array immunosensor for simultaneous detection of clinical analytes. *Analytical Chemistry* **1999**; 71(2):433-439.
- Salomon, B., Cohen, J. L., Masurier, C., and Klatzmann, D. Three populations of mouse lymph node dendritic cells with different origins and dynamics. *Journal of Immunology* **1998**; 160(2):708-717.
- Schägger, H., and von Jagow, G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Analytical Biochemistry* **1987**; 166(2):368-379.
- Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-colour fluorescent probe hybridization. *Genome Research* **1996**; 6(7):639-645.
- Schena, M. Genome analysis with gene expression microarrays. *Bioessays* **1996**; 18(5):424-431.
- Schneider, B. H., Dickinson, E. L., Vach, M. D., Hoijer, J. V., and Howard L. V. Highly sensitive optical immunoassays in human serum. *Biosensors and Bioelectronics* **2000**; 15(1-2): 13-22.

- Schroeder, H. R., Vogelhut, P. O., Carrico, R. J., Boguslaski, R. C., and Buckler, R. T. Competitive protein binding assay for biotin monitored by chemiluminescence. *Analytical Chemistry* **1976**; 48(13):1933-1937.
- Schubert, W. Ultrasensitive Hochdurchsatzanalyse von Proteomen auf Einzelzellniveau. *Transkript Laborwelt* **2000**; 4:42-43.
- Schuck, P., and Minton, A. P. Analysis of mass transport limited binding kinetics in evanescent wave biosensors. *Analytical Biochemistry* **1996**; 240(2):262-272.
- Sehr, P., Zumbach, K., Pawlita, M. A generic capture ELISA for recombinant proteins fused to glutathione S-transferase: validation for HPV serology. *Journal of Immunological Methods* **2001**; 253(1-2):153-162.
- Seitz, W. R. Immunoassay labels based on chemiluminescence and bioluminescence. *Clinical Biochemistry* **1984**; 17(2): 120-125.
- Shapiro, H. M. Practical flow cytometry. 3rd edition. Wiley-Liss, Inc., New York, New York, USA **1995**.
- Shepherd, P., and Dean, C. Monoclonal antibodies: a practical approach. Shepherd, P., and Dean, C., editors. Oxford University Press, Oxford, United Kingdom **2000**.
- Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Analytical Chemistry* **1996**; 68(5): 850-858.
- Shevchenko, A., Jennsen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proceedings of the National Academy of Sciences of the USA* **1996**; 93(25):14440-14445.
- Soldin, S. J. Receptor assays in the clinical laboratory. *Clinical Biochemistry* **1996**; 29(5):439-444.
- Stefan, R. I., van Staden, J. F., and Aboul-Enein, H. Y. Immunosensors in clinical analysis. *Fresenius Journal of Analytical Chemistry* **2000**; 366(6-7):659-668.
- Strohman, R. Epigenesis: The missing beat in biotechnology? *Biotechnology (New York)* **1994**; 12(2):156-164.
- Stults, J. T., Henzel, W. J., Wong, S. C., and Watanabe, C. Identification of electroblotted proteins by peptide mass searching of a sequence database. In: Mass spectrometry in the biological sciences. Burlingame, A. L., Carr, S. A., editors. Humana Press, Totowa, New Jersey, USA **1996**.
- Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yoshida, T. Protein and polymer analysis up to m/z 100,000 by laser ionization and time-of-flight mass spectrometry. *Proceeds of the 2nd Japan-China Joint Symposium on Mass Spectrometry*. Osaka, Japan **1987**:185-187.
- Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. and Yoshida, T. Protein and polymer analysis up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **1988**; 2(8):151-153.
- Tatusova, T. A., Karsch-Mizrachi, I., and Ostell, J. A. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* **1999**; 15(7-8):536-543.
- Tom-Moy, M., Baer, R. L., Spira-Solomon, D., and Doherty, T. P. Atrazin measurement using surface transverse wave devices. *Analytical Chemistry* **1995**; 67(7):1510-1514.
- Towbin, H., Staehelin, T., and Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the USA* **1979**; 76(9): 4350-4354.
- Turkova, J. Bioaffinity chromatography. In: Analytical and preparative separation methods of biomacromolecules. Aboul-Enein, H. Y., editors. Marcel Dekker, New York, New York, USA **1999**:99-165.
- Unlu, M., Morgan, M. E., and Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **1997**; 18(11): 2071-2077.
- Von Schulthess, G. K., Cohen, R. J., and Benedek, G. B. Laser light scattering spectroscopic immunoassay in the agglutination-inhibition mode for human chorionic gonadotropin (hCG) and human luteinizing hormone (hLH). *Immunochemistry* **1976**; 13(12):963-966.
- Vuong, G. L., Weiss, S. M., Kammer, W., Priemer, M., Vingron, M., Nordheim, A., and Cahill, M. A. Improved sensitivity proteomics by postharvest alkylation and radioactive labelling of proteins. *Electrophoresis* **2000**; 21(13):2594-2605.

- Warsinke, A., Benkert, A., and Scheller, F. W. Electrochemical immunoassays. *Fresenius Journal of Analytical Chemistry* **2000**; 366 (6-7):622-634.
- Washburn, M. P., Wolters, D., and Yates, J. R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **2001**; 19(3):242-247.
- Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **1995**; 16(7):1090-1094.
- Weber, K., and Osborn, M. The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry* **1969**; 244(16):4406-4412.
- Wheeler, D. L., Chappay, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2000**; 28(1): 10-14.
- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2001**; 29(1):11-16.
- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Research* **2002**; 30(1):13-16.
- Wickware, P., and Smaglik, P. Proteomics technology: character references. *Nature* **2001**; 413(6822):869-875.
- Wild, D. The immunoassay handbook. Wild, D., editor. Nature Publishing Group, New York, New York, USA **2001**.
- Williams, C., and Addona, T. A. The integration of SPR biosensors with mass spectrometry: Possible applications for proteome analysis. *Trends in Biotechnology* **2000**; 18(2):45-48.
- Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., and Williams, K. L. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews* **1996**; 13:19-50.
- Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology* **1999**; 112:531-552.
- Wolf, E., and Brem, G. "High-dose hook effect" as a pitfall in quantifying transgene expression in metallothionein-human growth hormone (MT-hGH) transgenic mice. *Clinical Chemistry* **1991**; 37(5):763-765.
- Wu, J. T., Huang, P., Li, M. X., and Lubman, D. M. Protein digest analysis by pressurized capillary electrochromatography using an ion trap storage/reflectron time-of-flight mass detector. *Analytical Chemistry* **1997**; 69(15):2908-2913.
- Yalow, R. S., and Berson, S. A. Assay of plasma insulin in human subjects by immunological methods. *Nature* **1959**; 184(4699):1648-1649.
- Yates, J. R., 3rd. Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometry* **1998**; 33(1):1-19.
- Young, M., Tang, N., Hempel, J. C., Oshiro, C. M., Taylor, E. W., Kuntz, I. D., Gibson, B. W., and Dollinger, G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(11):5802-5806.
- Yu, B. S., Choi, Y. K., and Chung, H. H. Development of immunoassay methods by use of liposomes. *Biotechnology and Applied Biochemistry* **1987**; 9(3):209-216.
- Zhang, X., Herring, C. J., Romano, P. R., Szczepanowska, J., Brzeska, H., Hinnebusch, A. G., and Qin, J. Identification of phosphorylation sites in proteins separated by polyacrylamide gel electrophoresis. *Analytical Chemistry* **1998**; 70(10): 2050-2059.

6

Methods for mRNA and protein expression analysis *in situ* and *in vivo*

6.1

Introduction

In the past few years, the focus of gene expression analysis has been on massively parallel approaches using either solid or bead-based arrays, PCR-based techniques or by means of serial analysis of gene expression and related technologies, that allow simultaneous measurement of the expression of very many genes. Such brute force approaches provide vast amounts of data on gene expression, but are a relatively blunt instrument. In general, little information on the topology of gene expression is provided, and the dynamic nature of gene expression in the living organism is often poorly represented.

However, a number of powerful methods exist for the analysis of gene expression in particular circumstances, such as the living cell or even the organism as a whole. While these methods usually allow the expression of only a single gene, or a small number of genes, to be analysed, they have the great advantage of providing a true reflection of the dynamism of living processes. Other methods are not applicable *in vivo*, but provide information of the location of expressed genes within tissues and cells.

Thus, this chapter begins with a description of *in situ* hybridisation and immunochemical techniques, which are still the work-horses for analysing gene expression in tissues. There follows a description of methods to visualise RNA in living cells by means of fluorescent probes. Indeed, all the remaining sections of this chapter may be applied to the living cell, to living tissue, or even to the whole organism. Methods such as positron emission tomography, add a further dimension, allowing metabolic processes to be followed in the living organism in real time.

A particularly exciting development is the use of light-based methods for analysis of gene expression. This chapter provides brief descriptions of two such methods, optical *in vivo* imaging and single photon emission computed tomography. Such light-based methods hold much promise for dynamic and non-invasive analysis of gene expression in living organisms.

6.2

***In situ* hybridisation, immunocytochemistry and immunohistochemistry**

6.2.1

Introduction

Localisation of specific mRNA and protein species using histochemical techniques may reveal new information about the expression of a gene within a cell or tissue and may be regarded as complementary to that generated by quantitative methods performed on processed tissue or cell samples. Thus, when used in an integrated fashion, both the amount and distribution of a specific mRNA and/or protein can be determined.

A variety of histochemical techniques is available for the study of gene expression. These include *in situ* hybridisation histochemistry (ISH) for the detection of mRNA, immunocytochemistry (ICC) or immunohistochemistry (IHC) for the detection of protein, enzyme histochemistry for measuring the catalytic activity of enzymes within tissues and ligand-binding auto-radiography for the study of receptor-ligand interactions. Although ICC/IHC and ISH reveal little about functionality, they are the techniques that are most widely used to localise the expression products of individual genes in cells and tissues, being preferred because of their superior durability and their wider target and tissue applicability. ICC/IHC and ISH methodologies adopt similar approaches in so far as they each utilise direct or indirect visualisation of labelled probes with specific affinity for the target molecule of interest (see figure 6.1).

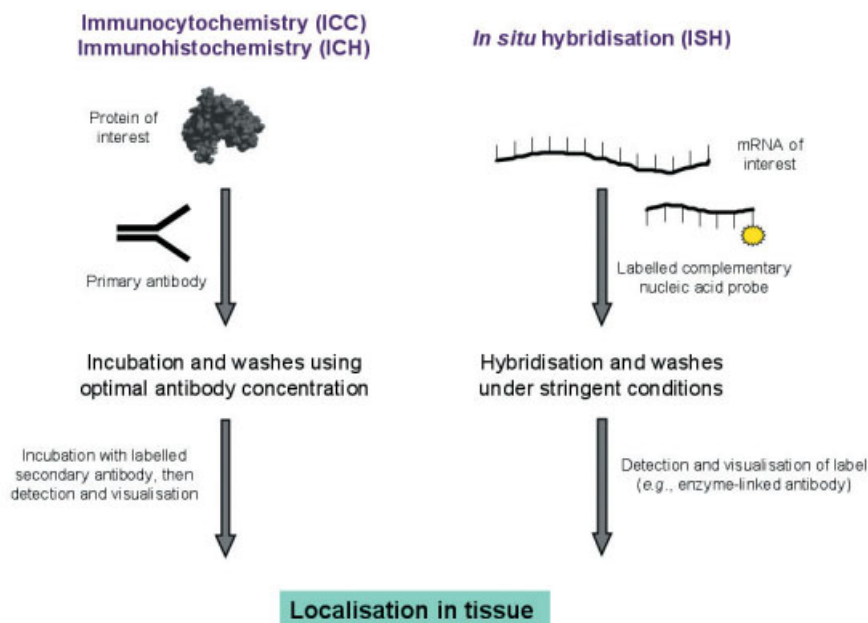


Figure 6.1 Immunocytochemistry/immunohistochemistry (ICC/IHC) and *in situ* hybridisation (ISH). Outline of typical methodologies.

6.2.2

***In situ* hybridisation**

ISH utilises nucleic acid probes with base sequence complementary to that of the target mRNA. An important feature of ISH probes is that they can be specifically designed to bind any nucleic acid sequence, thus conferring wide applicability to the technique. A critical component of all ISH protocols is the conditions under which hybridisation is carried out, as this determines the specificity of the target/probe hybrid and hence the specificity of the detection. Various sample preparation and signal amplification techniques are employed in order to optimise the sensitivity of the protocol. The key procedural components of ISH are depicted diagrammatically in figures 6.2.A and 6.2.B. A typical ISH experiment result is shown in figure 6.3, using rat inducible nitric oxide synthase (iNOS) as an example target.

6.2.2.1 Probe types

A number of different types of nucleic acid probes have been developed for ISH. Single-stranded antisense oligodeoxynucleotides, commonly between 25 and 40 bases long, are readily synthesised and labelled and make good probes for ISH; particularly when targeting abundant mRNA species (Hougaard *et al.*, 1997). Equimolar probe 'cocktails' of several different non-overlapping antisense sequences may increase the sensitivity of the detection. Single-stranded ribonucleotide probes, 'riboprobes', are becoming increasingly popular for ISH because of their relatively high sensitivity and ease of preparation and have been applied extensively in the study of many different mRNA targets in a variety of tissues (*e.g.*, Meltzer *et al.*, 1998, Baumgart *et al.*, 1997). Riboprobes are usually designed to be larger than 300 bases in length as longer fragments have reduced penetration of tissue sections. Riboprobes can be prepared from double-stranded sequences inserted into a plasmid, or by the generation of a template fragment by PCR using a thermal cycler. Both are characterised by the possession of flanking sequences for RNA polymerase promoters such as T3 and T7. Antisense and sense riboprobes with a high labelling index can be selectively generated by *in vitro* transcription using the specific polymerase enzyme to produce target and control probes with identical GC to AT ratios. In recent years, riboprobes have tended to replace double-stranded DNA probes in ISH protocols. Although cDNA probes can be efficiently labelled, sense and antisense strands are generated as the duplex, giving rise to the potential for non-specific hybridisation. Peptide nucleic acid (PNA) probes, based upon a peptide-like backbone, are a relatively recent development for ISH and are not yet in widespread use. The physicochemical properties of the PNA molecule confer high affinity and specificity, whilst reducing non-specific interaction. Such features, in combination with a sensitive label detection protocol, have enabled relatively low abundance targets to be detected successfully in routinely fixed tissues (Yokomori *et al.*, 2000).

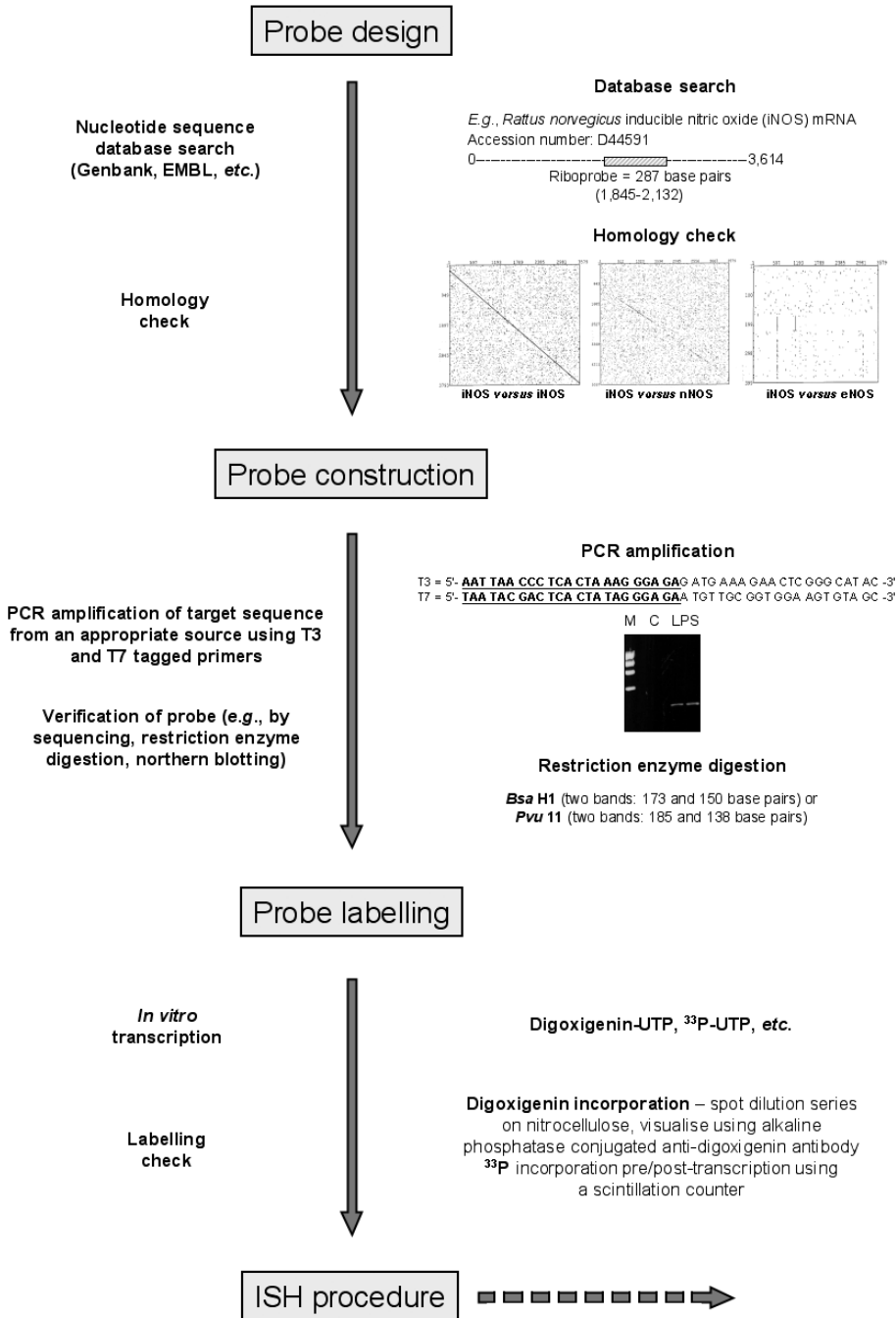


Figure 6.2.A

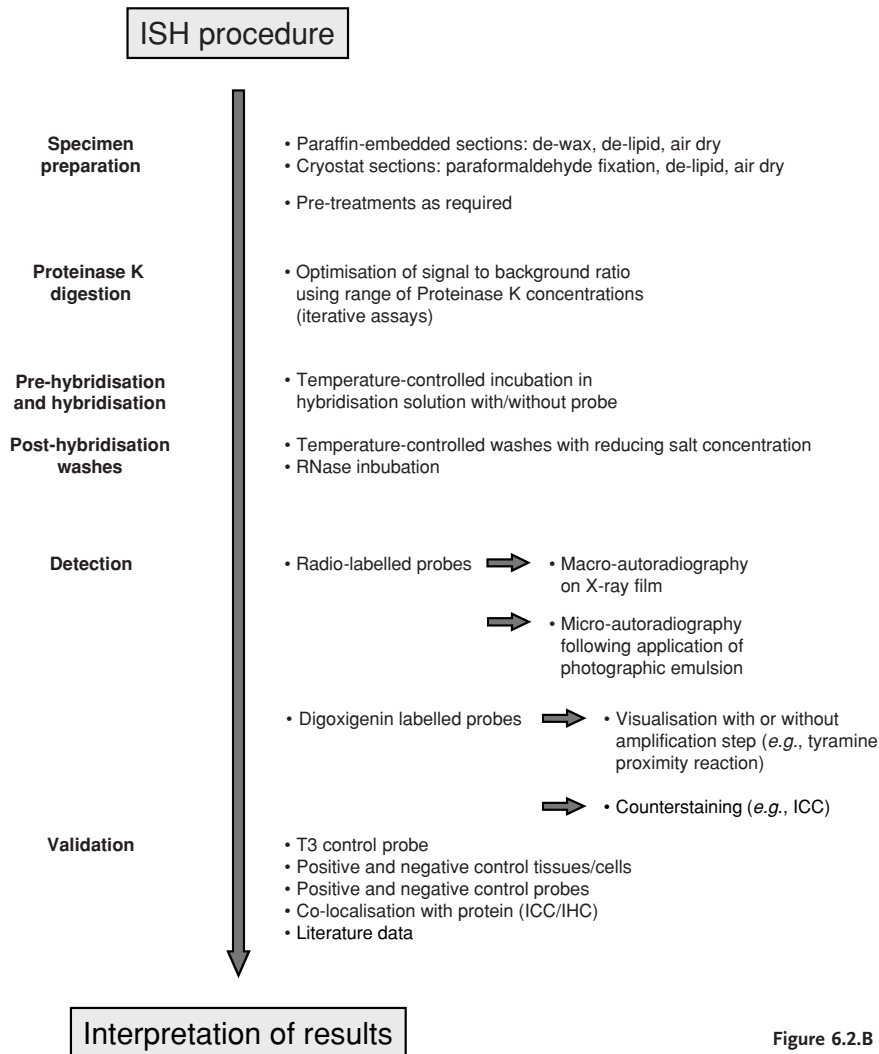


Figure 6.2.B

Figure 6.2 Principle of *in situ* hybridisation (ISH) analysis. A) Flow diagram showing generalised steps in construction of a riboprobe for ISH. B) Flow diagram showing generalised basic steps in an ISH experiment using a radiolabelled or digoxigenin-labelled riboprobe. Abbreviations

used in this figure: C, control sample; eNOS, endothelial nitric oxide synthase; iNOS, inducible nitric oxide synthase; LPS, lipopolysaccharide-treated sample; M, marker; nNOS, neuronal nitric oxide synthase. Sequences of T3 and T7 promoters are underlined.

6.2.2.2 Probe labels and detection methods

Numerous probe-labelling methods have been developed for ISH, but those remaining in common use are essentially restricted to the incorporation of radiolabelled or chemically modified nucleotides into a single or double-stranded template. ^{35}S and ^{33}P radioisotopes provide a good balance between half-life and

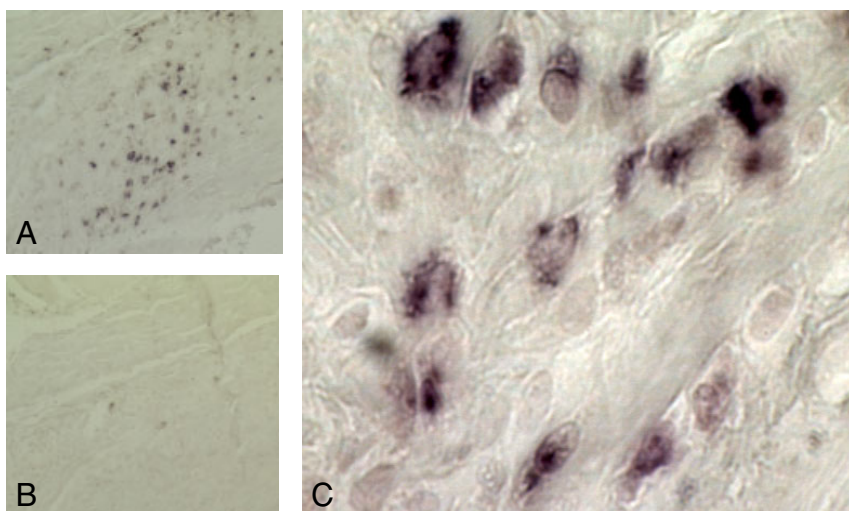


Figure 6.3 Tissue localisation of inducible nitric oxide synthase (iNOS) mRNA in arthritic rat joint using a digoxigenin-labelled antisense riboprobe *in situ* hybridisation (ISH). Visualisation was performed using an anti-digoxigenin alkaline phosphatase-conjugated antibody and nitro blue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP). **A)** T7 antisense probe. **B)** T3 sense probe (negative control). **C)** T7 antisense probe (high magnification). The negative control was performed using the corresponding sense riboprobe under the same conditions.

decay energy such as to give reasonable exposure time (generally a few weeks) with acceptable localisation of signal over target (grain scatter). Although reduced exposure times can be used with the higher energy ^{32} phosphorus radioisotope, grain scatter is usually unacceptably high, with consequential loss of localisation.

Non-isotopic probes offer a more convenient and rapid detection procedure than radio-labelling. Common non-isotopic labels include biotin, digoxigenin and various fluorochromes (Dirks 1996; Komminoth 1992). These are normally linked to a visualisation step using specific antibodies, streptavidin (biotin conjugation), or by direct fluorescence (*e.g.*, fluorescein, FITC, and rhodamine, TRITC). A simplified detection procedure can be adopted by direct labelling of probe (usually an oligonucleotide) with an enzyme such as alkaline phosphatase.

6.2.2.3 Probe design strategy

The design of a probe for ISH is to a large extent dictated by knowledge of the base sequence of the target mRNA (or DNA if a chromosome is the target). In order to produce a probe that will hybridise specifically with the target of interest, it is essential to perform a search of available sequence databases for possible homology with other mRNA species. Clearly, however, lack of specificity may still be a potential problem for homologous sequences not represented in the database. Additional factors can influence the effectiveness of the probe in ISH, for example, problems

of secondary structure and non-optimal GC/AT ratio. Such properties can be factored into search routines to identify sequences with properties most likely to produce good probes for ISH. A number of software programs are available to facilitate this process. Another common strategy used in probe design is to ensure the probe sequence targets transcribed regions that cross intron/exon boundaries, in order to avoid possible hybridisation to genomic DNA.

The sensitivity of the probe used is quite clearly a function of the number of detectable labels that can be incorporated into the probe without compromising either its specificity or ability to access and hybridise with the target. Achievable labelling density will depend upon the label, the method used for its incorporation and the type and length of probe. Radiolabelled nucleotides can be incorporated at high density and hence give probes with high sensitivity, although these, especially ³²phosphorus, have limited stability compared to chemical groups such as digoxigenin. Uniform incorporation of label throughout the probe is generally a feature of riboprobes and cDNA probes generated by enzymatic reaction, whereas oligonucleotides are usually labelled at the 3' or 5'-terminus either chemically during synthesis or enzymatically using a terminal transferase enzyme.

6.2.2.4 Sample acquisition and fixation

Tissues for ISH should be preserved as quickly as possible following removal in order to ensure that the quality of the mRNA is maintained. Combined perfusion-fixation and euthanasia procedures can be expected to produce high quality mRNA retention. Alternatively, rapid removal of tissues and submersion into fixative or snap freezing in liquid nitrogen should produce acceptable levels of preservation, whilst allowing flexibility of fixation method. Common cross-linking fixatives include buffered formalin, paraformaldehyde and glutaraldehyde. Protein precipitating fixatives, including ethanol, methanol and acetone in various proportions are often used for cell preparations. However, these tend to give relatively poor preservation of morphology when used for tissues. In order to produce uniform fixation throughout the specimen, tissues preserved by submersion in fixative should be trimmed in order to increase surface area relative to volume, thereby facilitating access of fixative to the entire tissue bulk. This is important in order to avoid artefacts resulting from variable hybridisation across the specimen. Frozen tissues sectioned on a cryostat can be fixed immediately following mounting onto the microscope slide. Similarly, cell preparations can be fixed *in situ*, either on microscope slides or in suitable cell culture plates (*e.g.*, chamber slides). It is important to achieve consistency between all fixation protocols so that the subsequent ISH procedure can be standardised.

6.2.2.5 Preparation and hybridisation of specimens

Various pre-treatments may be carried out on tissue or cell preparations prior to hybridisation with the probe. These are aimed primarily at optimising probe penetration and access to the target, minimising non-specific hybridisation and staining

problems, whilst retaining satisfactory tissue morphology (Oliver *et al.*, 1997). A key step in the preparation of fixed tissue sections for hybridisation is their partial digestion using a proteolytic enzyme such as proteinase K under optimal and reproducible conditions. This procedure often requires prior experimentation using a range of enzyme concentrations and depends on the degree of fixation and type of fixative used for tissue preservation. It is important to standardise the fixation procedure for a set of tissues in order to minimise differences in protein digestion requirements. Pre-treatment with proteases may also improve the ISH result by inactivation of nucleases. A number of additional pre-treatments such as acetylation and acidification are commonly incorporated into protocols, primarily in an attempt to reduce non-specific binding of probe and non specific staining due, for example, to the presence of endogenous forms of the enzyme (*e.g.*, peroxidases) used in the detection step.

Precise incubation conditions are required in order to ensure stability of the duplex formed between probe and the complementary target sequence whilst maintaining the required stringency for specific hybridisation. The main factors governing this include temperature, the salt and formamide concentrations of the hybridisation reaction and the length and base sequence of the probe. A guideline to suitable conditions can be obtained by calculating the melting temperature T_m , using the following equation:

$$T_m = 16.6 \cdot \log[\text{Na}^+] + 0.41 \cdot (\text{content of G/C base pairs in percent}) + 81.5 - (675 / \text{probe length}) - 0.65 \cdot (\text{formamide content in percent})$$

As a rule of thumb, optimal specific hybridisation will generally occur at around 10°C below the melting temperature.

Post-hybridisation washes are designed to remove all non-hybridised probe and probe hybridised to non-complementary sequences and generally entail temperature controlled washes in buffer at reducing salt concentration. An additional step often included with riboprobes is post-hybridisation incubation with ribonuclease in order to degrade and remove non-hybridised probe.

6.2.2.6 Optimisation and control of *in situ* hybridisation analysis

Preliminary experimentation is usually required to determine optimal conditions for producing maximum specific signal with acceptable tissue/cell morphology and background staining. A useful approach may be to perform an initial ISH using a probe for total mRNA (Pringle *et al.*, 1989) or other well-characterised target of known distribution and abundance.

Inclusion of controls is essential for successful analysis of *in situ* hybridisation. These should include the use of control tissues or cells (in which the mRNA target is known to be present/absent), and control probes (such as the corresponding sense sequence generated by transcription of labelled riboprobe using T3 polymerase). Additional supporting information can also be generated by other methods for analysis of gene expression such as RT-PCR and immunohistochemistry or im-

munocytochemistry. Useful negative control tissue sections can be generated by inclusion of a ribonuclease incubation stage prior to hybridisation. However, extreme care must be taken to ensure the enzyme does not contaminate other sections or subsequent steps in the ISH protocol.

6.2.2.7 Increasing sensitivity of detection

A major limitation of routine ISH analysis is its lack of sensitivity for the detection of low abundance mRNA species. Various approaches can be adopted to increase the sensitivity of the technique. These can be categorised into signal and target amplification.

Amplification of the signal produced by the labelled hybridised probe is an inherent component of most ISH methods and largely depends upon the label used. Most chromogenic staining methods employ enzymatic generation of coloured products. The time of this reaction can be extended to some degree without increasing background intensity. Similarly, exposure time can be increased with radio-isotopic labelling. More recently, proximity-based enzymatic amplification steps have been successfully employed for signal enhancement, using for example tyramine (Qian *et al.*, 2001; Zaidi *et al.*, 2000). Limited success has been achieved with selective and specific amplification of the target sequence based upon PCR methods. A key problem with this approach is the need to produce conditions within the tissue environment that enables amplification to occur whilst simultaneously preventing diffusion of the amplified sequence away from its origin. Although a number of PCR-amplified ISH methods have been developed, protocols tend not to be readily adaptable to general applications without considerable effort. These methods, together with their advantages and drawbacks, are well described by Komminoth *et al.* (1997).

6.2.3

Immunocytochemistry and immunohistochemistry

ICC/IHC methodologies are well established. A wide range of primary antibodies is available commercially and through specialist laboratories. A number of complete assay kits are also available in which all reagents necessary for the immunohistochemical detection of specific protein targets are included, usually with an optimised protocol. See figure 6.4 for a schematic illustration of ICC/IHC methodologies.

The main factors that determine the suitability of a primary antibody for a particular application are specificity of the antibody for the protein target, species cross-reactivity and sensitivity of the detection system. Also of great importance is the type of fixation used for the cell or tissue as this frequently determines the ability of the antibody to recognise and bind to the target epitope. A variety of sample preparation procedures are commonly utilised in order to maximise the accessibility of the antibody to its target. Similarly, numerous amplification techniques are available for enhancing visualisation of the label used in the detection system. Primary antibodies may be monoclonal or polyclonal in origin. Each

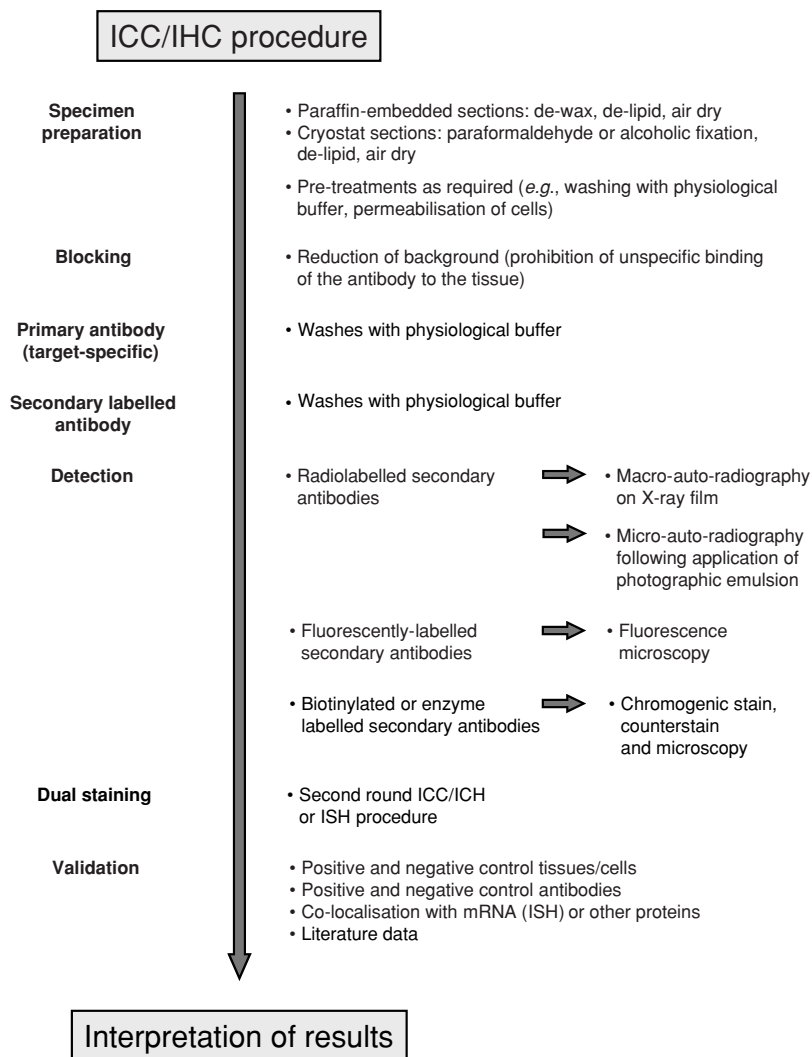


Figure 6.4 Principle of immunocytochemistry/immunohistochemistry (ICC/IHC) analysis. Flow diagram showing generalised basic steps in an ICC/IHC experiment using radiolabelled, fluorescently-labelled or biotinylated secondary antibodies.

of these has well documented advantages and disadvantages for use in ICC/IHC. Although the number of commercially available antibodies is limited, it is possible to produce antibodies directed against synthetic peptides, enabling a greater degree of flexibility for studying new targets of interest. The recent development and introduction of phage display antibody libraries has greatly expanded the potential availability of specific antibodies for immunocytochemistry and other applications

(Krebs *et al.*, 2001). Furthermore, these antibodies can be developed in a few weeks as compared to the period of several months required for conventional techniques.

6.2.4

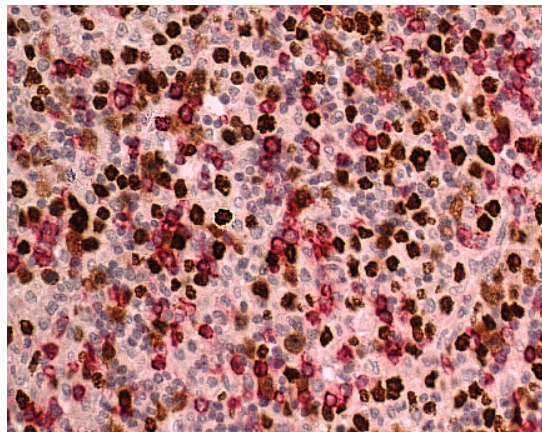
Advanced techniques

Simultaneous detection of multiple targets of interest, either by ISH, ICC/IHC or a combination of both is a powerful means of revealing additional information about the function of and interactions between cell populations. Many examples of dual and triple staining procedures can be found in the literature and these provide useful guidance on the approaches required (van der Loos, 1999; Stern *et al.*, 1998). Whereas chromogenic staining methods tend to be limiting with respect to the number of different colour combinations available, the expanding repertoire of fluorescent dyes can be utilised in relatively straightforward procedures to detect several targets at once, although specialised fluorescence microscopes and image capture systems are usually required. Examples of dual staining using an automated histochemistry system are shown in figure 6.5.

ISH and ICC/IHC are relatively labour-intensive techniques, typically requiring numerous manual manipulations. As such, protocols tend to be of protracted length, variable reproducibility and limited sample throughput. Although there is a clear need for automation of ISH and ICC/IHC, this has been accomplished with varying degrees of success. Whereas a number of automated systems have been developed commercially for ICC, fewer are available for the more complex operations required for ISH. A recent combined ISH and ICC analyser has been produced by Ventana Medical Systems, Inc. (Tucson, Arizona, USA), which automates the entire ISH and ICC/IHC process, including the removal of paraffin wax from the sections and the application of coverslips to the slides (Grogan *et al.*, 1995).

ISH and ICC/IHC have been adapted for use at the ultrastructural level, in combination with electron microscopy. These, however, remain essentially specialised

Figure 6.5 Combination of *in situ* hybridisation (ISH) and immunohistochemistry (IHC). Combined ISH (Epstein Barr Virus, brown stain) and IHC (CD43, red stain) on a tissue section, performed using a Ventana Discovery automated ICC/ISH analyser (Ventana Medical Systems, Inc., Tucson, Arizona, USA). Figure kindly provided by Dr. Fabien Gaire (Ventana Medical Systems SA, Illkirch CEDEX, France).



techniques with limited applicability. Ultrastructural ICC can be utilised to investigate aspects of the function of macromolecules such as protein trafficking within specific subcellular elements such as the cytoskeleton. In contrast, both ISH and ICC/IHC can be applied to large tissue specimens including whole embryo mounts. This has enabled substantial advances to be made in the identification of tissues and processes involved in development (Stern, 1998). The procedures applied are essentially the same as those for tissue sections, although incubation and wash times are usually considerably increased to allow for diffusion and removal of the various reagents.

Although ICC/IHC and ISH remain primarily qualitative techniques, quantitative data may be generated by careful application of densitometry, image analysis and other measuring systems. In order to ensure the integrity of data produced from quantitative analysis, rigorous validation and control procedures must be adopted (Gerfen, 1989; Kumar, 2000).

6.2.5

Application in the study of drug efficacy and toxicity

Histochemical techniques can be usefully applied to a number of different aspects of pharmaceutical research and development. Evaluation of the potential for a candidate drug to modulate a disease process and conversely, to cause undesirable side effects, is a key objective in the early stages of its progression. Thus, for example, characterisation of the induction and pharmacological inhibition of a key inflammatory mediator such as inducible nitric oxide synthase (iNOS) in an animal model of arthritis can provide important information on *in vivo* potency. Furthermore, localisation of iNOS expression within the disease tissue can help to identify key cellular elements involved (see figure 6.3). Localisation of stress-inducible protein expression can be a valuable tool for studying the *in vivo* effects of a wide variety of drugs and other agents with the potential to cause an intracellular stress response. Thus, for example, induction of heat shock protein 72 (*hsp72*) in the brain of rats following administration of the N-methyl-D-aspartate (NMDA) receptor antagonist MK-801 (dizocilpine maleate, see figure 6.6) occurs in those regions of the cingulate cortex and retrosplenial cortex known to be susceptible to neuronal vacuolation and necrosis (Kind, 1996). Furthermore, expression of *hsp72* and vacuolation has been co-localised within individual neurones (Sharp, 1991). Information gained from such investigations can be valuable as a method for targeting an area of tissue or population of cells for further analysis, either by co-localisation with other mRNA or protein species of interest, or for isolation and subsequent analysis (for example, using microdissection techniques). Importantly, altered gene expression associated with intracellular stress or perturbation of key cellular pathways may be detectable prior to and at lower doses than those required to produce morphological or functional signs of toxicity. This provides an opportunity to investigate the underlying mechanism of toxicity and identification of critical pathways leading to formation of the lesion.

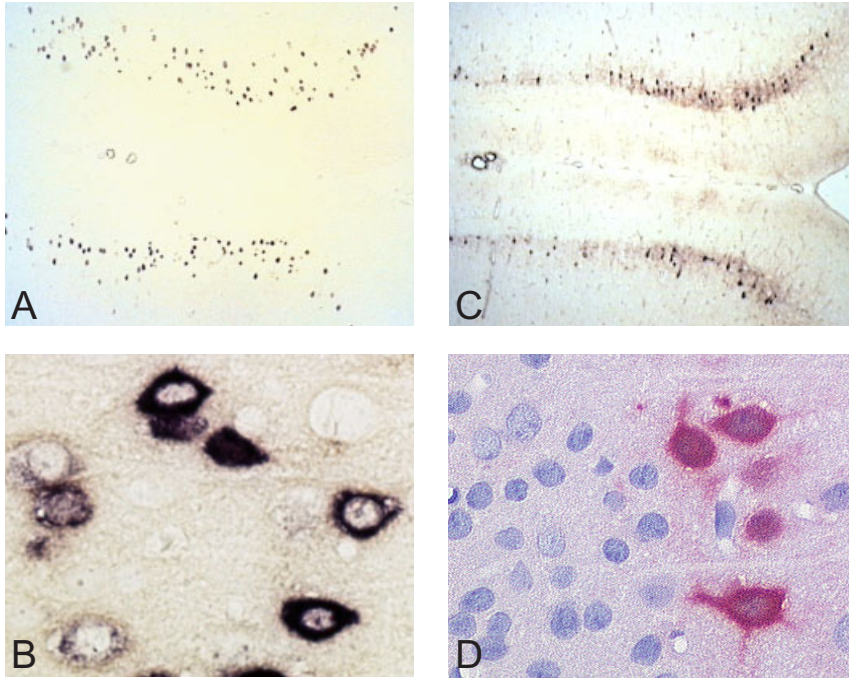


Figure 6.6 Combined *in situ* hybridisation (ISH) and immunohistochemistry (IHC) to analyse mRNA and protein expression of the cell-stress gene, heat shock protein 72 (*hsp72*). Induction of the *hsp72* mRNA (ISH) and protein (IHC) in rat posterior cingulate cortex following exposure to the N-methyl-D-aspartate (NMDA) receptor antagonist MK-801. **A) and B)** (low and high magnification, respectively): ISH was performed using a digoxigenin-labelled antisense riboprobe to rat *hsp72*, visualised using an anti-

digoxigenin alkaline phosphatase-conjugated antibody and nitro blue tetrazolium/5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP). **C) and D)** (low and high magnification, respectively): IHC was performed using a monoclonal antibody specific for rat *hsp72* (Stressgen Biotechnologies Corporation, Victoria, Canada) visualised using an alkaline phosphatase linked secondary antibody and Fast Red.

6.3

Visualising RNA in living cells

6.3.1

Introduction

Current optimised and standardised protocols of *in situ* hybridisation (ISH) have lead to a better understanding of gene expression patterns in several tissues. These protocols include the fixation and permeabilisation of the biological material, which may significantly alter the localisation of RNA within the cells. For this reason, protocols that are used for *in situ* hybridisation are of limited use in the study of living cells.

To examine whether the results of *in situ* hybridisation in fixed cells or tissue reflect the RNA distribution patterns *in vivo*, methods for visualising RNA molecules in living cells have been developed. These provide an opportunity to observe RNA processing and transport in the *in vivo* situation. These methods include standard procedures such as direct and indirect labelling techniques, and also newer applications such as fluorescence *in vivo* hybridisation (Dirks *et al.*, 2001) and the two-plasmid expression system have been applied (Bertrand *et al.*, 1998).

6.3.2

Fluorescence *in vivo* hybridisation

Fluorescence *in vivo* hybridisation appears to be a promising method because it is the technique in which the natural conditions in the cells are least disturbed. The concept of this method does not fundamentally differ from *in situ* hybridisation, but problems remain concerning the introduction of labelled material into the cell and the specific detection of fluorescence signals from hybridised probes only (Dirks *et al.*, 2001).

6.3.2.1 Probe selection

The first problem is the choice of an optimal probe for RNA hybridisation in living cells. Since unmodified phosphodiester oligodeoxynucleotides are known to be prone to degradation by endonucleases and exonucleases, they are not suitable for this method. Modified RNA molecules and RNA analogues such as those used in antisense protocols are better candidates. One of the more suitable derivatives are phosphorothioate oligodeoxynucleotides, in which a non-bridging oxygen of the phosphodiester nucleic acid backbone is replaced by sulphur. Other useful RNA analogues are 2'-O-methyl-RNAs in which the 2'-oxygen is methylated, and peptide nucleic acids in which the nucleic acid phosphodiester backbone is replaced by a 2-aminoethylglycine linkage. These RNA analogues are shown in figure 6.7.

Studies of the stability of phosphorothioate oligodeoxynucleotides in cells have shown that 70 percent of the phosphorothioate oligodeoxynucleotides are not degraded after one day, in contrary to unmodified oligodeoxynucleotides which have a half-life of only 15 to 20 minutes (Fisher *et al.*, 1993). In addition, it has been shown that the susceptibility to RNase H cleavage of RNAs hybridised with phosphorothioate oligodeoxynucleotides is low (Agrawal *et al.*, 1997), which also renders them more stable than unmodified RNAs.

Advantages of 2'-O-methyl-RNAs are their high affinity to complementary sequences and their resistance to nuclease activity. Such RNA derivatives are very suitable for visualising small nuclear RNAs (snRNA) (Carmo-Fonseca *et al.*, 1991). This was confirmed in studies performed by Dirks *et al.* (2001) and Moleenaar *et al.* (2001), where 2'-O-methyl-RNAs complementary to the small nuclear RNAs U1 and U3, 28S ribosomal RNA (rRNA) and poly(A) messenger RNAs (mRNAs) were microinjected into the cytoplasm of U2OS cells. The distribution

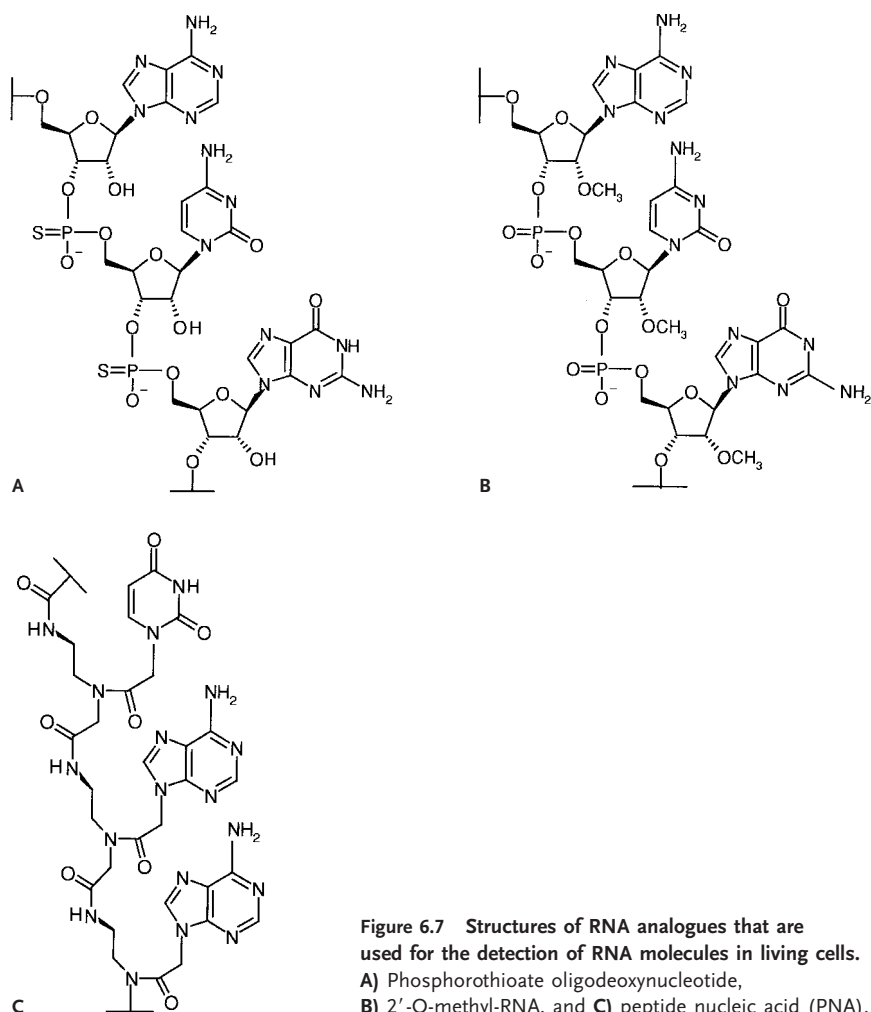


Figure 6.7 Structures of RNA analogues that are used for the detection of RNA molecules in living cells. **A)** Phosphorothioate oligodeoxynucleotide, **B)** 2'-O-methyl-RNA, and **C)** peptide nucleic acid (PNA).

patterns of these probes in the cells after injection and binding to their complementary cellular RNAs are shown in figure 6.8. In U2OS cells U1 snRNA (figures 6.8.A and 6.8.B) and poly(A) mRNA (figures 6.8.E and 6.8.F) show nearly the same speckle pattern, whereas U3 snRNA (figures 6.8.C and 6.8.D) and 28S rRNA (figures 6.8.G and 6.8.H) are localised mainly to nucleoli.

Peptide nucleic acids would also seem to be good candidates for detecting RNA in living cells because of their non-ionic state and the resistance to enzymatic cleavage. The marked structural changes seen in peptide nucleic acids (figure 6.7) allow these molecules to form triple-helices by forming the naturally occurring Watson-Crick base pairs, so that stable hybridisation is guaranteed. Despite these advantages, investigations using peptide nucleic acids for RNA detection in living cells have revealed poor sequence specificity (Dirks *et al*, 2001).

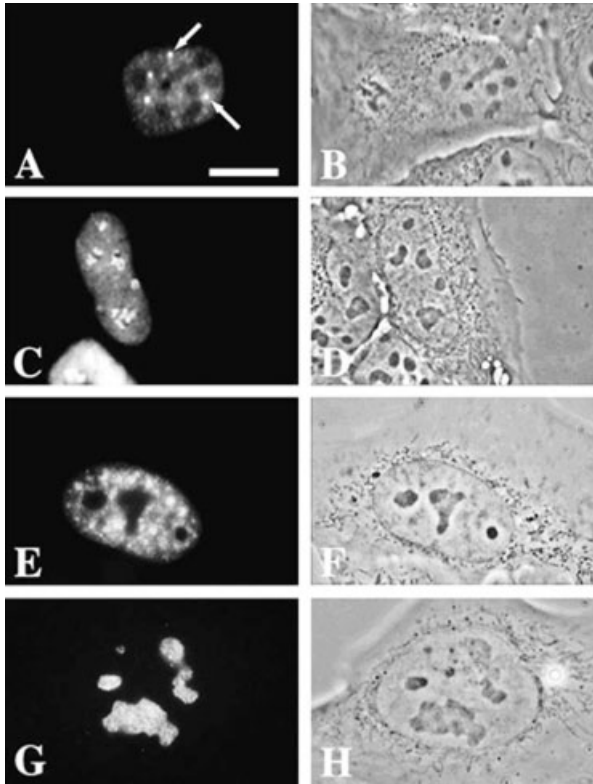


Figure 6.8 Cellular localisation of fluorescence-labelled 2'-O-methyl RNAs in living U2OS cells after microinjection in the cytoplasm. **A)** U1 small nuclear RNA-specific probes rapidly accumulated first in Cajal bodies (arrows) and a short while later also in speckles. **C)** U3 small nuclear RNA-specific probes were shown to localise mainly to nucleoli and to display a diffuse pattern throughout the nucleus. **E)** Poly(U) probes showed a speckled pattern within the

nucleoplasm excluding nucleoli. **G)** 28S ribosomal RNA-specific probes localised mainly to nucleoli. No hybridisation signals were observed in the cytoplasm of cells. **B), D), F), and H)** show the corresponding phase-contrast images. The bar indicates ten micrometers. The figure was kindly provided by Dr. Roeland W. Dirks (Leiden University Medical Centre, The Netherlands).

Once the problem of choosing the appropriate probe has been solved, the next question to be answered is the localisation of the hybridisation to the target RNA molecule. It is very important that the hybridisation does not take place at splice sites, sequences that promote polyadenylation or other sequences that contain regulatory elements, because this would disturb RNA processing. In addition, the target sequence must be free accessible to the probe. Although computer analysis of RNA structures may help to solve these problems (Scherr *et al.*, 2000), the creation of optimal probes is still a major source of errors.

6.3.2.2 Cell transfection

Transfection is the introduction of exogenous material, in this case RNA and RNA analogues, into a cell. Such introduction of exogenous material is a prerequisite for studying RNA processing and transport in living cells. Different methods of passing probes through membranes have been developed. The most common method is microinjection using micropipettes by thrusting through the cell membrane and directly injecting the exogenous material into the cytoplasm. Even injection into the nucleus is possible by thrusting through both the cell and the nuclear membrane (Alberts *et al.*, 1995).

Depending on the injected material, there are several means by which the molecules migrate within the cell. While linear oligodeoxynucleotides accumulate rapidly in the nucleus *via* free diffusion, plasmids require an energy-dependent mechanism to achieve nuclear localisation (Dowty *et al.*, 1995). In addition, the mobility of probes depends strictly on their size. After microinjection, a DNA molecule of 100 base pairs is fully mobile whereas one of two kilobases is quite immobile (Lukacs *et al.*, 2000). On the other hand, DNA molecules of any size are often immobilised in the nucleus due to sequence-dependent binding to nuclear proteins (Leonetti *et al.*, 1991; Politz *et al.*, 1998).

Another efficient method to introduce probes into a cell is so called 'glass bead loading', in which a mixture of the probe solution and glass beads is loaded onto the cells. By making small lesions in the plasma membrane, the glass beads provide a simple and sufficient means for probe material to enter the cell (McNeil & Werder, 1987). In contrast to microinjection, the disadvantage is that there is no possibility to determine how much of the probe has been taken up by the cell.

Other methods for transfecting cells exploit both pinocytosis and receptor-mediated endocytosis. These are naturally occurring processes for the uptake of macromolecules and other particles through parts of the cell membrane that enclose the exogenous material and form intracellular vesicles after separation from the plasma membrane. The uptake of liquids or diluted substances in small vesicles is called pinocytosis, whereas the uptake of large vesicles (*i.e.* microorganisms or cell debris) is called phagocytosis. Receptor-mediated endocytosis is the forming of a coated vesicle from a coated pit, which is a highly specialised part of the membrane that carries different receptors for ligands (*i.e.* insulin, lipoproteins, *etc.*) that have to be taken up by the cell (Alberts *et al.*, 1995). However, transfecting cells in this way is not very efficient, because large amounts of probes are needed and the vesicles have to be destroyed inside the cell by hypotonic treatment to set the probes free.

Lipofusion is another simple transfection method. In this procedure, liposomes encapsulating the probe material fuse with the cell membrane and thus deliver the probes to the cytoplasm (Thierry & Dritschilo, 1992). The disadvantage of this method is the change in the membrane constitution with extensive cell loading and the fact that the efficiency of fusion of the liposome with the membrane depends on the cell type, which means that each cell-specific application requires its own optimisation.

Methods for transfecting large cell populations work by inducing transient permeability of the entire cell membrane. This may be achieved by chemical means using a low concentration of detergent or by the use of a strong electric shock (Alberts *et al.*, 1995). The latter application is called electroporation. If the shock is well dosed, most of the cells are able to survive by simply repairing their membranes (Alberts *et al.*, 1995). Due to different susceptibilities of membranes, these methods are not applicable to every type of cell.

6.3.2.3 Detection of hybridised RNA in living cells

A great problem to be solved in fluorescence *in vivo* hybridisation is the specific detection of a fluorescence signal from hybridised probe material only, since autofluorescence from cell compartments such as mitochondria or lysosomes and signals from non-hybridised probes also occur (figures 6.9.A and 6.9.B). While autofluorescence can be minimised by several reagents (Andersson *et al.*, 1998), the exclusion of signals derived from non-hybridised probes is more problematic.

6.3.2.3.1 Molecular beacons

Molecular beacons are single-stranded nucleic acids with a stem-loop motif. The stem is derived by self-hybridisation of the bases at the ends of the nucleic acid sequence. The resulting loop contains a sequence of 15 to 20 base pairs that are complementary to the target molecule. The ends of the chain consist of a fluorochrome at the one end and a quencher at the other. Formation of the stem-loop motif leads to suppression of the fluorochrome by the quencher. Opening of the loop on target hybridisation leads to excitation and generation of a fluorescent signal (figures 6.9.C and 6.9.D).

Although this method seems promising, little experience has yet been gathered on the use of molecular beacons for the visualisation of RNA processing and transport in living cells (Matsuo, 1998; Sokol *et al.*, 1998). However, an obvious disadvantage is the spontaneous opening of the stem-loop motif after entering the nucleus, due to association with nuclear proteins (Dirks *et al.*, 2001), which may result in non-specific fluorescence.

6.3.2.3.2 Fluorescence resonance energy transfer (FRET)

Another elegant detection method for tracking RNA molecules in the whole cell is fluorescence resonance energy transfer (FRET) (Dirks *et al.*, 2001). In this method, two probe molecules with different fluorochromes must hybridise within a space of not greater than ten nanometers and further on the wavelengths of both fluorochromes spectra have to overlap each other to allow an energy transfer. In compliance with these requirements, the emitted light of the first fluorophore (donor) can excite the second fluorophore (acceptor) to fluorescence emission (figures 6.9.E and 6.9.F).

A further development of this application is fluorescence lifetime imaging (FLIM). While fluorescence resonance energy transfer identifies the increase of the signal intensity of the acceptor fluorochrome, fluorescence lifetime imaging

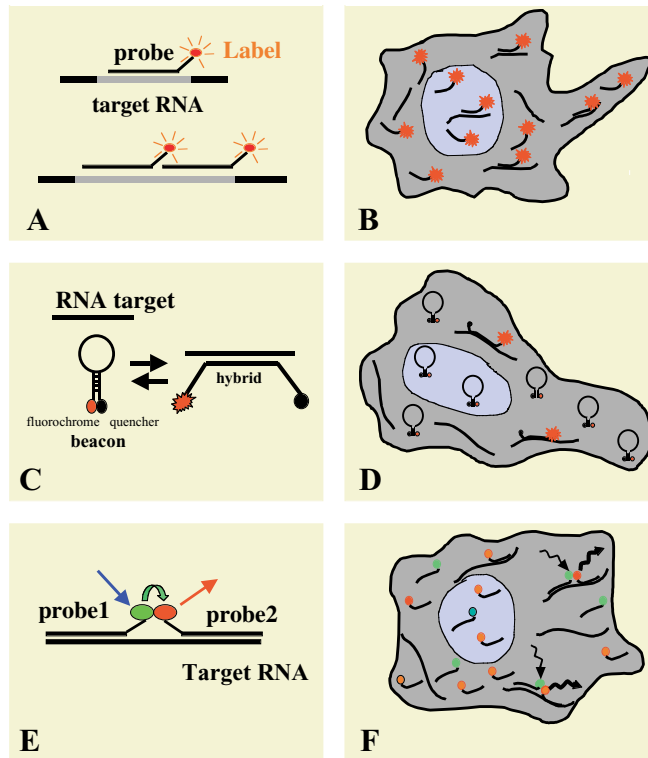


Figure 6.9 Schematic representation of fluorescence *in vivo* hybridisation approaches for living cell RNA detection. **A)** In the classic approach, one or more fluorescence-labelled oligonucleotides are used that hybridise to their complementary target RNAs. **B)** In living cells the problem is to discriminate between signals derived from hybridised and non-hybridised probes as both give the same fluorescent signal. **C)** and **D)** Using molecular beacons, fluorescence is observed only when the molecular beacon probes hybridise to their complementary target RNAs. **E)** With the more powerful

fluorescence resonance energy transfer approach, two hybridisation events must occur that bring a donor (green) and an acceptor molecule (red) in close proximity in order for energy transfer between the donor and acceptor to occur. **F)** Thus, fluorescent emission light from the acceptor molecule is visible only when the two different oligonucleotide probes hybridise close on the same messenger RNA target molecule. The figure was also kindly provided by Dr. Roeland W. Dirks and is reprinted with permission from *Histochemistry and Cell Biology*.

registers the decrease of the fluorescence signal of the donor fluorochrome. The combination of fluorescence resonance energy transfer and fluorescence lifetime imaging should allow more easy identification of background signals (Bastiaens & Squire, 1999; Lakowicz *et al.*, 1992).

Thus, with both fluorescence resonance energy transfer and fluorescence lifetime imaging the main task is to prevent unspecific signals, because the probability that non-hybridised fluorophores meet in such a way as to allow emission of light from the acceptor fluorophore is vanishingly small. The major problem in practise

is therefore to find suitable sections in the target molecule for the hybridisation of two different probe molecules.

6.3.3

Alternative methods for visualising RNA *in vivo*

In addition to fluorescence *in vivo* hybridisation, several other procedures have been developed in the last years for visualising RNA molecules in living cells: direct and indirect labelling of RNA, two plasmid expression, and microinjection of labelled RNA.

6.3.3.1 Direct and indirect labelling

Direct labelling of RNA molecules using fluorochrome-labelled nucleotides is a simple method for visualising RNA, because such nucleotides are efficiently incorporated in RNA transcripts by RNA polymerases.

An indirect labelling technique has been developed by Wansink *et al.* (1993) and Jackson *et al.* (1993). In this technique, the labelling of RNA transcripts with 5-bromo uridine 5'-triphosphate is followed by immunocytochemical staining with antibodies against 5-bromo uridine 5'-triphosphate.

Though these methods were very successful in the investigation of RNA distribution patterns in the nucleoplasm it cannot be excluded that the incorporated fluorophores alters the processing and transport of RNA within the cell nucleus.

6.3.3.2 Two plasmid expression

Bertrand *et al.* (1998) developed a method for the indirect real-time detection of RNA movement using chimeric plasmid constructs in living yeast cells (figure 6.10). In this technique, one plasmid is constructed containing DNA encoding a fusion protein which consists of (i) the capsid protein of the bacteriophage MS2, (ii) a green fluorescent reporter protein, and (iii) a nuclear localisation signal sequence. A second plasmid is then constructed containing DNA which encodes (i) an RNA containing six MS2-binding sites and, (ii) the target RNA. The RNA encoded by this former sequence binds avidly to MS2 capsid protein. Translation of the RNA encoded by this second plasmid is prevented by the introduction of a translation termination codon immediately in front of the sequence of the MS2-binding site. Because of the nuclear localisation signal, all unbound reporter molecules are restricted to the cell nucleus. Only those complexes which form between the MS2 capsid protein/reporter molecule construct and the RNA target/MS2-binding site transcript can leave the nucleus and be detected within the cytoplasm. Other results making use of this method are very promising (Beach *et al.*, 1999; Beach & Bloom, 2001). Additionally, the MS2 coat protein-based system has been successfully applied to mammalian cells in order to image mRNA transport in living neurons (Rook *et al.*, 2000).

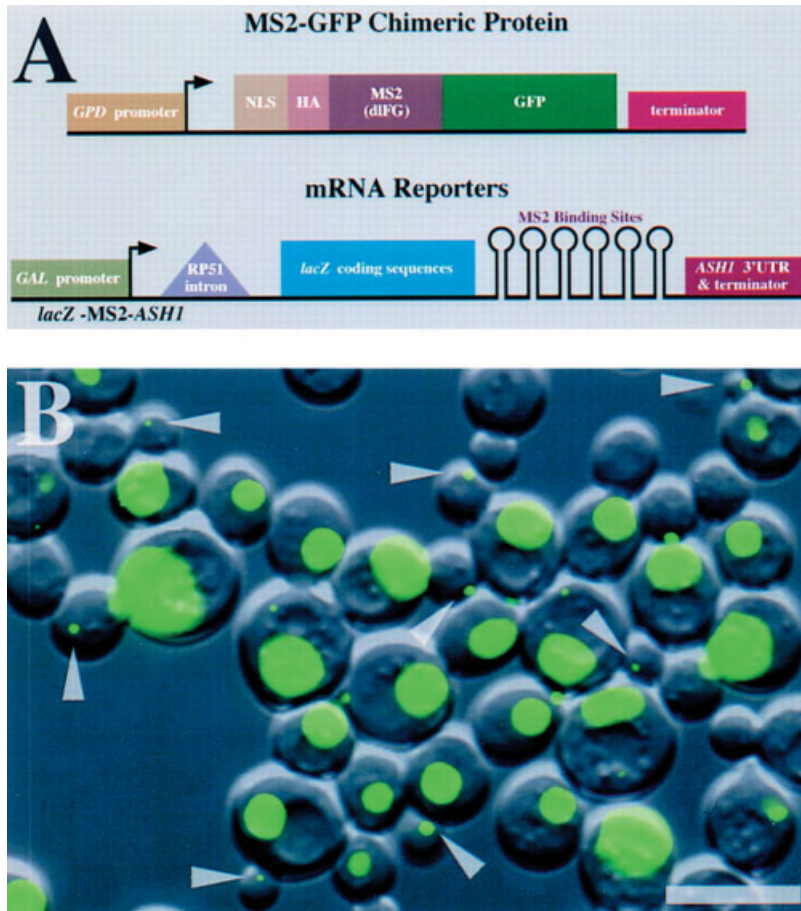


Figure 6.10 Visualisation of reporter messenger RNAs in live cells. **A)** Schematic presentation of the reporter constructs used in the two plasmid expression approach. The system is comprised of two plasmids, a reporter messenger RNA construct and a plasmid encoding a fusion protein consisting of the green fluorescent protein and the bacterial phage capsid protein MS2. The reporter messenger RNA construct contains six additional binding sites for the capsid protein, to which the MS2 part of the fusion protein binds. A complementary sequence of the 3'-untranslated region of the *ASH1* messenger RNA was used as a hybridisation probe. In addition, a nuclear localisation

signal was introduced at the amino terminus of the fusion protein so that only the green fluorescent/MS2 fusion protein, which is bound to the MS2 binding sites of the messenger RNA reporter, moves from the nucleus into the cytoplasm. **B)** Live cells expressing the green fluorescent/MS2 fusion protein and the MS2-*ASH1* reporter messenger RNA. Arrows indicate some of the labelled particles. The bar indicates five micrometers. The figure was kindly provided by Dr. Robert H. Singer (Albert Einstein College of Medicine, New York, United States of America) and is reprinted with permission from *Molecular Cell*.

6.3.2.4.3 Microinjection of labelled RNA

It is possible to study the transport processes and localisation of specific RNA molecules by microinjection of fluorescent labelled RNA molecules derived from *in vitro* transcription of cDNA (Wang *et al.*, 1991). Labelling is achieved either by incorporation of fluorescent nucleotides or by chemical modification at the 3'-end. Though RNA trafficking can be easily observed with this application, it is possible that exogenous RNA probes may follow transport routes which differ from those of endogenous RNA.

6.4

MRI – Magnetic resonance imaging

6.4.1

Introduction

Magnetic resonance imaging (MRI) is one of the most versatile techniques for *in vivo* imaging, both clinically and experimentally. Frequently, imaging can be done at high spatial resolution with great clarity of anatomical detail. MRI can be used to determine the size and location of tumours, and to determine metastatic burden. The fundamental principle underlying MRI is that unpaired nuclear spins (such as hydrogen atoms in water and organic compounds) align themselves when placed into a magnetic field. A temporary radiofrequency pulse is then given to change the alignment of the spins, and their return to baseline is recorded as a change in electromagnetic flux. The timing parameters of pulse excitation and recording can be altered, resulting in images with different types of magnetic contrast. These are known as T1 and T2 weighting, among others.

MRI also has considerable potential for imaging at the molecular level. A number of molecular probes based on paramagnetic metal cations (such as chelated gadolinium or dysprosium), or superparamagnetic nanoparticles have been developed as reporters to image specific molecular targets at the protein or RNA level. Although the different probes have variable sensitivities in detection, most recent generations of reporter probes are subject to amplification, *i.e.* their magnetic properties enhance after target binding (Bodanov *et al.*, 2001; Josephson *et al.*, 2001; Louie *et al.*, 2000). In this chapter, approaches to MRI of mRNA, endogenous gene expression by targeted imaging and imaging of specific exogenous 'magnetic resonance reporter genes' will be described. This chapter will also describe how MRI can be used to process samples at high throughput. For more in-depth backgrounds on the basics of MRI or molecular imaging, the reader is referred to additional articles (Stark & Bradley, 1999; Weissleder & Mahmood, 2001). Magnetic resonance spectroscopy is covered in a following chapter.

When is magnetic resonance useful in providing an image of gene expression? MRI is particularly useful and advantageous if one desires to image gene expression at high resolutions in deep tissues and *in vivo*. Originally developed for use in humans, small-scale MRI equipment has also been developed that allows

high-resolution imaging of animals such as mice. This is important, because as the genomic era provides us with better animal models of cancer and more specific therapeutics, better methods are needed to monitor disease in animals. Compared to other imaging modalities, MRI has several advantages including excellent three-dimensional spatial resolution – potentially down to the tens of the micrometer range (Smith *et al.*, 1994; Johnson *et al.*, 1993) – and the opportunity to determine a number of physiological parameters at the same time, which allows functional studies and measurement of the performance of perfusion.

6.4.2

Imaging mRNA with magnetic switches

The rapid detection and read-out of specific nucleic acid sequences has become a fundamental necessity in medicine and biotechnology. Several biological applications require high speed sample throughput, simplified sample preparation and measurements in turbid or opaque tissues. Many of the newer technologies to achieve this are covered in other sections of this chapter. It has recently been observed that superparamagnetic nanoparticles conjugated to oligonucleotides exert sensitive and reversible effects on the spin-spin relaxation of adjacent water protons upon hybridisation in a fluid phase (Josephson *et al.*, 2001). This has led to the development of magnetic nanosensors which might also work with other kinds of detectors such as sensitive oscillating magnetic field readers or magnetic force microscopy. Such nanosensors may also serve as prototypical magnetic resonance contrast agents.

Magnetic nanosensor Magnetic nanosensor“ consist of monodisperse, fluid-phase nanoparticles containing an icosahedral core of superparamagnetic, crystalline $\text{Fe}_2\text{O}_3/\text{Fe}_3\text{O}_4$ particles of three nanometers in size in a cage of dextran cross-linked by epichlorohydrin. This particle (CLIO- NH_2) is functionalised by the addition of amine groups to which thiolated oligonucleotides are coupled (figure 6.11). The nanosensors are potent enhancers of the spin-spin and spin-lattice relaxation processes, reducing T2 relaxation times by as much as 20 milliseconds at femtomolar concentrations. Similar changes in relaxivity are also observed in turbid and tissue-like samples. The magnetic effects are fully reversible. Oligonucleotides can be recovered from the sensors. Recently, it has been shown that mRNA expression in cell lysates can be detected using MRI (Josephson *et al.*, 2001).

Overall, magnetic nanosensors have unique features compared to other particulate reporter systems: (i) No primary amplification step is required. (ii) Since light is not used, substances interfering with light do not affect assay values and experiments can be carried out in turbid and heterogeneous samples. (iii) The assay is homogeneous, employing neither a wash to remove unbound analyte nor the detection of analyte at a vessel wall. (iv) There is no immobilisation of the oligonucleotide onto a glass microscope slide, resulting in faster hybridisation kinetics. (v) There is great flexibility of target selection and the technique may be adapted to sense other interactions, such as protein/DNA and protein/protein interactions, with minimal sample preparation. (vi) The technology is relatively inexpensive. Re-

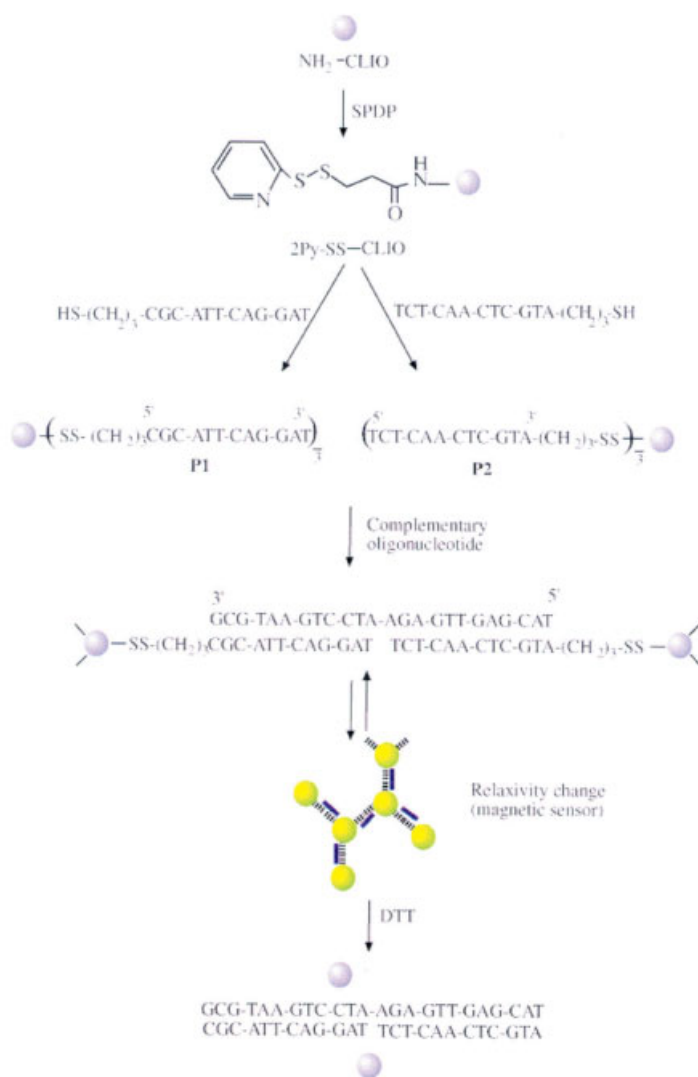


Figure 6.11 Magnetic nanosensors for magnetic resonance imaging (MRI) of oligonucleotide hybridisation. Thioalkane-substituted oligonucleotides were treated with N-succinimidyl 3-(2-pyridyldithio)propionate-activated nanoparticles to form the nanosensors P1 and P2. P1 and P2 hybridise with complementary oligonucleotides resulting in oligomerisation,

which changes the relaxivity allowing detection by MRI. Hybridisation of particles is temperature and 1,4-dithio-DL-threitol-dependent (DTT), so that isolated oligonucleotides or mRNAs can also be magnetically isolated. Reprinted with permission from *Nature Biotechnology* (Louie *et al.*, 2000).

duced sample sizes required by modern magnetic bench top readers will enable the attomolar detection of oligonucleotides at high-throughput while higher sensitivity detection systems such as magnetic force microscopy will allow detection of single strands of hybridised oligonucleotides. Finally, since nanosensors are generally nontoxic to mammalian cells, non-degradable oligonucleotide analogues (*e.g.*, peptide nucleic acids) may be coupled to nanoparticles and used to image sequences of nucleic acids *in vivo*.

6.4.3

Imaging proteins with targeted probes

Various MRI agents have been developed in order to target specific proteins in the body. In a typical scenario, an affinity ligand is attached to either a superparamagnetic iron oxide nanoparticle (Weissleder *et al.*, 2000) or to a synthetic carrier molecule bearing a large number of paramagnetic lanthanides (*e.g.*, gadolinium diethylenetriamine penta-acetic acid (Gd-DTPA)-labelled poly-L-lysine or Gd-DTPA-labelled liposomes). Tissue and cell specificity has been achieved using proteins, peptides or small molecules as targeting molecules (see Enochs *et al.* (1996) for in depth review). For example, the human transferrin receptor (see below), the hepatocyte-localised asialoglycoprotein receptor and the secretin receptor have been targeted with superparamagnetic nanoparticles. Using paramagnetic liposomes the endothelial integrin $\alpha V\beta 3$ has also been targeted (Sipkins *et al.*, 1998).

6.4.4

High-throughput imaging

The ability to image intact animals *in vivo* and biological samples at high-throughput has important applications in biotechnology. Recently, clinical MRI systems have been adapted to function as high-resolution magnetic plate readers. In one study, the use of an magnetic resonance imager permitted 1,920 samples to be analysed in about 50 minutes acquiring more than 38,000 individual data points in multi-echo trains (figure 6.12) (Högemann *et al.*, 2002). Using this approach it appears feasible to image upwards of 20,000 experimental samples per day, a number that is currently limited only by the availability of better magnetic coils and stronger gradients so as to achieve higher spatial resolutions. In the same study, MRI provided very accurate measurements, and was used to generate receptor binding and internalisation data, which was validated by radioactive assays. The technique thus allows the screening of libraries of peptide/nanoparticle conjugates directed against target cells and allows the identification of conjugates that may be subsequently used as reporter agents *in vivo*. This technology will accelerate the development of target and cell-specific magnetic resonance contrast agents.

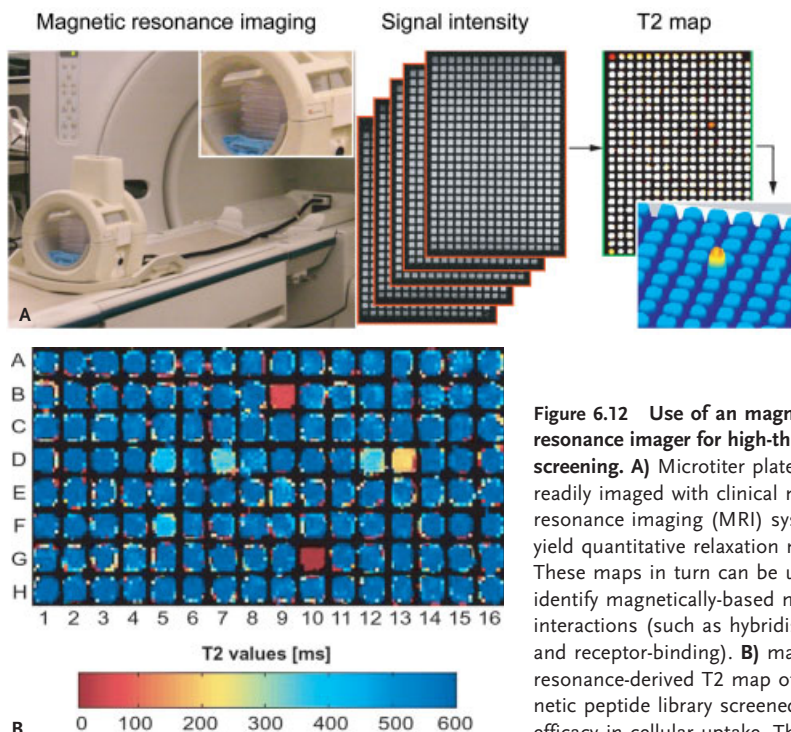


Figure 6.12 Use of an magnetic resonance imager for high-throughput screening. **A)** Microtiter plates can be readily imaged with clinical magnetic resonance imaging (MRI) systems to yield quantitative relaxation maps. These maps in turn can be used to identify magnetically-based molecular interactions (such as hybridisation and receptor-binding). **B)** magnetic resonance-derived T2 map of a magnetic peptide library screened for its efficacy in cellular uptake. The lead compound in **B** was the oligopeptide GRKKRRQRRRGYK(FITC)C-NH₂)₁₁-SC-CLIO (with FITC as the fluorescent marker fluorescein isothiocyanate, see text for explanation of the abbreviation CLIO). Reprinted with permission from *Bioconjugate Chemistry* (Högemann *et al.*, 2002).

6.4.5

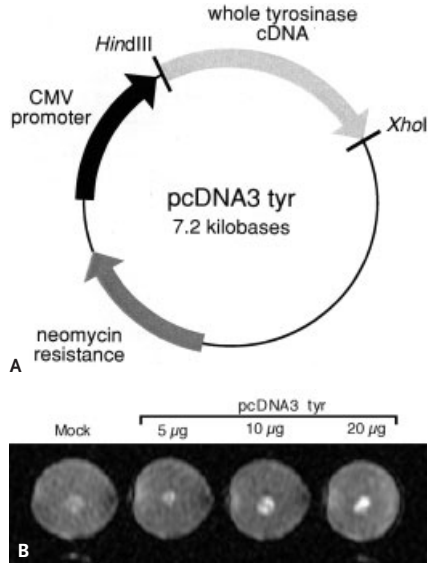
Magnetic resonance-sensitive marker genes

Different marker genes and/or substrates, which can be used for monitoring of specific transgene expression have been explored for magnetic resonance imaging purposes ('imaging marker genes'; for in depths review, see Weissleder & Mahmood, 2001).

6.4.5.1 Tyrosinase

Melanins have high metal-binding capacities, which are responsible for the high signal intensity of some melanomas on T1-weighted images (Enochs *et al.*, 1989 and 1997; Isiklar *et al.*, 1995; Okazaki *et al.*, 1985). Enzymes involved in melanogenesis such as tyrosinase can potentially be exploited to detect gene expression *in vivo* (Weissleder *et al.*, 1997). Tyrosinase catalyses the hydroxylation of tyrosine to dioxyphenylalanine (DOPA) and the subsequent oxidation of DOPA to DOPA quinone (Sanchez-Ferrer *et al.*, 1995), which finally yields melanins by spontaneous cyclisation and polymerisation. Cells transfected with a vector bearing a

Figure 6.13 Tyrosinase as a magnetic resonance marker gene. **A)** Eukaryotic expression vector pcDNA3tyr for human tyrosinase, an imaging marker gene. The vector contained the entire 1.8 kilobases cDNA encoding the human tyrosinase gene as well as a neomycin marker gene and a cytomegalovirus (CMV) promoter. **B)** T1-weighted magnetic resonance image of human cells transfected with pcDNA3 tyr (right) or mock-transfected (left) three days prior to magnetic resonance imaging. Cells were grown in iron-containing cell culture medium. Note the increase of cellular signal intensity after gene transfer due to over-expression of tyrosinase producing melanin, which scavenges iron. Reprinted with permission from *Radiology* (Enochs *et al.*, 1997).



cDNA insert coding for human tyrosinase yielded an signal which was eight times greater than that detected in a non transfected control cell line (figure 6.13) (Kwon *et al.*, 1988; Weissleder *et al.*, 1997). Recently, a number of different tyrosinase variants have been generated that are characterised by higher enzyme activity and low endogenous toxicity (Simonova *et al.*, 2000). When compared to the wild-type tyrosinase transfectants, truncated mutants resulted in higher mRNA levels in transfected healthy human epithelial cells and in tumour cells. This increase in mRNA expression paralleled the higher enzyme activity of the truncated mutants. Overall, this approach promises to serve as an efficient means of imaging marker genes in prodrug activation systems (figure 6.14) (Simonova *et al.*, 2000).

6.4.5.2 Activatable enzyme-sensitive magnetic resonance probes

Activatable or 'smart' magnetic resonance contrast agents undergo enzyme-mediated conformational changes, which result in a significant alteration of their imaging properties (*i.e.* shorter T1 relaxation). To date, various approaches are under investigation to develop smart magnetic resonance probes capable of detecting small amounts of specific enzyme activity *in vivo*.

A recently developed magnetic resonance signal amplification strategy is based on enzyme-mediated polymerisation of paramagnetic substrates into oligomers with higher relaxivity (Bodanov *et al.*, 2002). The substrates consist of chelated gadolinium covalently bound to phenols, which serve as electron donors during enzymatic hydrogen peroxide reduction by peroxidase (Bodanov *et al.*, 2002). The converted monomers undergo rapid condensation into paramagnetic oligomers leading to a threefold increase in atomic relaxivity (figure 6.14) (Bodanov *et al.*, 2002). *In*

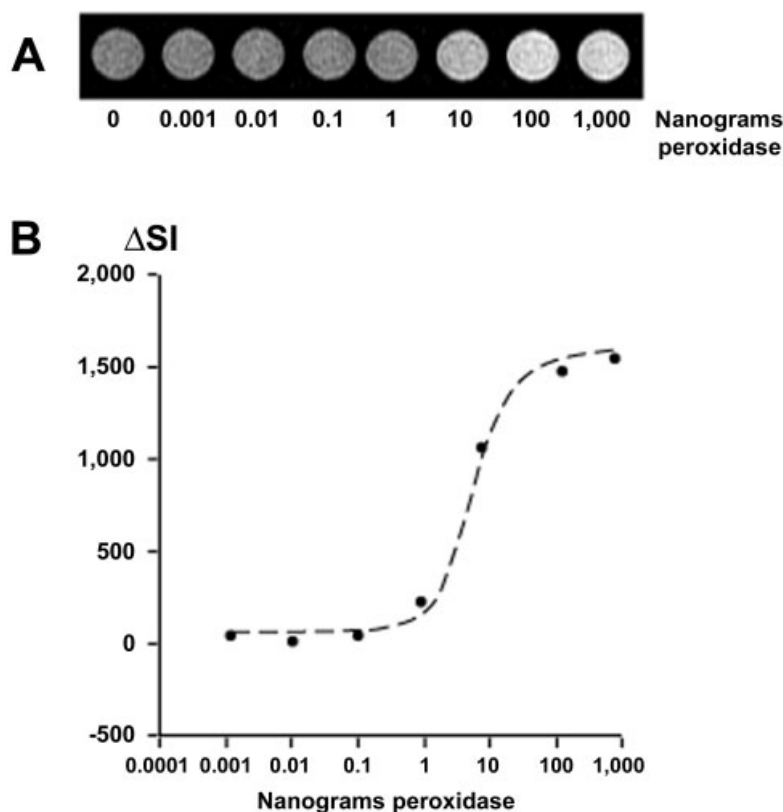


Figure 6.14 Detection of peroxidase using magnetic resonance signal amplification strategy. **A)** T1-weighted magnetic resonance image containing gadolinium 1,4,7,10-tetraazacyclododecane-N,N',N'',N'''-tetraacetic acid (D-DOTA) in the presence or the absence of the enzyme. **B)** The corresponding signal intensity increases over the control with no peroxidase present. Reprinted with permission from *Molecular Imaging* (Bodanov *et al.*, 2002).

vitro, this system allowed sensitive detection of E-selectin expression on the surface of interleukin-1 β -stimulated human endothelial cells. This substrate thus should allow probing for marker-enzyme activity *in vivo* as a tool for *in vivo* imaging of gene expression.

A different approach to using activatable contrast agents for *in vivo* imaging of gene expression was recently described using β -galactosidase as a converting enzyme (Louie *et al.*, 2000). In this magnetic resonance probe, the access of water to the chelated gadolinium is blocked by a β -galactosidase substrate (galactopyranose) that is removed by cleavage through β -galactosidase. Following cleavage, the paramagnetic ion can interact directly with water protons resulting in a T1 decrease of about 50 percent (Ahrens *et al.*, 1998; Louie *et al.*, 2000). The feasibility of detecting β -galactosidase expression has been demonstrated in a *Xenopus laevis* embryo model in which the expression vector for a β -galactosidase and the contrast

agent were co-injected (Louie *et al.*, 2000). Regions of higher intensity in the magnetic resonance image correlated with regions expressing marker enzyme.

6.4.5.3 Internalising receptors

A different approach to imparting cellular specificity into MRI is to selectively shuffle magnetic substances into the cell by over-expression of internalising cell receptors.

The native transferrin receptor is over-expressed in variety of tumour cells and has been extensively investigated in cancer research (Cotten *et al.*, 1993; Thorstensen & Romslo, 1993) and for imaging applications (Kayyem *et al.*, 1995; Koretsky *et al.*, 1996).

The synthesis of native transferrin receptor by a cell is tightly regulated by an iron-dependent negative feedback mechanism, so that high levels of intracellular iron result in down-regulation of the native transferrin receptor (Casey *et al.*, 1988 and 1989). By employing a genetically engineered transferrin receptor that lacks this negative feedback mechanism, this approach can be utilised as an imaging marker gene in MRI. The introduction of this genetically engineered transferrin receptor into 9L glioma cells resulted in the accumulation of transferrin-labelled superparamagnetic nanoparticles within the tumour cells (Weissleder *et al.*, 2000). Expression levels of the receptor correlated with magnetic resonance signal intensities on T2-weighted images in cells. Tumours expressing this up-regulated transferrin receptor were easily distinguishable *in vivo* from the same tumour cell line lacking this receptor by virtue of lower magnetic resonance signal intensities compared to the control (figure 6.15). This example represents the first

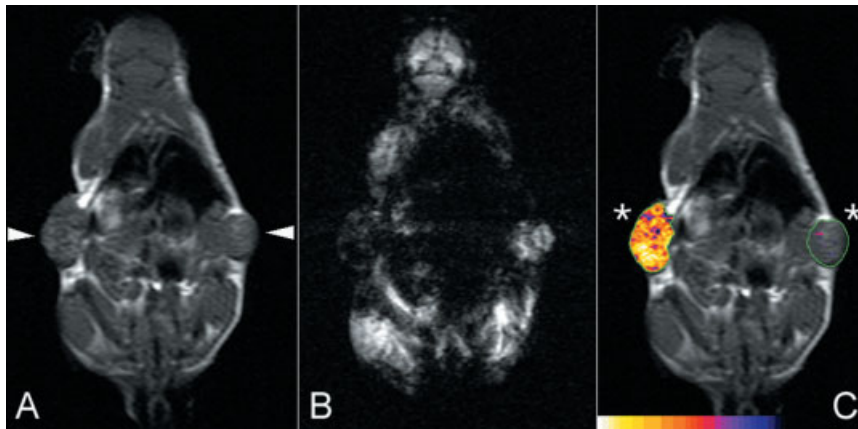


Figure 6.15 *In vivo* imaging of native transferrin receptor expression 24 hours after intravenous injection of transferrin-labelled monocrySTALLINE iron oxide nanoparticle. **A)** T1-weighted, **B)** T2-weighted, and **C)** T1-weighted image with colour-encoded magnetic resonance signal change). 9L gliosarcoma cells were transfected with an engineered transferrin

receptor (left tumour in **B** and **C**) that shuttled targeted superparamagnetic iron oxides into the cells. Note the strong signal decrease in the transferrin receptor-positive tumour on the left compared to the control tumour on the right. Reprinted with permission from *Nature Medicine* (Weissleder *et al.*, 2000).

proof of principle that imaging of transgene expression *in vivo* by MRI is feasible (Weissleder *et al.*, 2000). New probe design has now improved the labelling process of the targeted iron oxides. The number of transferrin molecules attached per iron oxide nanoparticle has been increased four-fold and binding and uptake by cells has been increased by more than ten-fold. This translated into an imaging probe that was 16 times better for imaging gene expression *in vivo* than the first generation of transferrin receptor-targeted iron oxides. This novel probe for MRI may substantially increase the sensitivity of detection of endogenous or genetically induced transferrin receptor expression in a small numbers of cells (Högemann *et al.*, 2000).

6.5

MRS – Magnetic resonance spectroscopy

6.5.1

Introduction

During the last four decades, nuclear magnetic resonance spectroscopy (NMRS) or magnetic resonance spectroscopy (MRS) has seen increasing use in chemical, physical and biochemical research for the analysis of the molecular structure or composition of small probes. *In vivo* MRS provides biochemical information on the function of animal and human organs, on the composition of body fluids and on the concentration of metabolites. Like magnetic resonance imaging, (MRI) MRS has the advantage of being noninvasive, safe and reproducible. Because the specific energy of the radiowaves used for magnetic resonance studies are much smaller than the binding energy of chemical bonds, dynamic biochemical changes can be observed without harming the object of investigation.

The term ‘clinical MRS’ is used to refer to spectroscopic investigations performed with scanners designed for imaging purposes in the clinic. The most common applications utilise on hydrogen and phosphorus as tracers. Human whole-body MRS is usually carried out in 1.5 Tesla scanners, although clinical research systems that operate at up to eight Tesla also exist. Increasing magnetic field strengths improve sensitivity and resolution, but the production costs of high-field whole body systems that operate at more than two Tesla are prohibitive. *In vitro* MRS investigations of tissue samples, body fluids or cell samples are usually carried out at up to 14 Tesla, far beyond the range of wide-bore clinical systems for human or animal imaging. This difference is important, because the signal-to-noise ratio increases linearly with field strength. Therefore, biomedical and especially clinical MRS applications are hampered by lower signal-to-noise ratios.

On the following pages we will give a short overview on the principles of MRS and current applications especially for detection of gene expression *in vivo*. An in-depth discussion of the technical aspects of magnetic resonance would be beyond the topic of this overview. For such discussion, the reader is referred to a recent excellent review (Boesch, 1999).

6.5.2

Principles and basics

Nuclear magnetic resonance (NMR), first demonstrated in 1946 (Bloch, 1946), is defined as the enhanced absorption – and subsequent emission – of energy that occurs when nuclei of atoms within an external magnetic field are exposed to radio-frequency energy at a specific frequency, called the resonance frequency. Every magnetic resonance emission signal has three components: amplitude, frequency, and phase. With MRS the frequency of the signal is used to detect nuclei of different elements within a sample. However, only compounds present in millimolar concentrations are detectable with standard *in vivo* MRS. Currently, MRS is being developed as a noninvasive method to assess transgene expression indirectly by means of magnetic resonance visible intracellular markers. These markers are intracellular metabolites (so-called ‘marker metabolites’) associated with exogenous enzyme expression and function (Bell & Taylor-Robinson, 2000). Because MRS is applicable not only *in vitro* but also *in vivo*, experiments to detect the spatial and temporal distribution of gene expression within living organisms are feasible with this technique.

Nuclei of various atoms, *e. g.*, hydrogen (^1H), nitrogen (^{15}N), fluorine (^{19}F), sodium (^{23}Na) and phosphorus (^{31}P), produce radiofrequency signals in a strong magnetic field as an answer to an electromagnetic pulse (Gadian, 1982). This signal, called free induction decay, can be resolved by a mathematical operation (Fourier transformation) into a frequency spectrum. The free induction decay contains all the information about the magnetisation of the nuclei, in particular about different contributors in a mixture of chemical compounds. The basic principle of MRS is that different elements resonate at different frequencies. But even nuclei of the same element may resonate at slightly different frequencies if their chemical environment is different. The frequency difference of the same nucleus is the result of the influence of nearby nuclei and is called ‘chemical shift’. As an example, the six hydrogen nuclei in ethanol resonate at three different frequencies, which is explained in figure 6.16. Because of the chemical shift, a spectrum of resonance frequency of a single nucleus has the potential to detect different chemical substances within the sample. For this purpose, comparison of chemical shift values in a sample with values in a table allows identification of chemical substances. For example, a ^{31}P phosphorus spectrum of a human muscle allows detection of seven different signals: phosphomonoesters, inorganic phosphate, phosphodiester, phosphocreatinine, and the three phosphates of 5'-adenosine triphosphate (ATP). Because the intensity of the signal is correlated with the concentration of each substance, MRS can measure metabolic activities and some gene products in noninvasive fashion.

For MRS in living organisms, localisation of spectra is usually of interest. In order to detect the magnetic resonance signal, every magnetic resonance system uses either receiver coils which can detect the radiofrequency signal from the whole body (‘body coils’) or coils that measure signals from the surface of the organism (‘surface coils’). The use of a surface coil is the simplest means of localisation, since it only detects signals from nearby anatomic structures. However, this localisation is crude and restricted to areas close to the body surface and does not allow flex-

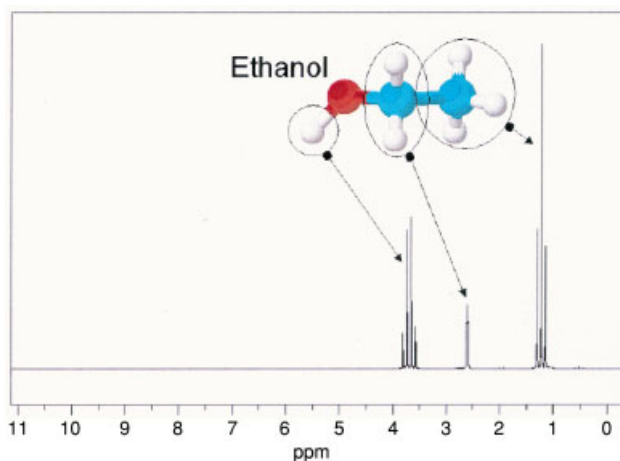


Figure 6.16 Proton spectrum of ethanol. The frequency spectrum is shown on the horizontal axis and signal intensity is shown on the vertical axis. The resonance frequency of the hydrogen nuclei (white balls) of the ethanol molecule differ because of chemical shift. The chemical shift on the horizontal axis is expressed in units of 10^{-6} or parts per million (ppm) of the applied magnetic field. Hydrogen atoms in different local environments of the sample molecule give different peaks according to the electronic character of their environment. These signals can be assigned to methyl, methylene and the hydroxy groups. Chemical

shift is relative to an internal standard such as tetramethylsilane or 3-trimethylsilylpropionic acid, which has an intense signal to the right from almost all other hydrogen nuclei and defines the parts per million scale of zero. Usually, no vertical scale is given since the vertical amplification is influenced by many parameters. For absolute quantification reference substances of known concentration must be measured. The spectrum was taken from the Integrated Spectral Data Base System for Organic Compounds (<http://www.aist.go.jp/RIODB/SDBS/>).

ible adaptation to the shape and size of the volume of interest. Nevertheless, this technique is still in use today because of its high signal-to-noise ratio. With modern systems, it is possible to define a volume of interest on an anatomic image and then to acquire spectra from that region only. This is called ‘image-localised spectroscopy’. In principle, MRS uses the same strategies to achieve spatial resolution as MRI. Selective excitation and phase-encoding are regularly used in MRS but the application of gradients – directed local changes in the static magnetic field – during signal acquisition is not possible. These readout gradients would destroy the spectral information within the magnetic resonance signal. In fact, the absence of this readout gradient could be seen as the fundamental difference between imaging and spectroscopy on the level of the pulse sequences. However, sophisticated techniques allow the acquisition of spectral information even in the presence of magnetic gradients (Dixon, 1984; Sepponen *et al.*, 1984). In conclusion, localised *in vivo* spectroscopy is a combination of diagnostic MRI and high-resolution NMR. The spatial resolution of localised ^1H -NMR spectroscopy *in vivo* is in the centimetre range for human applications and in the millimetre range for dedicated systems used for small animal MRS. The spatial resolution for other nuclei is even lower.

MRS is undoubtedly a very complex and demanding technique. However, it offers a unique window on various aspects of physiology and pathology. For example, real-time observation of biochemical reactions and of compartmentalisation and transport processes is possible. MRS has also been valuable in studies of the blood-brain barrier and of tissue microstructure. Studies of other physiological processes, such as glucose uptake in muscles, muscle force and tumour viability are also feasible. For descriptions of these sophisticated experiments, involving such techniques as spin saturation and anisotropy of relaxation times, the reader is referred to other literature (Boesch, 1999).

6.5.3

In vivo spectroscopy of mRNA and protein expression

Analysis of transgene expression *in vivo* currently requires destructive and invasive molecular assays of tissue specimens in nearly all cases. A noninvasive methodology for assessing the location, magnitude, and duration of transgene expression *in vivo* would facilitate the correlation of therapeutic response with transgene expression and would allow optimisation of gene therapy protocols. Analysis of mRNA and protein expression *in vivo* by means of MRS requires the use of reporter proteins or metabolites (figure 6.17). An ideal reporter gene must satisfy a number of criteria. (i) It must provide a unique signal so that transgene expression

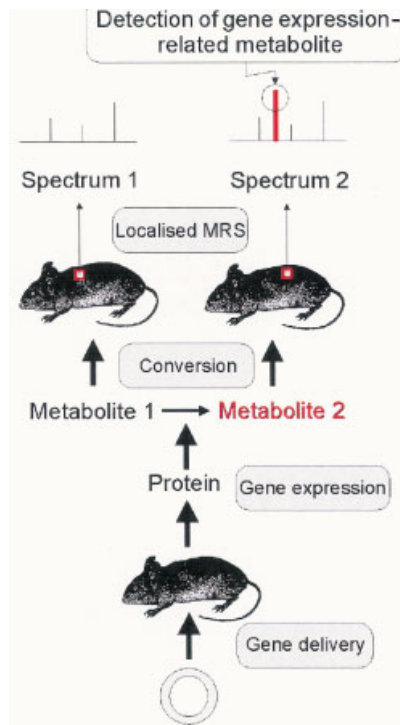
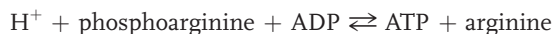


Figure 6.17 Principle of detection of gene expression by the means of localised magnetic resonance spectroscopy (MRS). A transgene is expressed encoding an enzyme that catalyses a reaction forming metabolites detectable by MRS. Localised MRS not only detects gene expression but also localises it within the organism (see text for details).

can be monitored against the background of physiological gene expression. (ii) The reporter gene should be small enough to be co-expressed with therapeutic genes in bicistronic constructs so that it can serve as a proportional marker for the expression of the therapeutic gene. (iii) It should not interfere with normal tissue function. (iv) It should ideally code for a syngeneic marker protein in order to prevent an immunological response.

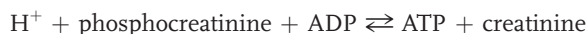
A recent study has demonstrated the potential of MRS for visualising transgene expression in cancer cell in a human tumour xenografts model of mice (Stegman *et al.*, 1999). In this study, the enzymatic activity of yeast cytosine deaminase expressed in subcutaneous tumours was quantified using ^{19}F -MRS to observe the conversion of 5-fluorocytosine to 5-fluorouracil and major metabolites. The pro-drug conversion reaction catalysed by cytosine deaminase is MRS-observable because 5-fluorouracil and other metabolites are fluorinated, and ^{19}F produces an MRS signal with no background due to endogenous compounds. In carcinomas expressing yeast cytosine deaminase, MRS depicted fluorinated nucleotides and fluoro- β -alanine, a catabolic breakdown product of 5-fluorouracil, in comparison to control tumours which showed a 5-fluorouracil peak only. These results demonstrate the feasibility of using MRS to noninvasively monitor therapeutic transgene expression *in vivo*.

Another approach was used to detect the expression of *Drosophila melanogaster* arginine kinase within the hindlimb muscle of neonatal mice *in vivo* (Walter *et al.*, 2000). Arginine kinase is the invertebrate correlate of creatinine kinase and is not expressed in muscle and nerves. It catalyses the reaction:



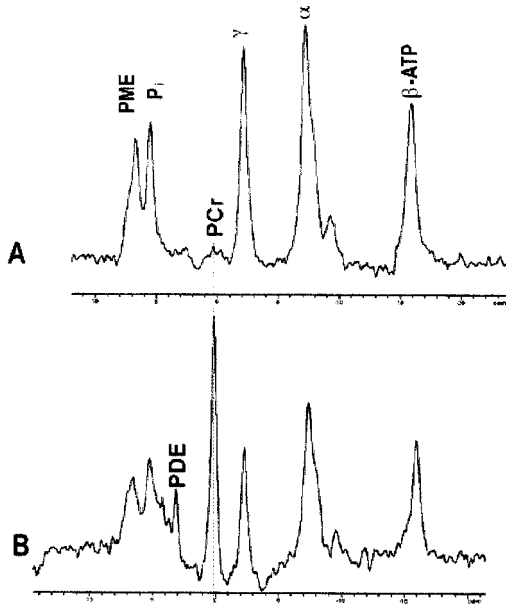
Because arginine kinase is not normally present in mammalian muscle tissue, its expression can be detected with ^{31}P -MRS due to the presence of the marker metabolite phosphoarginine. Indeed the presence of this marker metabolite was detected up to eight months after adenovirus-mediated gene delivery. The authors concluded that this nontoxic gene product allows noninvasive and repeated monitoring of gene expression after virus-mediated gene transfer to muscle.

Whereas yeast cytosine deaminase and the arginine kinase used by Walter *et al.* are xenogenes in the model organisms used in their experiments, other approaches are based on the expression of syngeneic enzymes. Koretsky *et al.* established the enzyme creatinine kinase as a possible marker gene for MRS, based on the observation that creatinine kinase is normally not expressed in liver, kidney, and pancreas (Koretsky *et al.*, 1990). Creatinine kinase catalyses the reaction:



^{31}P magnetic resonance detected the marker metabolite phosphocreatinine in liver extracts from mice. This approach can also be used for *in vivo* detection of creatinine kinase gene expression, as was shown in a recent work with the brain isoform of this enzyme (figure 6.18) (Auricchio *et al.*, 2001). The authors were

Figure 6.18 *In vivo* ^{31}P -magnetic resonance spectroscopy (MRS) detection of murine livers transduced with the creatine kinase B gene. **A)** Representative spectrum derived from control livers ($n = 6$). **B)** Spectrum derived from transduced livers ($n = 6$). The following abbreviations are used in the figure. PCr: phosphocreatinine; PME, phosphomonoesters; P_i , inorganic phosphate; PDE, phosphodiester; α , β and γ -ATP, α , β and γ resonance of adenosine triphosphate. The figure is reproduced with permission from the *Proceedings of the National Academy of Sciences of the USA* (Auricchio *et al.*, 2001).



able to detect phosphocreatinine in a concentration as low as approximately one micromole per gram of wet liver tissue. However, the technique was rather invasive because there was a need for surgical exposure of the liver surface to the surface coil. Creatinine kinase may also be suitable as a marker gene even in tissues that normally express the enzyme, because the increment in creatinine kinase activity can be determined by magnetization transfer experiments. With this technique the muscle creatinine kinase isoform and the brain isoform can be distinguished because they are localised in different subcellular compartments (Roman *et al.*, 1997).

A different approach is to use MRS to detect side effects of gene therapy as an indirect estimation of gene expression. In one recent study, the accumulation of polyunsaturated fatty acids was linked to apoptosis induced by ganciclovir during therapy of *Herpes simplex* virus thymidine kinase expressing tumours (Hakumäki *et al.*, 1999). MRS helped to show that apoptosis was closely associated with the accumulation of lipids.

6.5.4

Discussion

Magnetic resonance spectroscopy shows promise as a noninvasive method to assess transgene expression indirectly by means of magnetic resonance-visible intracellular markers. These markers are metabolites associated with exogenous enzyme expression and function (Bell & Taylor-Robinson, 2000). Because they are catalysts that can turn over a large amount of substrate, the use of enzymes as marker

proteins may be seen as an *in vivo* amplification strategy. This amplification allows detection of gene expression even with the relatively insensitive magnetic resonance techniques MRI or MRS, which have an *in vivo* detection limit in the low millimolar or in the high micromolar range depending on the experimental conditions (Weissleder & Mahmood, 2001).

The ability of MRS to follow prodrug conversion dynamically and to distinguish between individual metabolites is a significant advantage over other molecular imaging methods such as positron emission tomography (PET) and MRI. This is especially true for gene-directed enzyme-prodrug therapy. Gene-directed enzyme-prodrug therapy is designed to overcome the dose-limiting systemic toxicity of cancer chemotherapy by expressing nonmammalian enzymes in tumour cells; these enzymes convert low-toxicity prodrugs to cytotoxic metabolites. The success of this therapy depends on adequate distribution and expression of the therapeutic transgene in the target tissue. Because the method is noninvasive, several measurements over time can be conducted, greatly enhancing the performance of the model systems. This was shown to be useful for yeast cytosine deaminase gene therapy as explained above (Stegman *et al.*, 1999). In principle, similar MRS methodologies might also be developed for other gene-directed enzyme-prodrug therapeutic approaches.

Whereas MRS is able to detect many metabolites within an organism, other metabolites require special editing techniques because their spectrum is 'hidden' due to signal overlap. In order to detect metabolites with ^1H -MRS beneath the very large signal from water and fat, the resonance of these substances must first be removed using a so-called 'crusher gradient'. Examples of hidden resonances that can be extracted by spectral editing techniques are lactate against strong fat background signals (Bloch *et al.*, 1995) and γ -aminobutyric acid in brain tissue (Rothman *et al.*, 1993). Only metabolites that are freely moving in solution are readily accessible to MRS. NMR signals from solids cannot be detected by commonly used biomedical magnetic resonance methods because of their extremely short T2 relaxation time (Boesch, 1999). On the other hand, some metabolites may only be visible under pathological conditions, thus providing diagnostic information.

One inherent problem of MRS is quantification of nuclei *in vivo*. As mentioned in the legend of figure 6.16, spectral intensities are not proportional to the concentration of the nuclei. Whereas in theory the area under the specific spectrum represents the concentration of the spins, with the low field magnetic resonance scanners used for *in vivo* spectroscopy a substantial overlay of spectra occurs in most cases. Therefore, exact quantification of the area under the peak is not always possible within crowded *in vivo* spectra. In addition, the peak amplitude depends on the instrument used. For these reasons, an internal standard, *i.e.* a metabolite of known concentration, has to be used for calculation of the approximate concentration of the nuclei under investigation. Whereas the signal of water as an internal standard after correction for relaxation effects is quite stable and robust for quantification with ^1H -MRS, signals from other nuclei are more difficult to quantify. Many *in vivo* MRS studies therefore report results in terms of metabolite ratios.

Another problem is that clinical magnetic resonance scanners with spectroscopy options suitable for human *in vivo* applications are not yet commonly available. As a result, few radiologists have experience with MRS. However, this may change rapidly within the near future. High-field clinical scanners with a static magnetic field of three Tesla have recently been approved by the Food and Drug Administration (FDA) for human use and will be available within the next years. Altogether, the field of biomedical MRS needs a high degree of specialisation and skills. At the moment, it is separated into three different fields: (i) Basic pathophysiological studies in animals, isolated organs, and cell extract performed in small-bore high-field magnetic resonance systems, (ii) basic pathophysiological studies in humans using high-field whole-body magnetic resonance systems, and (iii) diagnostic clinical applications using standard clinical scanners. Only a combination of these fields within a single cooperative research effort will advance the application of *in vivo* magnetic resonance spectroscopy to the analysis of (trans)gene and protein expression.

In conclusion, magnetic resonance spectroscopy may come to play an important role in monitoring gene therapy, both as a tool to detect gene expression and as a means to measure associated changes within the targeted organs and cells, thus providing a measure of efficacy. In addition, MRS can be performed as an adjunct to MRI studies providing modern high-field clinical scanners are used, combining the advantages of both methods to study gene expression and therapeutic response *in vivo*.

6.6

PET – Positron emission tomography

6.6.1

Introduction

Positron emission tomography (PET) is based on the detection of two 511-kilo-electron-volt annihilation photons that originate from radionuclides emitting a positron (β^+ -particle). The most commonly used positron-emitting radionuclides are characterised in table 6.1. Unfortunately, the commonly used PET radioisotopes have half-lives in the minute range. For this reason, production and labelling facilities must be located near the PET scanner. Positron emitters are proton-rich isotopes that exhibit energy differences between themselves and their daughter nuclides of at least 1.02 mega-electron-volts. Formally, a proton is converted to a neutron (which remains in the atomic nucleus) and a positron (which is expelled into the surrounding matter). The energetic positron, or β^+ -particle, loses energy while passing through matter. When almost all energy has been lost, it combines with an atomic electron of the medium, and both particles are annihilated, giving rise to two photons of 511 kilo-electron-volt (*i.e.*, the energy analogue of the positron and electron masses, respectively) emitted at an angle of 180 degrees to each other. These photons are the so-called annihilation radiation (figure 6.19).

Table 6.1 Characteristics of typical PET radionuclides. The following abbreviations are used: MeV, mega-electron-volt.

Radionuclide	Half-life time	Mode of decay	Energy of positron [MeV]
Carbon (^{11}C)	20.4 minutes	β^+ (100 percent)	0.385
Nitrogen (^{13}N)	Ten minutes	β^+ (100 percent)	0.491
Oxygen (^{15}O)	Two minutes	β^+ (100 percent)	0.735
Fluorine (^{18}F)	110 minutes	β^+ (97 percent)	0.242
Iron (^{52}Fe)	Eight hours	β^+ (55 percent)	0.340
Copper (^{64}Cu)	13 hours	β^+ (18 percent)	0.278
Gallium (^{68}Ga)	68 minutes	β^+ (89 percent)	0.740
Arsenic (^{72}As)	26 hours	β^+ (87 percent)	1.167
Rubidium (^{82}Rb)	75 seconds	β^+ (95 percent)	1.409
Yttrium (^{86}Y)	15 hours	β^+ (32 percent)	0.554
Iodine (^{124}I)	Four days	β^+ (23 percent)	0.819

PET imaging is based on the detection of these two 511-kilo-electron-volt annihilation photons originating from positron emitters within an object (such as a patient or an animal). Thus, the PET scanner does not detect positrons directly, but uses the features of positron annihilation to determine their special location. In contrast to conventional nuclear imaging, where mechanical collimators are used to filter out all photons which will not hit the camera at a ninety-degree angle, the two 511 kilo-electron-volt photons in PET are detected by two detectors (figure 6.19.A). Since the annihilation photons are emitted in opposite directions, the detection of a coincidence event localises the positron annihilation to a point lying somewhere on the line that joins the two detectors (figure 6.19.B). Only two incidents detected within a very short timeframe are considered to result from the same annihilation process, which must lie on the connecting line between the two detectors hit by the two 511-kilo-electron-volt photons within this timeframe.

A typical PET scanner consists of hundreds of such detectors forming a ring that surrounds the object. Data are collected over 360 degrees around the axis of the study object (*i.e.*, an animal or a patient) and are used to reconstruct the spatial distribution of the activity within this object. During a typical PET experiment, several million coincidence events are recorded, forming a large number of intersecting coincidence lines and providing information about the quantity and location of positron emitters within the body. The image reconstruction is performed by filtered back-projection or in an iterative manner, leading to an image with a typical matrix in the range of 64×64 to 512×512 . This image enables one to localise the signal within the organism under investigation.

For technical reasons, the spatial resolution of PET scanners is limited to approximately six millimetres for human applications. However, dedicated systems

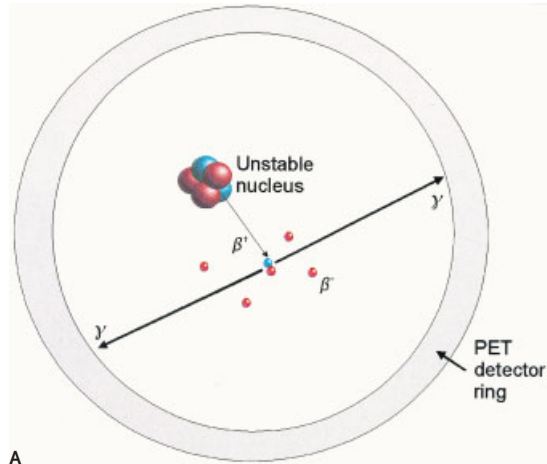
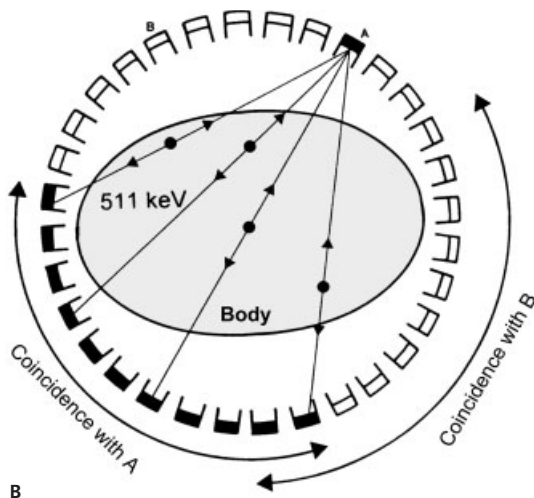


Figure 6.19 Basic principle of positron emission tomography (PET).

A) An unstable nucleus decays by emitting a positron (a positively charged electron, β^+ , light blue). The positron interacts with an electron of the organism (β^- , light red) to emit annihilation radiation (γ) at 511 kilo-electron-volt. **B)** Activity quantification by PET. The annihilation radiation is detected by the PET detector ring. Millions of such events are computed to a PET image which provides information on the localisation of the detected signals within the organism under investigation. The following abbreviation is used in this figure: keV, kilo-electron-volt. The figure is modified from Behr *et al.* (1999) and reprinted with permission from Harwood Academic Publishers.



for small animal imaging reach spatial resolutions up to two millimetres, which is close to the maximum theoretical resolution determined by the path length of the positron. Whereas this spatial resolution is usually less than with other imaging techniques, as for example magnetic resonance imaging (MRI) or computed tomography (CT), the advantage of PET is its extremely high sensitivity which enables to detect reporter probes in the picomolar range. This makes PET a valuable tool for *in vivo* molecular imaging.

6.6.2

From clinical PET to quantitative small laboratory animal tomography

PET may serve as an *in vivo* analogue to tissue dissection, well-counter quantification, auto-radiography, and other techniques involving either sectioning with subsequent imaging or counting of excised and dissected tissue samples from animals into which radioactively-labelled probes have been introduced previously. The technology allows for temporal and spatial resolution of the bio-distribution of molecular probes. Since the same living being can be used repeatedly, inter-individual variability can be minimised; and often, each animal can serve as its own control. In addition, molecular imaging approaches change the model organism under investigation from a 'black box' to a transparent one. This is because, in contrast to traditional approaches, as for example gene expression analysis using light microscopy to detect reporter gene products or *in situ* PCR, we can actually see the interaction of the molecular probe with the organism in real time, for example, in order to detect *in vivo* gene expression.

Whereas the principle of PET imaging is almost identical for animals and humans, PET scanner design is very important for animal applications. Clinical PET scanners have a spatial resolution that is much too low for imaging of small animals. Considering the 2,000-fold reduction in size from human beings to mice, one can estimate the technical challenge of constructing a PET scanner for imaging in this small rodent. Nevertheless, several PET imagers for this application are already commercially available (Chatziioannou, 2002).

6.6.3

Examples of molecular imaging and imaging of gene expression with PET

A wide variety of applications of animal and clinical PET for imaging of molecular events has been reported. Examples are single or longitudinal investigations of glucose metabolism after various interventions, investigations of the dopaminergic system in the brain, imaging of cell surface antigens by radiolabelled antibodies and their fragments, or the development of methods for imaging reporter gene expression *in vivo* (figure 6.21).

A paradigmatic approach is the imaging of glucose metabolism with 2-[¹⁸F]-fluoro-2-deoxy-D-glucose (FDG), an imaging approach originally developed in the 1970s (Sokoloff *et al.*, 1977). The use of FDG is based on the observation that malignant tumours have an increased glycolytic rate. Whereas FDG initially follows the glycolytic pathway, it is not a substrate for subsequent reactions and remains trapped within the cell with an accumulation rate that is proportional to the rate of glucose utilisation (Price, 2001). The relatively high levels of FDG within tumours after systemic injection of this PET imaging probe in comparison to surrounding normal tissue is paralleled by a higher PET signal in tumours and enables us to detect even small lesions. These properties appear to be common for many malignancies. In lung and colorectal cancer, melanoma, and lymphoma

PET has already been shown to improve the detection and staging of tumours in a clinical setting (Phelps, 2000a).

6.6.4

Quantitative imaging in oncology

There is a range of PET probes that are of interest in cancer models, including FDG, 3'-deoxy-3'-[^{18}F]-fluorothymidine (FLT) as a marker of DNA replication and cell proliferation (Shields *et al.*, 1998), labelled antibodies, and their fragments (generated enzymatically or by means of genetic engineering) against tumour-associated antigens (Kairemo, 1993). Studying (xeno-)transplanted tumours is often simple since they can be placed well separated from all major organs in the back, thigh, shoulder, front or hind legs of the animals. Under these circumstances, corrections for partial volume effects are more easily possible than in, otherwise more preferable, (patho-)physiological metastatic models. By PET imaging and quantification, the usual variability between animals is effectively circumvented and each animal can serve as its own (time-wise) longitudinal control (Sundaresan *et al.*, 2001). Since PET is capable of surveying the whole animal, it may be used for monitoring of together with morphological imaging modalities (*e.g.*, small animal magnetic resonance imaging or computer tomography), local tumour growth (or regression) or for assessing the spread of metastatic disease.

6.6.5

Quantitative imaging of transgene and endogenous gene expression

Increasing interest is focussing on methods to measure transgene expression *in vivo*. In analogy to β -galactosidase, used as a standard reporter gene in molecular biology, a PET reporter gene (PRG) expresses a protein capable of trapping or binding a positron-emitter-labelled probe, called a PET reporter probe (PRP). The most important example of this approach is the use of the *Herpes simplex* virus thymidine kinase (HSV-TK) as a PET reporter gene (figure 6.20). This enzyme, in comparison to human thymidine kinase, has a broader substrate specificity and can phosphorylate nucleoside analogues that are structurally different from its natural substrate deoxythymidine. Beside of its use as a therapeutic agent, transgenic HSV-TK can be used for imaging purposes, as was first shown by the group of Juri G. Tjuvajev (Tjuvajev *et al.*, 1995). Nuclear imaging of HSV-TK with PET requires injection of a detectable nucleoside analogue, for example, 8-[^{18}F]-fluoroganciclovir (FGCV) or [^{124}I]-2'-deoxy-2'-fluoro-5-iodod-1 β -D-arabinofuranosyluracil (FIAU). These analogues are phosphorylated as the physiological substrate and therefore become trapped within the cell, because the phosphorylated metabolites cannot penetrate the cell membrane. Since the non-converted nucleoside analogue is rapidly excreted from the organism, the detection of the ^{18}F fluorine-induced PET signal correlates well with reporter gene expression (figure 6.20) (Hospers *et al.*, 2000). Gambhir *et al.* (1999a) recently showed that this approach of transgene expression analysis is indeed quantitative. In this study, microPET measurements of gene ex-

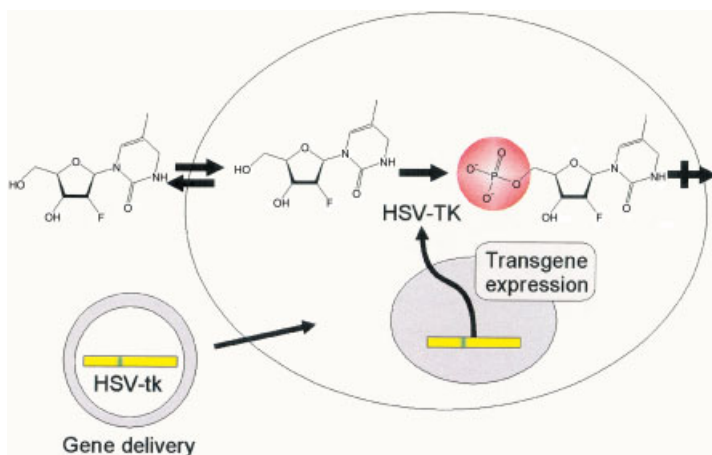


Figure 6.20 Example of *Herpes simplex virus thymidine kinase (HSV-TK)* as an **positron emission tomography (PET) reporter gene**. The reporter gene codes for HSV-TK, which converts the freely diffusible reporter probe, in this case [^{124}I]-2'-deoxy-2'-fluoro-5-iodo-1β-D-arabino-furanosyluracil (FIAU), to the non-diffusible monophosphate form. The molecule is further phosphorylated to the diphosphate and tri-

phosphate forms by cellular kinases and subsequently trapped within cells that express the transgene. If there is no transgene expression, FIAU diffuses out of the cells and is cleared to the bladder. If the PET image is acquired within the correct time window, the PET signal is proportional to the level of transgene expression.

pression correlated highly with *in vitro* measurements (Northern blot analysis and biochemical measures of enzyme activity) of reporter gene expression. The use of such reporter genes allows gene therapy protocols to be monitored in both animals and humans (Alauddin *et al.*, 2001). Other examples of PET reporter genes are receptors such as the somatostatin or dopamine receptors, which can be detected using labelled ligands due to the inherently high sensitivity of PET. In this case, the gene coding for the receptor is the PET reporter gene and the radiolabelled ligand for the receptor is the PET reporter probe; gene expression is therefore assayed by imaging of the ligand-receptor interaction. 3-(2'-[^{18}F]fluoroethyl)spiperone (FESP) is a ligand for the dopamine D2 receptor. For this PET reporter gene/probe system a good correlation between *in vivo* and *in vitro* measurements was also shown (figure 6.21) (MacLaren *et al.*, 1999). Other approaches, as for example the combination of therapeutic and reporter genes in bicistronic expression vectors, are also under investigation to quantify targeted gene expression *in vivo* (Yu *et al.*, 2000).

Different imaging approaches are under investigation to detect and quantify endogenous gene expression. An interesting approach is the development of radiolabelled antisense oligonucleotides (RASONS), which, in similar fashion to the labelled antisense oligonucleotides used in *in situ* hybridisation experiments, have the theoretical ability to detect physiological concentrations of mRNA *in vivo* (Gambhir *et al.*, 1999a). Tavitian *et al.* (1998) labelled oligonucleotides with or with-

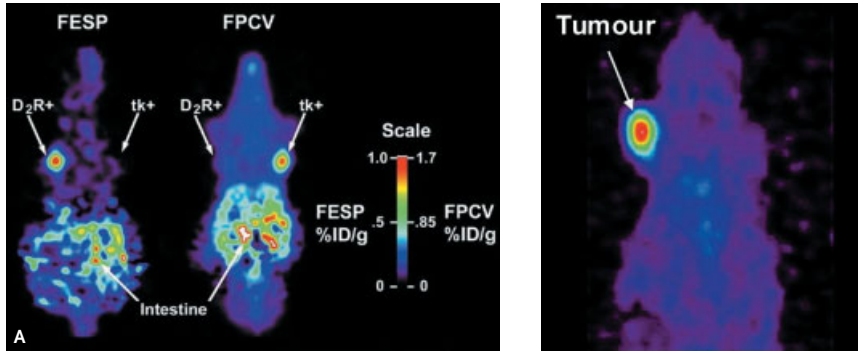


Figure 6.21 Experiments using positron emission tomography (PET) for the analysis of gene expression.

A) Image of a mouse bearing two tumours, each of which expresses a different reporter gene (left: dopamine receptor D₂; right: *Herpes simplex virus 1* thymidine kinase (HSV1-TK) reporter gene). On two separate days, the same mouse was injected with two different probes (3-(2'-[¹⁸F]fluoroethyl)spiperone, FESP, and 8-[¹⁸F]fluoropenciclovir, FPCV), each of which was specific for one of the two reporter genes. The figure is modified from Chatziioannou (2002) and is reprinted with permission from the *European Journal of Nuclear Medicine*. **B)** Coronal PET scan of a mouse injected with a ¹²⁴I-labelled anti-carcinoembryonic antigen (CEA) antibody. Eighteen hours after the immunoconjugate injection, the carcinoembryonic antigen-expressing human colon tumour LS174T xenograft is clearly visible. The figure is modified from Chatziioannou (2002) and is reprinted with permission from the *European Journal of Nuclear Medicine*. The following abbreviations are used in this figure: D₂R+, dopamine receptor D₂-positive; tk+, thymidine kinase-positive; ID/g; injected dose per gram body mass in percent.

out backbone modification (phosphorothioate) with ¹⁸fluorine and were able to detect differences in the bio-distribution of these RASONS in the baboon. However, their study was designed to evaluate the pharmacokinetics but not the hybridisation of these RASONS. The main problem still to be resolved is how to differentiate specific binding of RASONS to mRNA from the background signal due to unspecific binding within the body. However, this technique, which can be named *in vivo* hybridisation, still needs confirmation.

Targeting the gene product of the endogenous gene of interest has the advantage of a relatively higher concentration of this protein in comparison to mRNA or even DNA. This protein may be a receptor or an enzyme. If the protein is a receptor, no further signal amplification occurs, because one labelled probe binds to only one receptor molecule. If the protein is an enzyme, some amplification of the signal may occur, because one enzyme is able to produce many labelled metabolites of the radiolabelled probe. The disadvantage of this method is that a specific labelled ligand or substrate has to be developed and evaluated for each gene of interest. In some cases a usable PET reporter probe may even be impossible to construct. With the RASON approach imaging of any gene can theoretically be achieved by standard molecular biology methods (Phelps, 2000a).

Another approach is the use of a reporter gene that is driven by the same promoter as the gene of interest (Phelps, 2000b). The retention of the positron-emitter-labelled probe by the protein product of the PET reporter gene is usually proportional to the quantitative level of reporter gene expression. In turn, this reflects the level of expression of the gene of interest, having the same promoter or being coupled directly to the PET reporter gene. In this manner, the location, magnitude of expression, and time course of expression levels of any particular gene that is introduced into a laboratory animal (e.g., a mouse) can be monitored *in vivo*. This allows for monitoring of gene therapy protocols in animal models and, possibly, in human patients. The same approach can be used in transgenic mice, where every cell in the mouse carries the PET reporter gene. In this case, the signal is detectable only in those locations where the promoter driving the PET reporter gene is physiologically active. This enables study of endogenous gene expression in animal models.

6.6.6

Discussion and conclusions

In summary, PET is an imaging modality based on the detection of radionuclides in very low mass amounts. For molecular imaging purposes, PET reporter probes are designed by insertion of these radionuclides into molecules of specific interest or by replacement of non-radioactive atoms by positron-emitting ones. PET then assays the changing tissue concentrations of these labelled molecules or labelled products *in vivo*. With specially designed probes capable to bind to molecules of interest, PET is able to detect protein or, in the near future, even mRNA concentrations within an organism. In combination with specific PET reporter genes, this method is able to detect transgene expression *in vivo*. The major advantages of PET in comparison to other molecular imaging methods is its high sensitivity and its intrinsic ability to quantify. Major disadvantages are the relatively low spatial resolution, a long imaging time, and the short half-lives of the nuclides used for imaging purposes, making it necessary to have a technically demanding imaging probe production facility in close proximity to the scanner.

The applications of this methodology are widespread. ‘Genetic tagging’ of tumour cells that can be followed over time after injecting them into an animal, investigation of the efficiency of gene therapy vectors for delivering genes into experimental animals, study of genotypic interaction between cancer cells and the immune system, or interaction of gene expression patterns during development to see when certain genes of interest are switched on or off are some examples. This opens powerful research opportunities that take advantage of the ability of PET to longitudinally image and quantify gene expression throughout an organism over its entire life-span. Additionally, it provides an experimental platform to develop the principles for applying imaging of gene expression with PET scanners to humans (Haberborn *et al.*, 2002).

6.7

SPECT – Single photon emission computed tomography

6.7.1

Introduction

The history of nuclear medicine and scintigraphic tools starts with the discovery of radioactivity by Antoine Henri Becquerel in 1896 (Becquerel, 1999; Feld & De Roo, 2000). In 1923, the father of nuclear medicine, the Hungarian chemist Georg Karl von Hevesy, was the first to use radioactive isotopes of phosphorus (^{32}P), lead (^{210}Pb) and bismuth (^{210}Bi) to follow physiological processes in plants, animals and humans (Feld & De Roo, 2000; von Hevesy, 1964). In 1938, the first iodine uptake studies in the thyroid were carried out by Saul Hertz, Arthur Roberts and Robley Evans using ^{131}I (Evans, 1975; Feld & De Roo, 2000; Hertz *et al.*, 1938). In 1958, Hal O. Anger developed the γ -camera which allowed for the first time the measurement of the spatial distribution of radioactivity in living animals and humans (Anger, 1958; Feld & De Roo, 2000; Sørensen & Phelps, 1987). In 1963, single photon emission computed tomography (SPECT) was developed by David E. Kuhl and Roy Q. Edwards. This technique allowed three-dimensional reconstruction of the spatial distribution of radioactivity (Feld & De Roo, 2000; Kuhl & Edwards, 1963). A further step forward was the introduction of positron emission tomography (PET) by Michel M. Ter-Pogossian (Feld & De Roo, 2000; Ter-Pogossian *et al.*, 1975) (see also the corresponding section in this book). Since then, the field has progressed due to technical improvements in the scanning modalities and the development of more specific radio-pharmaceuticals.

The development of nuclear medicine as a medical subject in its own right is mainly a result of the ability of isotopic techniques to follow physiological processes *in vivo* in a noninvasive fashion. A target of interest, the imaging or (endo-radio-)therapeutic probe, is labelled with a radionuclide, injected, and the accumulation of the radioactivity – which is related to the accumulation of the molecule – is monitored externally using a γ -camera. In this way, the expression of receptors, antigens or other biomolecules can be detected in different tissues.

6.7.2

Principles and basics

In order to be suitable for imaging, the radionuclide used for labelling of the probe must emit γ -radiation in the energy range of 60 to 440 kilo-electron-volts. The energy for pure scintillation counting by a γ -camera is suitable as well at lower or higher energies. The mode of operation of a γ -camera is shown in figure 6.22. A lead collimator installed in front of the γ -camera filters out all rays that are not perpendicular to the plane of detection of the camera resulting in an image with spatial resolution. When the γ -rays strike the sodium iodide crystal, a flash of light is produced which is detected by photomultipliers. The image is then reconstructed by a computer. An improvement on this technique is single photon emission com-

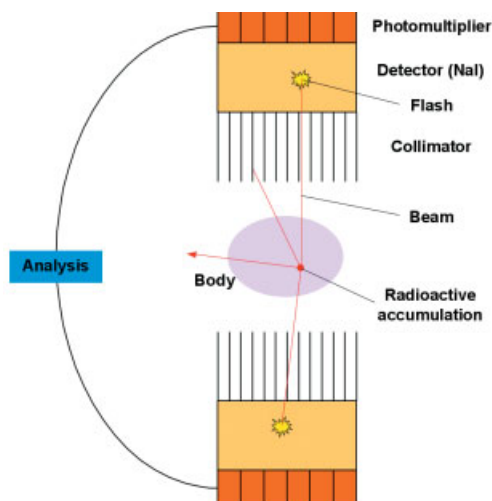


Figure 6.22 Mode of operation of a γ -camera used for single photon emission computed tomography (SPECT). A lead collimator is installed in front of the camera to filter γ -rays that are not perpendicular to the plane of the camera. Only γ -rays striking the sodium iodide crystal generate flashes of light that are detected by photomultiplier tubes. Because light intensity is proportional to beam energy, and

because the angles of the incoming beams are restricted by the lead collimator, the γ -camera provides an image of the location and amount of a radionuclide within the body. With SPECT a series of planar images is obtained over a 360 degree arc, *i.e.* the heads of the γ -camera rotate around the body, allowing reconstruction of transaxial images. The following abbreviation is used in this figure: NaI, sodium iodide.

puted tomography (SPECT). In the SPECT technique, the heads of the γ -camera rotate around the body through an angle of 360 degrees. A three-dimensional image is then constructed in a fashion similar to that used in computer tomography in radiology. The resolution of the method with clinical cameras is in the range of one centimetre and is determined by the radionuclide and the collimator. Newer developments promise to improve the resolution to the low millimetre range for use in animals (Wirrwar *et al.*, 2001, Green *et al.*, 2001).

The characteristics of the radioisotopes that are currently in use are described in table 6.2. The advantage of iodine isotopes is their ability to bind covalently to carbon atoms and the possibility of using them to replace some chemical groups in biological molecules. *In vivo*, deiodases may deiodinate these compounds, leading to mixed bio-distributions between the free iodide (mainly in thyroid and stomach) and the labelled compound. The synthesis of stable compounds is therefore crucial and normally requires several synthetic steps. Labelling with radiometals is performed *via* a chelator that is coupled to the biomolecule of interest. The resulting complex is a relatively large chemical structure, the size of which may affect the biological and pharmacological behaviour of the substance to a greater extent than iodine isotopes. However, labelling with radiometals is more stable than iodine labelling, and if the radiometal/biomolecule complex is taken up by a cell

Table 6.2 A selection of the most commonly used γ -emitting radioisotopes in alphabetical order.

Nuclide	γ -energy [kilo-electron-volt]	Half-life time
Chromium (^{51}Cr)	320	27,7 days
Cobalt (^{58}Co)	122	271, 8 days
	136	
Gallium (^{67}Ga)	93	78,3 hours
	300	
Indium (^{111}In)	173	2,81 days
	247	
Iodine (^{123}I)	159	13,2 hours
Iodine (^{131}I)	364	8,02 days
Technetium ($^{99\text{m}}\text{Tc}$)	140	Six hours
Thallium (^{201}Tl)	167	3,05 days

(*i.e.*, internalised), since externalisation rates are very slow (residual labelling). This leads to a better accumulation of the activity in target tissue over time.

The main advantage of scintigraphy is the noninvasive *in vivo* detection of targets such as RNA, receptors, transporters and antigens using labelled biomolecules. The first step is to identify a target and to find a biomolecule that binds specifically to it. In the second step, based on the biomolecule a tracer molecule is synthesised that can be labelled with the radioisotope in an easy and fast way. This tracer is then examined for binding affinity and other properties such as metabolic and chemical stability (figure 6.23).

6.7.3

Direct imaging of gene expression

The idea of imaging gene expression directly with antisense molecules was born as a result of the increased interest of antisense therapy. Imaging techniques have been developed that allow the distribution of the antisense molecule to be tracked *in vivo* in a noninvasive fashion. Synthetic stabilised and radiolabelled single-strand oligodeoxyribonucleic acids (DNAs) or oligoribonucleic acids (RNAs) are designed to bind to a specific sequence during the translation or transcription mechanism of the gene and interfere with the expression of the gene. These radiolabelled antisense oligodeoxynucleotides are called RASONS. The pioneers in this field, Mrinal K. Dewanjee and Donald J. Hnatowich, began in the 1990s by labelling antisense molecules with radioactive iodine (^{131}I), technetium ($^{99\text{m}}\text{Tc}$) and indium (^{111}In) isotopes (Dewanjee *et al.*, 1991 and 1994a; Hnatowich *et al.*, 1995). They showed that the labelled antisense oligonucleotide continued to bind normally to the sense molecules. Mrinal K. Dewanjee succeeded in specifically detecting the *c-myc* oncogene

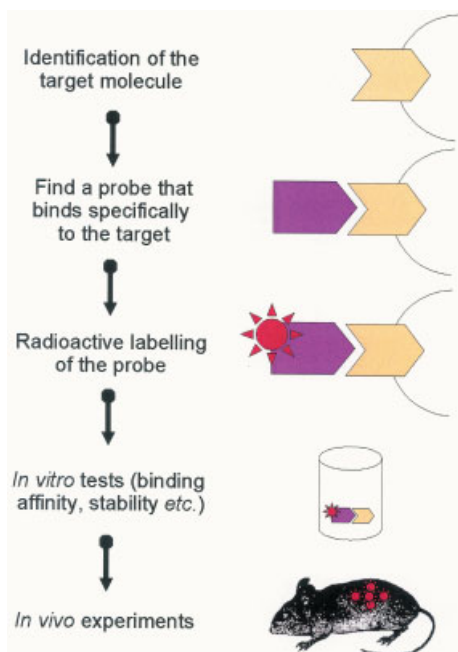


Figure 6.23 Illustration of the principle steps of molecular imaging with single photon emission computed tomography (SPECT). After identification and characterisation of a target molecule, a probe is designed that binds to the target with high specificity. After radioactive labelling of the probe *in vitro*, tests of the probe and target/probe interaction are conducted to determine if the radioactive labelling has altered the physicochemical properties of the probe. *In vivo* experiments are performed only after the probe has been shown to interact with the target with sufficient specificity and sensitivity *in vitro*.

at the mRNA level *in vivo* in 1994 (Dewanjee *et al.*, 1994b) by injecting ^{111}In -labelled sense and antisense molecules in BALB/c mice (animals that carry an autosomal recessive mutation associated with decreased sphingomyelinase and glucocerebrosidase activity and storage of sphingomyelin and glucocerebroside as well as cholesterol in lysosomes) with mammary tumours. A significantly higher uptake of the antisense probe was observed in the tumour demonstrating specific uptake. However, the method suffered from the general problem of transporting the antisense oligonucleotide to the site of the sense molecules limiting its widespread use.

6.7.4

Indirect imaging *via* expressed proteins

Indirect imaging *via* expressed proteins has been used in nuclear medicine for many years. The most well-known clinical example is the examination of the expression of the sodium/iodine symporter in different thyroid diseases using radioiodine. Radioactive compounds such as $^{123}\text{I}^-$, $^{131}\text{I}^-$ or $^{99\text{m}}\text{TcO}_4^-$ are given to a patient with suspected thyroid disease. Several hours or a day later, an image is taken with a γ -camera in order to show the distribution of the expression of the sodium/iodine symporter in the thyroid, or in the case of metastasis, in the whole body. Afterwards, the β -emitter ^{131}I may be used to eliminate undesired (*e.g.*, malignant) cells. Based on this technique, a method was proposed for treating malignant cells by first transfecting with the sodium/iodine symporter. The success of transfection and expression of the symporter was assessed by measuring the ability of tumour

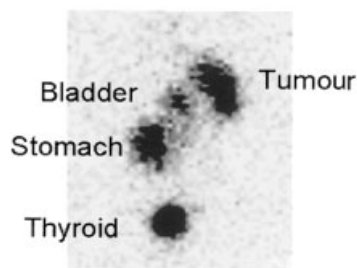
cells to take up radioisotopes ^{123}I , ^{125}I or ^{131}I (figure 6.24) *in vitro* (Spitzweg *et al.*, 1999; Haberkorn *et al.*, 2001, Spitzweg *et al.*, 2001a) and *in vivo* (Spitzweg *et al.*, 2001b). The uptake of radioactive iodine is directly correlated with the expression of the symporter. However, in non-thyroid cells, this method suffers from a lack of binding of the iodine to tyrosine structures (organification) resulting in fast externalisation of the isotopes. Boland *et al.* (2002) co-transfected the organification system (human thyroperoxidase), but, surprisingly, failed to observe a significant increase in uptake of radioiodine.

Labelled biomolecules such as antibodies or peptides can be used to target expressed proteins. Because of the high intrinsic sensitivity of scintigraphy, detection of intra and extracellular receptors is feasible despite the relatively low concentration of these proteins *in vivo* (picomolar to nanomolar range). This sensitivity is currently impossible using magnetic resonance imaging (MRI).

The radioiodine-labelled compound FIAU (2'-deoxy-5-iodo-1- β -D-arabinofuranosyluracil), an agonist of FMAU (2'-fluoro-5-methyl-1- β -D-arabinofuranosyluracil), is used as a marker of HSV1-TK (*Herpes simplex* virus type 1 thymidine kinase) expression. Mice bearing subcutaneous transfected cells were injected with radiolabelled FIAU and the stability of the expression *in vivo* was verified by imaging with a γ -camera (Blasberg *et al.* 1999; Haubner *et al.*, 2000; Tjuvajev *et al.*, 1995; Wiebe *et al.*, 1999).

Buchsbaum *et al.* (1999) performed *in vivo* transfection of human type 2 somatostatin receptor (*sstr2*) into non-small cell lung cancer tumour xenografts in mice. Using $^{99\text{m}}\text{Tc}$ -P829, a $^{99\text{m}}\text{Tc}$ -labelled synthetic radiopharmaceutical oligopeptide and ligand for *sstr2*, they demonstrated the feasibility of a therapeutic approach using ^{188}Re -P829, a ^{188}Re -labelled synthetic radiopharmaceutical oligopeptide ligand for *sstr2* (Buchsbaum *et al.*, 1999; Vallabhajosula *et al.*, 1996; Zinn *et al.*, 2000). Mice bearing a *sstr2*-negative tumour on both flanks received by direct injection into the tumour an adenoviral vector encoding *sstr2* on one side and a control vector on the other side. The expression of *sstr2* was detected after 48 hours by the injection of the $^{99\text{m}}\text{Tc}$ -labelled *sstr2* agonist followed by imaging three hours later.

Figure 6.24 Example of a ^{123}I iodine single photon emission planar scan. Shown is a scan of a mouse transfected with a construct expressing the sodium/iodide symporter protein (NP-1) and P-1 prostate cancer xenografts as sodium/iodide symporter-negative control tumours. The image was generated 24 hours after intrapulmonary injection of 500 microcurie of ^{123}I iodine. Whereas the sodium/iodide symporter-expressing NP-1 tumours trapped about 25 to 30 percent of the total radioiodine administered (right), sodium/iodide symporter-negative control tumour P-1 did not show radioiodine uptake (left). ^{123}I also accumulated physiologically in the bladder, stomach, and thyroid gland. This figure is republished from Spitzweg *et al.* (2000) with permission of the American Association for Cancer Research.



The use of labelled antibodies allows detection of most membrane proteins. Radiolabelled antibodies against cluster of differentiation 20 (CD20) are being used to evaluate the expression of the CD20 protein for diagnostic purposes and to target the cells for therapeutic applications (Press *et al.*, 2000).

Another interesting approach is the *in vivo* evaluation of antibodies against the HER2/neu oncogene (also called V-ERB-B2 avian erythroblastic leukemia viral oncogene homologue 2, ERBB2, or c-erbB2). The HER2/neu oncogene encodes a 185 kilodalton transmembrane phosphorylated glycoprotein of the epidermal growth factor (EGF) receptor family. It is expressed in about 30 percent of breast cancer patients and is indicative of a poor prognosis. Herceptin[®], an antibody against this oncogene, is clinically used for treatment alone or in combination with chemotherapy. Uptake of the radiolabelled antibody in breast cancer and heart tissue was shown to correlate directly to expression of the HER2/neu oncogene and to predict positive effects of the unlabelled Herceptin[®] antibody on therapeutic efficacy or a negative influence of the antibody in terms of myocardial complications (Behr *et al.*, 2001).

6.7.5

Conclusions

Scintigraphy has a high sensitivity allowing detection of low concentrations of probes at given sites. In comparison to the nuclides used in positron emission tomography, the nuclides used in scintigraphy generally have longer half-lives and allow the bio-distribution of a labelled biomolecule to be followed for a longer time. However, apart from the iodine, most of these isotopes are metallic and tend to alter the properties of the biomolecules they label. This problem also arises with other imaging modalities such as magnetic resonance imaging and, to a lesser extent, positron emission tomography. The relative lack of morphological information may be eliminated by a combination of scintigraphy with magnetic resonance imaging or computer tomography (called 'image fusion'). However, only centres that are specialised in this field have the necessary equipment for these experiments at the present time.

The detection of RNA and protein expression with radionuclides is a growing field and even today allows the detection of gene expression.

6.8

Optical *in vivo* imaging

6.8.1

Introduction

Light offers unique wavelength-dependent interactions with tissue and has been used to investigate both structural and functional tissue characteristics. While most traditional techniques depend on microscopy with limited depth penetration,

recent developments in 'optical' applications have made possible imaging of deeper tissues. Intravital confocal and multiphoton microscopy allows imaging to depths of about 400 micrometres. Most recently however, *in vivo* techniques have been developed that allow sensing up to several centimetres in depth (Ntziachristos *et al.*, 2000). Underpinning these developments are technologies that detect near-infrared fluorescence, a spectral window within which haemoglobin and water show minimal absorption (figure 6.25). In particular, new laser technology and highly sensitive charge coupled device (CCD) cameras have significantly increased the sensitivity of photon detection. Finally, novel reconstruction algorithms have been developed that allow tomographic reconstruction of optical images (diffuse optical tomography, DOT; fluorescence mediated tomography, FMT) and thus three-dimensional, quantitative visualisation of fluorochromes *in vivo* (Ntziachristos *et al.*, 1998 and 2000).

In parallel to the development of sensitive photon detection hardware, novel fluorescent and bioluminescent reporter probes have been developed that can sense molecular targets and interactions *in vivo*. Reporter genes with an optical signature (such as green fluorescent protein (GFP) or its variants) can be linked to regulatory elements that may reveal spatial and temporal information about a variety of biological processes at the transcriptional level (Contag *et al.*, 1995 and 1998). Luciferases and other bioluminescent proteins have also been used. Finally, a variety of novel near-infrared fluorescent optical reporter probes have been introduced that are suitable for intravenous injection and that can be activated by enzymes or that bind to specific targets (Becker *et al.*, 2001; Weissleder *et al.*, 1999). The combination of these modern *in vivo* optical imaging modalities now allows the sensitive and quantitative detection of optical molecular beacons *in vivo*.

6.8.2

Fluorescent proteins

Green fluorescent proteins and/or its spectrally shifted mutants can be linked to specific genes of interest or to therapeutic vectors that monitor the expression of the latter. GFP has been extensively used to track individual proteins and to study cellular dynamics (Chalfie *et al.*, 1994; Cole, *et al.*, 1996; Girotti & Banting, 1996; Prashar *et al.*, 1992; Presley *et al.*, 1997; Simonova *et al.*, 1999; Yokoe & Meyer, 1996).

Recently, GFP was used to monitor *in vivo* tumour cell growth and metastasis (Yang *et al.*, 2000). In addition, the role of vascular endothelial growth factor (VEGF) in tumour growth and wound healing was illustrated in a transgenic mouse strain that expresses GFP under the control of the VEGF promoter (Fukumura *et al.*, 1998). This model showed that most of the VEGF in tumours is produced by cells of the surrounding stroma rather than by the tumour cells themselves (Fukumura *et al.*, 1998).

While first results of GFP imaging are promising, one major drawback of this fluorochrome is the poor depth penetration since it fluoresces in the range of visible light (excitation at 489 nanometres, emission at 508 nanometres; figure 6.25),

which is highly absorbed by animal tissue. Recently, a red-shifted variant (*Discosoma* species red fluorescent protein; DsRed) of the protein was isolated from a tropical coral (*Anthozoa*) with an absorption maximum of 558 nanometres and emission at 583 nanometres (Matz *et al.*, 1999). Since red-shifted light penetrates the tissue more efficiently (figure 6.25), this protein is better suited for *in vivo* applications but still not ideal. Ideal proteins for *in vivo* applications would have their peak fluorescence at 700 to about 800 nanometres. However, first studies show the feasibility of monitoring protein localisation and gene expression in small organisms *in vivo* (Jakobs *et al.*, 2000).

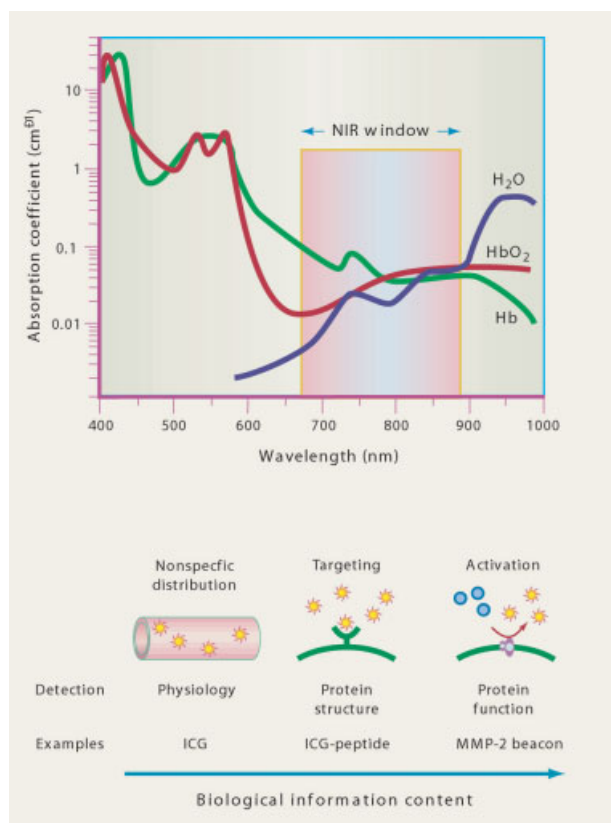


Figure 6.25 Optical imaging in the near-infrared window. Note the relatively low light attenuation by water (H₂O), haemoglobin (HbO₂) and deoxyhaemoglobin (Hb) allowing for maximum tissue penetration of the photons. *In vivo* imaging approaches in the near-infrared (NIR) include the use of non-specific dye (e.g., indo-

cyanine green, ICG), targeted probes or activatable, enzyme-sensing probes. Reprinted with permission from *Nature Biotechnology* (Weissleder, 2001). The following abbreviation is used in this figure: MMP-2, matrix metalloproteinase 2.

6.8.3

Bioluminescence

The oxidation of luciferin by luciferase results in the emission of photons, which can be detected by sensitive CCD technology. Different luciferase genes have been isolated from various organisms such as fireflies (*Photinus pyralis*), corals (*Renilla*), jellyfishes (*Aequorea*), and dinoflagellates (*Gonyaulax*). The luciferase of the North American firefly, for instance, emits green light on oxidation of luciferin. Luciferase can thus be used as an imaging marker gene acting like an internal sensor that emits light from within the body. Using sensitive bioluminescence imaging systems, tumour cells that express luciferase as a transgene can be monitored *in vivo* (Sweeney *et al.*, 1999). Tumour growth and regression of human cervical carcinoma (HeLa) cell lines transfected with luciferase was monitored *in vivo* using a sensitive intensified CCD camera (Edinger *et al.*, 1999; Sweeney *et al.*, 1999). This model thus allows for real-time, quantitative, spatial analysis of luciferase marker gene expression *in vivo*. Since luciferase is not found in mammalian organisms, this detection system operates with no background noise (Contag *et al.*, 1995, 1997 and 1998; Siragusa *et al.*, 1999; Zhang *et al.*, 1999). However, its use is likely to be confined to research applications since luciferin cannot be administered to humans.

6.8.4

NIRF – Near-infrared fluorescent imaging

Near-infrared fluorescent imaging (NIRF) is based on the excitation and detection of emitted near-infrared photons of intravenous-injectable fluorescent dyes (Ntzia-christos *et al.*, 2000), targeted probes (Becker *et al.*, 2001), enzyme activatable NIRF sensors (Weissleder *et al.*, 1999) or DNA/RNA sensing beacons (Tyagi *et al.*, 2000). Since in the near-infrared spectrum tissue absorption of light is minimal ('diagnostic window'), depth penetration required for *in vivo* applications is best achieved using light within this spectrum (figure 6.25).

6.8.4.1 Optical imaging of mRNA with molecular beacons

A molecular beacon is a dual-labelled oligonucleotide with a fluorescent reporter group at one end and a fluorescence quencher group at the other end. In the absence of a target, the molecule forms an internal hairpin that brings the reporter and quencher groups into physical proximity resulting in efficient quenching of the reporter (Tyagi *et al.*, 2000). In the presence of the complementary target sequence, the probe domain hybridises to the target resulting in disruption of the uni-molecular stem-loop structure. In this new two-molecule conformation, the fluorescent reporter group is no longer in physical proximity to the quencher group and the reporter can emit light at the wavelength characteristic of that fluorochrome (figure 6.26). Wavelength-shifting molecular beacons are substantially brighter than conventional molecular beacons that contain a fluorochrome that cannot efficiently absorb energy from the available monochromatic light source.

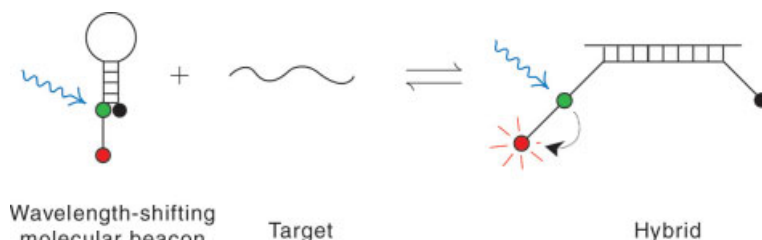


Figure 6.26 Molecular beacons for optical imaging of mRNA. A harvester fluorochrome that absorbs in the wavelength range of the monochromatic light source, an emitter fluorochrome of the desired emission colour, and a non-fluorescent quencher. In the absence of complementary nucleic acid targets, the probes are dark, whereas in the presence of targets, they fluoresce not in the emission range of the harvester fluorochrome that ab-

sorbs the light, but rather in the emission range of the emitter fluorophore. This shift in emission spectrum is due to the transfer of the absorbed energy from the harvester fluorochrome to the emitter fluorochrome by fluorescence resonance energy transfer, and takes place only in probes that are bound to targets. Reprinted with permission from *Nature Biotechnology* (Tyagi *et al.*, 2000).

In vivo use of the molecular beacons is challenging because of delivery barriers into cells and degradation. Enzyme-sensing probes however, also based on the fluorescence resonance energy transfer (FRET) principle, have been used successfully for imaging a variety of biological abnormalities (see below).

6.8.5

Enzyme-sensing, activatable NIRF probes

Recently, the concept of protease-sensing NIRF contrast agents has been introduced (Weissleder *et al.*, 1999). This technique is based on a quenching/dequenching mechanism of biocompatible near-infrared fluorochromes. These fluorochromes are mounted on a macromolecular carrier through peptide spacers and in close proximity to each other so that quenching occurs (Weissleder *et al.*, 1999). Only after enzymatic cleavage, the fluorochromes can be released with generation of a strong fluorescence signal that can be detected by a sensitive CCD camera (process of dequenching; see figure 6.27). The first generation of these probes

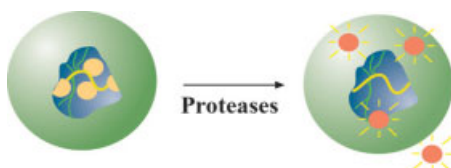
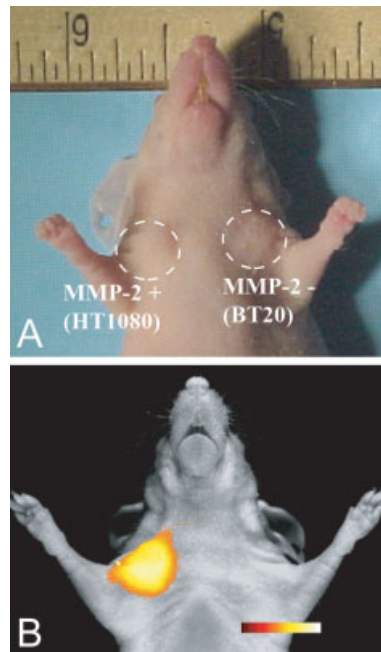


Figure 6.27 Protease-sensing optical probes. Multiple fluorochromes are bound to a long circulating macromolecular carrier resulting in mutual quenching of the fluorescence (left). After proteolytic cleavage of the probe and release of the fluorochromes a strong (up to

several 100-fold) increase in fluorescence can be detected (right). Enzyme activity can be sensitively imaged *in vivo* using this approach. Reprinted with submission from *Nature Biotechnology* (Weissleder *et al.*, 1999).

is activated by lysosomal enzymes such as cathepsin B (Weissleder *et al.*, 1999). *In vivo* these smart optical contrast agents allowed sensitive detection of small cell carcinoma (LX-1) xenografts in the submillimetre range in the nude mouse (Weissleder *et al.*, 1999). Moreover, the cathepsin B-sensitive probe was efficiently activated by colonic and intestinal polyps in a mouse modelling the human adenomatosis polyposis coli disease and clearly identified adenomas as little of as 50 micrometres across (Marten *et al.*, 2002). In an experimental breast cancer model, enzyme expression levels measured by Western blotting correlated closely to the NIRF signal observed *in vivo* suggesting that this technique may help to differentiate grades of tumour invasiveness (Bremer *et al.*, 2002). In order to investigate other key enzymes, the probe design was modified, so that cathepsin D and matrix metalloproteinases (MMPs) could be imaged with high selectivity (Bremer *et al.*, 2002; Tung *et al.*, 1999 and 2000). Tumours over-expressing matrix metalloproteinase 2 (HT1080 fibrosarcomas) were easily differentiated from BT20 adenocarcinomas, which revealed low expression levels of the enzyme (figure 6.28) (Bremer *et al.*, 2001a). Moreover, therapeutic inhibition of matrix metalloproteinase 2 was monitored using this approach (Bremer *et al.*, 2001b). Enzyme-sensing probes have now been synthesised for a variety of other proteases including caspase 1, caspase 3, thrombin, cathepsin K, *Herpes simplex* virus protease, human immunodeficiency virus (HIV) protease, prostate-specific antigen, nuclear factor κ B (NF κ B) among others (C. Tung, personal communication).

Figure 6.28 Near-infrared fluorescent (NIRF) image of a nude mouse. Mouse co-implanted with a matrix metalloproteinase 2-positive human fibrosarcoma (HT1080 tumour; left) and a matrix metalloproteinase 2 (MMP-2)-negative mammary adenocarcinoma (BT-20 tumour; right) as seen by reverse transcription-based polymerase chain reaction and Western blotting. Both tumours measured approximately three millimetres **A**). While the fibrosarcoma generated a strong fluorescent signal after intravenous injection of the matrix metalloproteinase 2-sensitive probe, the NIRF signal of the BT-20 tumour was only slightly higher than the background fluorescence of the skin **B**) colour-encoded fluorescence map superimposed on white light image. Reprinted with submission from *Radiology* (Bremer *et al.*, 2001).



6.8.5.1 Receptor-targeted NIRF-probes

Other strategies for achieving molecular specificity in optical imaging have been proposed by using receptor-targeted fluorochromes. Recently, a peptide/dye-conjugate consisting of a cyanine dye and the somatostatin analogue octreotate was synthesised serving as a contrast agent for optical tumour imaging (Becker *et al.*, 2001; Licha *et al.*, 2001). The probe showed an accumulation in mouse xenografts detected by an *in vivo* NIRF reflectance imager with a three-fold higher tumour fluorescence compared to non-diseased tissue (Becker *et al.*, 2001). The targeting conjugate was also specifically internalised by primary human neuroendocrine tumour cells (Becker *et al.*, 2001). While showing strong target specificity, this approach operates with a lower signal-to-background ratio, since non-bound probes are fully fluorescent and thus contribute to background noise.

6.9 Summary

The previous chapters of this book have been devoted to more general methods for analysing gene expression. The present chapter describes methods for analysing gene expression in tissue sections, in living cells and in the whole organism. While some of these methods may not be immediately obvious to the classical molecular biologist, their potential should not be underestimated. Physical-based methods such as magnetic resonance imaging are noninvasive and do not require labelled probes. Positron emission tomography requires use of radioactive tracers, but images metabolic processes in real-time. Optical methods allow real-time imaging of metabolic processes in living organisms without the use of radioactivity.

In our view, the real power of the methods described in this chapter will be unleashed when they are combined with the more standard methods described in earlier chapters. For example, some of these methods might be used to monitor the real-time activity of a metabolic pathway containing a gene of interest identified in a standard microarray-based experiment.

The ultimate aim of all expression analysis is to understand how gene expression functions in the living cell – how the machine works, as it were. The methods described in this chapter will make an important contribution to this task.

6.10

References

Internet resources

Integrated Spectral Data Base System for Organic Compounds (National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan) (Hayamizu, 2001); available at <http://www.aist.go.jp/RIODB/SDBS/menu-e.html>.

"Let's play PET!" is an educational website (also available as CD-ROM) illustrating positron emission tomography; available at <http://www.crump.ucla.edu/software/lpp/>.

Molecular imaging central (Crump Institute for Molecular Imaging, University of California – Los Angeles, UCLA; Los Angeles, California, USA); available at <http://www.micentral.org/>.

spectroscopyNOW.com by John Wiley & Sons Ltd. (Chichester, United Kingdom); available at <http://www.spectroscopynow.com/Spy/basehtml/SpyH/>.

The basics of NMR by Joseph P. Hornak (Magnetic Resonance Laboratory, Center for Imaging Science, Rochester Institute of Technology, Rochester, New York, USA); available at <http://www.cis.rit.edu/htbooks/nmr/>.

The basics of MRI by Joseph P. Hornak (Magnetic Resonance Laboratory, Center for Imaging Science, Rochester Institute of Technology, Rochester, New York, USA); available at <http://www.cis.rit.edu/htbooks/mri/>.

Society of Nuclear Medicine (SNM; Reston, Virginia, USA); accessible at <http://www.snm.org/>.

Cited publications

Agrawal, S., Jiang, Z., Zhao, Q., Shaw, D., Cai, Q., Roskey, A., Channavajjala, L., Saxinger, C., and Zhang, R. Mixed-backbone oligonucleotides as second generation antisense oligonucleotides: *in vitro* and *in vivo* studies. *Proceedings of the National Academy of Sciences USA* **1997**; *94*(6):2620-2625.

Ahrens, E. T., Rothbacher, U., Jacobs, R. E., and Fraser, S. E. A model for MRI contrast enhancement using T1 agents. *Proceedings of the National Academy of Sciences of the USA* **1998**; *95*(15):8443-8448.

Alauddin, M. M., Shahinian, A., Gordon, E. M., Bading, J. R., and Conti, P. S. Preclinical evaluation of the penciclovir analog 9-(4-[(18F]fluoro-3-hydroxymethylbutyl)guanine for *in vivo* measurement of suicide gene expression with PET. *Journal of Nuclear Medicine* **2001**; *42*(11):1682-1690.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. Molecular biology of the cell. Garland Publishing, Inc., New York, New York, USA **1995**.

Andersson, H., Baechli, T., Hoechl, M., and Richter, C. Autofluorescence of living cells. *Journal of Microscopy* **1998**; *191*(1):1-7.

Anger, H. O. Scintillation camera. *Review of Scientific Instruments* **1958**; *29*(1):27-33.

Auricchio, A., Zhou, R., Wilson, J. M., and Glickson, J. D. *In vivo* detection of gene expression in liver by ³¹P nuclear magnetic resonance employing creatine kinase as a marker gene. *Proceedings of the National Academy of Sciences of the USA* **2001**; *98*(9):5205-5210.

Bastiaens, P. I. H., and Squire, A. Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. *Trends in Cell Biology* **1999**; *9*(2):48-56.

Baumgart, E., Schad, A., Volki, A., and Fahimi, H. D. Detection of mRNAs encoding peroxisomal proteins by non-radioactive *in situ* hybridisation with digoxigenin-labelled cRNAs. *Histochemistry and Cell Biology* **1997**; *108*(4-5):371-379.

Beach, D. L., Salmon, E. D., Bloom, K. Localisation and anchoring of mRNA in budding yeast. *Current Biology* **1999**; *9*(11):569-578.

Beach, D. L., Bloom, K. *ASH1* mRNA Localisation in three acts. *Molecular Biology of the Cell* **2001**; *12*(9):2567-2577.

Becquerel, A. H. On radioactivity, a new property of matter. Nobel Lecture, December 11, 1903. In: Nobel lectures in physics, 1901-1995. Elsevier Publishing Company, Amsterdam, The Netherlands **1999**.

Becker, A., Hessenius, C., Licha, K., Ebert, B., Sukowski, U., Semmler, W., Wiedenmann, B., and Grotzinger, C. Receptor-targeted optical imaging of tumors with near-infrared fluorescent ligands. *Nature Biotechnology* **2001**; *19*(4):327-31.

- Behr, T. M., Dunn, R. M., Schmidberger, H., Gratz, S., Goldenberg, D. M., and Becker, W. Dosimetry in clinical radioimmunotherapy: practical approaches to reliable dose estimates, biological effects in relation to absorbed doses. In: *Cancer radioimmunotherapy – Present and future*. Riva, P., editor. Harwood Academic Publishers, Amsterdam, The Netherlands **1999**:147-190.
- Behr, T. M., Béhé, M., and Wörmann, B. Trastuzumab and breast cancer. *New England Journal of Medicine* **2001**; 345(13):995-996.
- Bell, J. D., and Taylor-Robinson, S. D. Assessing gene expression *in vivo*: magnetic resonance imaging and spectroscopy. *Gene Therapy* **2000**; 7(15):1259-1264.
- Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., and Long, R. M. Localisation and anchoring of ASH1 mRNA particles in living yeast. *Molecular Cell* **1998**; 2(4):437-445.
- Blasberg, R. G., and Tjuvajev, J. G. *Herpes simplex virus thymidine kinase as a marker/reporter gene for PET imaging of gene therapy*. *Quarterly Journal of Nuclear Medicine* **1999**; 43(2):163-169.
- Bloch, F. Nuclear induction. *Physical Review* **1946**; 70(7-8):460-474.
- Bloch, G., Jouvenal, L., and Carlier, P. G. ^1H NMR determination of lactate ^{13}C -enrichment in skeletal muscle: using a double quantum filter for the simultaneous editing of ^{13}C -coupled and ^{13}C -uncoupled methyl protons resonances. *Magnetic Resonance in Medicine* **1995**; 34(3):353-358.
- Bodanov, A., Matuszewski, L., Bremer, C., Petrovsky, A., and Weissleder, R. Oligomerization of paramagnetic substrates results in signal amplification and can be used for MR imaging of molecular targets. *Molecular Imaging* **2002**; 1(1):16-23.
- Boesch, C. Molecular aspects of magnetic resonance imaging and spectroscopy. *Molecular Aspects of Medicine* **1999**; 20(4-5):185-318.
- Boland, A., Magnon, C., Filetti, S., Bidart, J. M., Schlumberger, M., Yeh, P., and Perricaudet, M. Transposition of the thyroid iodide uptake and organification system in nonthyroid tumor cells by adenoviral vector-mediated gene transfers. *Thyroid* **2002**; 12(1):19-26.
- Bremer, C., Tung, C. H., and Weissleder, R. *In vivo* molecular target assessment of matrix metalloproteinase inhibition. *Nature Medicine* **2001a**; 7(6):743-748.
- Bremer, C., Bredow, S., Mahmood, U., Weissleder, R., and Tung, C. H. Optical imaging of matrix metalloproteinase-2 activity in tumors: feasibility study in a mouse model. *Radiology* **2001b**; 221(2):523-529.
- Bremer, C., Tung, C. H., Bogdanov, A., and Weissleder, R. Imaging of differential protease expression in breast cancers for detection of aggressive tumor phenotypes. *Radiology* **2002**; 222(3):814-818.
- Buchsbaum, D. J., Rogers, B. E., Khazaeli, M. B., Mayo, M. S., Milenic, D. E., Kashmiri, S. V., Anderson, C. J., Chappell, L. L., Brechbiel, M. W., and Curiel, D. T. Targeting strategies for cancer radiotherapy. *Clinical Cancer Research* **1999**; 5(10 Supplement):3048s-3055s.
- Carmo-Fonseca, M., Pepperkok, Sproat, B. S., Ansorge, W., Swanson, M. S., and Lamond, A. I. *In vivo* detection of snRNA-rich organelles in the nuclei of mammalian cells. *EMBO Journal* **1991**; 10(7):1863-1873.
- Casey, J. L., Koeller, D. M., Ramin, V. C., Klausner, R. D., and Harford, J. B. Iron regulation of transferrin receptor mRNA levels requires iron-responsive elements and a rapid turnover determinant in the 3' untranslated region of the mRNA. *EMBO Journal* **1989**; 8(12):3693-3699.
- Casey, J. L., Hentze, M. W., Koeller, D. M., Rouault, T. A., Harford, J. B., Klausner, R. D. Iron-responsive elements: regulatory RNA sequences that control mRNA levels and translation. *Science* **1988**; 240(4954):924-928.
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **1994**; 263(5148):802-805.
- Chatziioannou, A. F. Molecular imaging of small animals with dedicated PET tomographs. *European Journal of Nuclear Medicine and Molecular Imaging* **2002**; 29(1):98-114.
- Cole, N. B., Smith, C. L., Sciaky, N., Terasaki, M., Edidin, M., and Lippincott-Schwartz, J. Diffusional mobility of Golgi proteins in membranes of living cells. *Science* **1996**; 273(5276):797-801.
- Contag, C. H., Contag, P. R., Mullins, J. I., Spilman, S. D., Stevenson, D. K., and Benaron, D. A. Photonic detection of bacterial pathogens in living hosts. *Molecular Microbiology* **1995**; 18(4):593-603.
- Contag, C. H., Spilman, S. D., Contag, P. R., Oshiro, M., Eames, B., Dennery, P., Steven-

- son, D. K., and Benaron, D. A. Visualizing gene expression in living mammals using a bioluminescent reporter. *Photochemistry and Photobiology* **1997**; 66(4):523-531.
- Contag, P. R., Olomu, I. N., Stevenson, D. K., and Contag, C. H. Bioluminescent indicators in living mammals. *Nature Medicine* **1998**; 4(2):245-247.
- Cotten, M., Wagner, E., and Birnstiel, M. L. Receptor-mediated transport of DNA into eukaryotic cells. *Methods in Enzymology* **1993**; 217:618-644.
- Dewanjee, M. K., Ghafouripour, A. K., Werner, R. K., Serafini, A. N., and Sfakianakis, G. N. Development of sensitive radioiodinated antisense oligonucleotide probes by conjugation technique. *Bioconjugate Chemistry* **1991**; 2(4):195-200.
- Dewanjee, M. K., Ghafouripour, A. K., Kapadvanjwala, M., and Samy, A. T. Kinetics of hybridization of mRNA of *c-myc* oncogene with ¹¹¹In-labeled antisense oligodeoxynucleotide probes by high-pressure liquid chromatography. *Biotechniques* **1994**; 16(5):844-850.
- Dewanjee, M. K., Ghafouripour, A. K., Kapadvanjwala, M., Dewanjee, S., Serafini, A. N., Lopez, D. M., and Sfakianakis, G. N. Noninvasive imaging of *c-myc* oncogene messenger RNA with indium-111-antisense probes in a mammary tumor-bearing mouse model. *Journal of Nuclear Medicine* **1994**; 35(6):1054-1063.
- Dirks, R. W. RNA Molecules lighting up under the microscope. *Histochemistry and Cell Biology* **1996**; 106(2):151-166.
- Dirks, R. W., Molenaar, C., and Tanke, H. J. Methods for visualising RNA processing and transport pathways in living cells. *Histochemistry and Cell Biology* **2001**; 115(1):3-11.
- Dixon, W. T. Simple proton spectroscopic imaging. *Radiology* **1984**; 153(1):189-194.
- Dowty, M. E., Williams, P., Zhang, G., Hagstrom, J. E., and Wolff, J. A. Plasmid DNA entry into postmitotic nuclei of primary rat myotubes. *Proceedings of the National Academy of Sciences of the USA* **1995**; 92(10):4572-4576.
- Edinger, M., Sweeney, T. J., Tucker, A. A., Olomu, A. B., Negrin, R. S., and Contag, C. H. Noninvasive assessment of tumor cell proliferation in animal models. *Neoplasia* **1999**; 1(4):303-310.
- Enochs, W. S., Hyslop, W. B., Bennett, H. F., Brown, R. D., 3rd, Koenig, S. H., Swartz, and H. M. Sources of the increased longitudinal relaxation rates observed in melanotic melanoma. An *in vitro* study of synthetic melanins. *Investigative Radiology* **1989**; 24(10):794-804.
- Enochs, W. S., Petherick, P., Bogdanova, A., Mohr, U., and Weissleder, R. Paramagnetic metal scavenging by melanin: MR imaging. *Radiology* **1997**; 204(2):417-423.
- Enochs, S., and Weissleder, R. Organ and tissue directed MR contrast agents. In: Clinical magnetic resonance imaging. Edelman, R. R., Hesselink, J. R., and Zlatkin, M., editors. WB Saunders, Philadelphia, Pennsylvania, USA **1996**:192-220.
- Evans, R. D. Early history (1936-1946) of nuclear medicine in thyroid studies at Massachusetts General Hospital. *Medical Physics* **1975**; 2(3):105-109.
- Feld, M., and De Roo, M. Geschichte der Nuklearmedizin in Europa. Schicha, H., and Bergdolt, M., editors. Schattauer Verlag, Stuttgart, Germany **2000**.
- Fisher, T. L., Terhorst, T., Cao, X., and Wagner, R. W. Intracellular disposition and metabolism of fluorescently-labeled unmodified and modified oligonucleotides microinjected into mammalian cells. *Nucleic Acids Research* **1993**; 21(16):3857-3865.
- Fukumura, D., Xavier, R., Sugiura, T., Chen, Y., Park, E. C., Lu, N., Selig, M., Nielsen, G., Taksir, T., Jain, R. K., and Seed, B. Tumor induction of VEGF promoter activity in stromal cells. *Cell* **1998**; 94(6):715-725.
- Gadian, D. G. Nuclear magnetic resonance and its application to living systems. Clarendon Press, Oxford, United Kingdom **1982**.
- Gambhir, S. S., Barrio, J. R., Phelps, M. E., Iyer, M., Namavari, M., Satyamurthy, N., Wu, L., Green, L. A., Bauer, E., MacLaren, D. C., Nguyen, K., Berk, A. J., Cherry, S. R., and Herschman, H. R. Imaging adenoviral-directed reporter gene expression in living animals with positron emission tomography. *Proceedings of the National Academy of Sciences of the USA* **1999a**; 96(5):2333-2338.
- Gambhir, S. S., Barrio, J. R., Herschman, H. R., and Phelps, M. E. Imaging expression: principles and assays. *Journal of Nuclear Cardiology* **1999b**; 6(2):219-233.
- Gerfen, C. R. Quantification of *in situ* hybridisation histochemistry for analysis of brain function. In: Methods in neurosciences. Conn, P. M., editor. Academic Press, San Diego, California, USA **1989**; 1(*Gene probes*).

- Girotti, M., and Banting, G. TGN38-green fluorescent protein hybrid proteins expressed in stably transfected eukaryotic cells provide a tool for the real-time, *in vivo* study of membrane traffic pathways and suggest a possible role for ratTGN38. *Journal of Cell Science* **1996**; 109(12):2915-2926.
- Green, M. V., Seidel, J., Vaquero, J. J., Jagoda, E., Lee, I., and Eckelman, W. C. High resolution PET, SPECT and projection imaging in small animals. *Computerized Medical Imaging and Graphics* **2001**; 25(2):79-86.
- Grogan, T. M., Rangel, C., Rimsza, L., Bellamy, W., Martel, R., McGraw, B., Richards, W. Kinetic-Mode, Automated double-labeled immunohistochemistry and *in-situ* hybridisation in diagnostic pathology. In: *Advances in pathology and laboratory medicine*. St. Louis, Missouri, USA **1995**; 8:86-96.
- Haberkorn, U., Henze, M., Altmann, A., Jiang, S., Morr, I., Mahmut, M., Peschke, P., Kubler, W., Debus, J., and Eisenhut, M. Transfer of the human NaI symporter gene enhances iodide uptake in hepatoma cells. *Journal of Nuclear Medicine* **2001**; 42(2):317-325.
- Haberkorn, U., Altmann, A., and Eisenhut, M. Functional genomics and proteomics – the role of nuclear medicine. *European Journal of Nuclear Medicine* **2002**; 29(1):115-132.
- Hakumäki, J. M., Poptani, H., Sandmair, A.-M., Ylä-Herttuala, S., and Kauppinen, R. A. ¹H MRS detects polyunsaturated fatty acid accumulation during gene therapy of glioma: Implications for the *in vivo* detection of apoptosis. *Nature Medicine* **1999**; 5(11):1323-1327.
- Haubner, R., Avril, N., Hantzopoulos, P. A., Gansbacher, B., and Schwaiger, M. *In vivo* imaging of *Herpes simplex* virus type 1 thymidine kinase gene expression: early kinetics of radiolabelled FIAU. *European Journal of Nuclear Medicine* **2000**; 27(3):283-291.
- Hayamizu, K. An input tool by a personal computer for the NMR spectral database (SDBS-NMR). *Journal of Computer Aided Chemistry* **2001**; 2:1-10.
- Hertz, S., Roberts, A., and Evans, R. D. Radioiodine as indicator in study of thyroid physiology. *Proceedings of the Society of Experimental Biology and Medicine* **1938**; 38(4):510-513.
- Hnatowich, D. J., Winnard, P., Jr., Virzi, F., Fogarasi, M., Sano, T., Smith, C. L., Cantor, C. R., and Rusckowski, M. Technetium-99m labeling of DNA oligonucleotides. *Journal of Nuclear Medicine* **1995**; 36(12):2306-2314.
- Högemann, D., Josephson, L., Weissleder, R., and Basilion, J. P. Improvement of MRI probes to allow efficient detection of gene expression. *Bioconjugate Chemistry* **2000**; 11(6):941-946.
- Högemann, D., Ntziachristos, V., Josephson, L., and Weissleder, R. High throughput magnetic resonance imaging for evaluating targeted nanoparticle probes. *Bioconjugate Chemistry* **2002**; 13(1):116-121.
- Hospers, G. A., Calogero, A., van Waarde, A., Doze, P., Vaalburg, W., Mulder, N. H., and de Vries, E. F. Monitoring of *Herpes simplex* virus thymidine kinase enzyme activity using positron emission tomography. *Cancer Research* **2000**; 60(6):1488-1491.
- Hougaard, D. M., Hansen, H., and Larsson, L. Non-radioactive *in situ* hybridisation for mRNA with emphasis on the use of oligodeoxynucleotide probes. *Histochemistry and Cell Biology* **1997**; 108(4-5):335-344.
- Isiklar, I., Leeds, N. E., Fuller, G. N., and Kumar, A. J. Intracranial metastatic melanoma: correlation between MR imaging characteristics and melanin content. *American Journal of Roentgenology* **1995**; 165(6):1503-1512.
- Jackson, D. A., Hassan, A. B., Errington, R. J., and Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO Journal* **1993**; 12(3):1059-1065.
- Jakobs, S., Subramaniam, V., Schonle, A., Jovin, T. M., and Hell, S. W. EFEG and DsRed expressing cultures of *Escherichia coli* imaged by confocal, two-photon and fluorescence lifetime microscopy. *FEBS Letters* **2000**; 479(3):131-135.
- Johnson, G. A., Benveniste, H., Black, R. D., Hedlund, L. W., Maronpot, R. R., and Smith, B. R. Histology by magnetic resonance microscopy. *Magnetic Resonance Quarterly* **1993**; 9(1):1-30.
- Josephson, L., Perez, M., and Weissleder, R. Magnetic nanosensor for the detection of oligonucleotide sequences. *Angewandte Chemie International Edition* **2001**; 40(17):3204-3206.
- Kairemo, K. J. Positron emission tomography of monoclonal antibodies. *Acta Oncologica* **1993**; 32(7-8):825-30.
- Kayyem, J. F., Kumar, R. M., Fraser, S. E., and Meade, T. J. Receptor-targeted co-transport of

- DNA and magnetic resonance contrast agents. *Chemistry and Biology* **1995**; 2(9): 615-620.
- Kind, C. N., Pearson, E. C., Evans, J. G., and Hammond, T. G. Expression of HSP 72kD in cingulate neurones of the rat following acute exposure to MK801. *Human Experimental Toxicology* **1996**; 15:657.
- Komminoth, P. Digoxigenin as an alternative probe labeling for *in situ* hybridisation. *Diagnostic Molecular Pathology* **1992**; 1(2): 142-150.
- Komminoth, P., and Werner, M. Target and signal amplification: approaches to increase the sensitivity of *in situ* hybridisation. *Histochemistry and Cell Biology* **1997**; 108(4-5): 325-333.
- Koretsky, A. P., Brosnan, M. J., Chen, L., Chen, J., and van Dyke, T. NMR detection of creatine kinase expressed in liver of transgenic mice: Determination of free ADP levels. *Proceedings of the National Academy of Sciences of the USA* **1990**; 87(8):3112-3116.
- Koretsky, A., Lin, Y., Schorle, H., and Jaenisch, R. Genetic control of MRI contrast by expression of the transferrin receptor. Presented at the 4th Scientific Meeting of the Society of Magnetic Resonance, New York, New York, USA, April 27th to May 3rd, **1996**.
- Krebs, B., Rauchenberger, R., Reiffert, S., Rothe, C., Tesar, M., Thomassen, E., and Cao, M. High-throughput generation and engineering of recombinant human antibodies. *Journal of Immunological Methods* **2001**; 254(1-2):67-84.
- Kuhl, D. E., and Edwards, R. Q. Image separation radioisotope scanning. *Radiology* **1963**; 80(4):653-661.
- Kumar, V. Immunofluorescence and enzyme immunomicroscopy methods. *Journal of Immunoassay* **2000**; 21(2-3):235-253.
- Kwon, B. S., Haq, A. K., Kim, G. S., Pomerantz, S. H., and Halaban, R. Cloning and characterization of a human tyrosinase cDNA. *Progress in Clinical and Biological Research* **1988**; 256:273-282.
- Lakowicz, J. R., Szmajcinski, H., Nowaczyk, K., Berndt, K. W., and Johnson, M. Fluorescence lifetime imaging. *Analytical Biochemistry* **1992**; 202(2):316-330.
- Leonetti, J. P., Mechti, N., Degols, G., Gagnor, C., and LeBleu, B. Intracellular distribution of microinjected antisense oligonucleotides. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(7):2702-2706.
- Licha, K., Hassenius, C., Becker, A., Henklein, P., Bauer, M., Wisniewski, S., Wiedenmann, B., Semmler, W. Synthesis, characterization, and biological properties of cyanine-labeled somatostatin analogues as receptor-targeted fluorescent probes. *Bioconjugate Chemistry* **2001**; 12(1):44-50.
- Louie, A. Y., Huber, M. M., Ahrens, E. T., Rothbacher, U., Moats, R., Jacobs, R. E., Fraser, S. E., and Meade, T. J. *In vivo* visualization of gene expression using magnetic resonance imaging. *Nature Biotechnology* **2000**; 18(3):321-325.
- Lukacs, G. L., Haggie, P., Seksek, O., Lechardeur, D., Freedman, N., and Verkman, A. S. Size-dependent DNA mobility in cytoplasm and nucleus. *Journal of Biological Chemistry* **2000**; 275(3):1625-1629.
- MacLaren, D. C., Gambhir, S. S., Satyamurthy, N., Barrio, J. R., Sharfstein, S., Toyokuni, T., Wu, L., Berk, A. J., Cherry, S. R., Phelps, M. E., and Herschman, H. R. Repetitive, non-invasive imaging of the dopamine D2 receptor as a reporter gene in living animals. *Gene Therapy* **1999**; 6(5):785-791.
- Marten, K., Bremer, C., Khazaie, K., Sameni, M., Sloane, B., Tung, C.-H., and Weissleder, R. Detection of dysplastic intestinal adenomas using enzyme sensing molecular beacons. *Gastroenterology* **2002**; 122(2):406-414.
- Matz, M. V., Fradkov, A. F., Labas, Y. A., Savitsky, A. P., Zaisky, A. G., Markelov, M. L., and Lukyanov, S. A. Fluorescent proteins from nonbioluminescent *Anthozoa* species. *Nature Biotechnology* **1999**; 17(10):969-973.
- Matsuo, T. *In situ* visualization of messenger RNA for basic fibroblast growth factor in living cells. *Biochimica et Biophysica Acta* **1998**; 1379(2):178-184.
- McNeil, P. L., and Warder, E. Glass beads load macromolecules into living cells. *Journal of Cellular Sciences* **1987**; 88(5):669-678.
- Meltzer, J. C., Sanders, V., Grimm, P. C., Stern, E., Rivier, C., Lee, S., Rennie, S. L., Gietz, R. D., Hole, A. K., Watson, P. H., Greenberg, A. H., and Nance, D. M. Production of digoxigenin-labelled RNA probes and the detection of cytokine mRNA in rat spleen and brain by *in situ* hybridisation. *Brain Research Protocols* **1998**; 2(4):339-351.
- Molenaar, C., Marras, S. A., Slats, J. C., Truffert, J. C., Lemaitre, M., Raap, A. K., Dirks, R. W., and Tanke, H. J. Linear 2'-O-Methyl RNA

- probes for the visualization of RNA in living cells. *Nucleic Acids Research* **2001**; 29(17):E89.
- Ntziachristos, V., Yodh, A. G., Schnall, M., and Chance, B. Concurrent MRI and diffuse optical tomography of breast after indocyanine green enhancement. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(6):2767-2772.
- Ntziachristos, V., Hielscher, A. H., Yodh, A. G., and Chance, B. Diffuse optical tomography of highly heterogeneous media. *IEEE Transactions on Medical Imaging* **2001**; 20(6):470-478.
- Okazaki, M., Kuwata, K., Miki, Y., Shiga, S., and Shiga, T. Electron spin relaxation of synthetic melanin and melanin-containing human tissues as studied by electron spin echo and electron spin resonance. *Archives of Biochemistry and Biophysics* **1985**; 242(1):197-205.
- Oliver, K. R., Heavens, R. P., and Sirinathsinghji, D. J. Quantitative comparison of pretreatment regimens used to sensitize *in situ* hybridisation using oligonucleotide probes on paraffin-embedded brain tissue. *Journal of Histochemistry and Cytochemistry* **1997**; 45(12):1707-1713.
- Phelps, M. E. PET: The merging of biology and imaging into molecular imaging. *Journal of Nuclear Medicine* **2000a**; 41(4):661-681.
- Phelps, M. E. Positron emission tomography provides molecular imaging of biological processes. *Proceedings of the National Academy of Science of the USA* **2000b**; 97(16):9226-9233.
- Politz, J. C., Browne, E. S., Wolf, D. E., and Pederson T. Intracellular diffusion and hybridisation state of oligonucleotides measured by fluorescence correlation spectroscopy. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(11):6043-6048.
- Prasher, D. C., Eckenrode, V. K., Ward, W. W., Prendergast, F. G., and Cormier, M. J. Primary structure of the *Aequorea victoria* green-fluorescent protein. *Gene* **1992**; 111(2):229-233.
- Presley, J. F., Cole, N. B., Schroer, T. A., Hirschberg, K., Zaal, K. J., and Lippincott-Schwartz, J. ER-to-Golgi transport visualized in living cells. *Nature* **1997**; 389(6646):818-815.
- Press, O. W., and Rasey, J. Principles of radio-immunotherapy for hematologists and oncologists. *Seminars in Oncology* **2000**; 27(6 Supplement 12):62-73.
- Price, P. PET as a potential tool for imaging molecular mechanisms of oncology in man. *Trends in Molecular Medicine* **2001**; 7(10):442-446.
- Pringle, H. J., Primrose, L., Kind, C. N., Talbot, I. C., and Lauder, I. *In-situ* hybridisation demonstration of poly-adenylated RNA sequences in formalin-fixed paraffin sections using a biotinylated oligonucleotide poly d(T) probe. *Journal of Pathology* **1989**; 158(4):279-286.
- Qian, X., Bauer, R. A., Xu, H. S., and Lloyd, R. V. *In situ* hybridisation detection of calcitonin mRNA in routinely fixed, paraffin-embedded tissue sections: a comparison of different types of probes combined with tyramide signal amplification. *Applied Immunohistochemistry and Molecular Morphology* **2001**; 9(1):61-69.
- Roman, B. B., Wieringa, B., and Koretsky, A. P. Functional equivalence of creatine kinase isoforms in mouse skeletal muscle. *Journal of Biological Chemistry* **1997**; 272(28):17790-17794.
- Rook, M. S., Lu, M., and Kosik, K. S. CaMKIIalpha 3' untranslated region-directed mRNA translocation in living neurons: visualization by GFP linkage. *Journal of Neuroscience* **2000**; 20(17):6385-93.
- Rothman, D. L., Petroff, O. A. C., Behar, K. L., and Mattson, R. H. Localised ¹H NMR measurements of gamma-aminobutyric acid in human brain *in vivo*. *Proceedings of the National Academy of Sciences of the USA* **1993**; 90(12):5662-5666.
- Sanchez-Ferrer, A., Rodriguez-Lopez, J. N., Garcia-Canovas, F., and Garcia-Carmona, F. Tyrosinase: a comprehensive review of its mechanism. *Biochimica et Biophysica Acta* **1995**; 1247(1):1-11.
- Scherr, M., Rossi, J. J., Sczakiel, G., and Patzel, V. RNA accessibility prediction: a theoretically approach is consistent with experimental studies in cell extracts. *Nucleic Acids Research* **2000**; 28(13):2455-2461.
- Sepponen, R. E., Sipponen, J. T., and Tantt, J. I. A method for chemical shift imaging: demonstration of bone marrow involvement with proton chemical shift imaging. *Journal of Computer Assisted Tomography* **1984**; 8(4):585-587.
- Sharp, F. R., Jasper, P., Hall, J., and Noble, L. MK-801 and ketamine induce heat shock protein HSP72 in injured neurons in poster-

- ior cingulate and retrosplenial cortex. *Annals of Neurology* **1991**; 30(6):801-809.
- Shields, A. F., Grierson, J. R., Dohmen, B. M., Machulla, H. J., Stayanoff, J. C., Lawhorn-Crews, J. M., Obradovich, J. E., Muzik, O., and Mangner, T. J. Imaging proliferation *in vivo* with [F-18]FLT and positron emission tomography. *Nature Medicine* **1998**; 4(11):1334-1336.
- Simonova, M., Weissleder, R., Sergeyev, N., Vilissova, N., and Bogdanov, A., Jr. Targeting of green fluorescent protein expression to the cell surface. *Biochemical and Biophysical Research Communication* **1999**; 262(3):638-642.
- Simonova, M., Wall, A., Weissleder, R., and Bogdanov, A., Jr. Tyrosinase mutants are capable of prodrug activation in transfected nonmelanotic cells. *Cancer Research* **2000**; 60(23):6656-6662.
- Sipkins, D. A., Cheresch, D. A., Kazemi, M. R., Nevin, L. M., Bednarski, M. D., and Li, K. C. Detection of tumor angiogenesis *in vivo* by alphaVbeta3-targeted magnetic resonance imaging. *Nature Medicine* **1998**; 4(5):623-626.
- Siragusa, G. R., Nawotka, K., Spilman, S. D., Contag, P. R., and Contag, C. H. Real-time monitoring of *Escherichia coli* O157:H7 adherence to beef carcass surface tissues with a bioluminescent reporter. *Applied and Environmental Microbiology* **1999**; 65(4):1738-1745.
- Smith, B. R., Johnson, G. A., Groman, E. V., and Linney, E. Magnetic resonance microscopy of mouse embryos. *Proceedings of the National Academy of Sciences of the USA* **1994**; 91(9):3530-3533.
- Sokol, D. L., Zhang, X., Lu, P., and Gewirtz, A. M. Real time detection of DNA. RNA hybridisation in living cells. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(20):11538-11543.
- Sokoloff, L., Reivich, M., Kennedy, C., Des Rosiers, M. H., Patlak, C. S., Pettigrew, K. D., Sakurada, O., and Shinohara, M. The [14C]deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat. *Journal of Neurochemistry* **1977**; 28(5):897-916.
- Sorenson, J. A., and Phelps, M. E. Physics in nuclear medicine. WB Saunders, Philadelphia, Pennsylvania, USA **1987**.
- Spitzweg, C., Zhang, S., Bergert, E. R., Castro, M. R., McIver, B., Heufelder, A. E., Tindall, D. J., Young, C. Y., and Morris, J. C. Prostate-specific antigen (PSA) promoter-driven androgen-inducible expression of sodium iodide symporter in prostate cancer cell lines. *Cancer Research* **1999**; 59(9):2136-2141.
- Spitzweg, C., and Morris, J. C. Approaches to gene therapy with sodium/iodide symporter. *Experimental and Clinical Endocrinology and Diabetes* **2001**; 109(1):56-59.
- Spitzweg, C., Dietz, A. B., O'Connor, M. K., Bergert, E. R., Tindall, D. J., Young, C. Y., and Morris, J. C. *In vivo* sodium iodide symporter gene therapy of prostate cancer. *Gene Therapy* **2001**; 8(20):1524-1531.
- Stark, D. D., Bradley, W. G. Magnetic resonance imaging. 3rd edition. Mosby Year Book, St. Louis, Missouri, USA **1999**.
- Stegman, L. D., Rehemtulla, A., Beattie, B., Kievit, E., Lawrence, T. S., Blasberg, R. G., Tjuvajev, J. G., and Ross, B. D. Noninvasive quantitation of cytosine deaminase transgene expression in human tumor xenografts with *in vivo* magnetic resonance spectroscopy. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(17):9821-9826.
- Stern, C. D. Detection of multiple gene products simultaneously by *in situ* hybridisation and immunohistochemistry in whole mounts of avian embryos. *Current Topics in Developmental Biology* **1998**; 36:222-243.
- Sundaresan, G., Yazaki, P. J., Shively, J. E., Toyokuni, T., Nguyen, K., Finn, R., Larson, S. M., Raubitschek, A. A., Gambhir, S. S., and Wu, A. M. I-124 radiolabeled genetically engineered anti-CEA antibody fragments for tumor imaging with microPET. *Journal of Nuclear Medicine* **2001**; 42(Supplement 1):42.
- Sweeney, T. J., Mailander, V., Tucker, A. A., Olomu, A. B., Zhang, W., Cao, Y. A., Negrin, R. S., and Contag, C. H. Visualizing the kinetics of tumor-cell clearance in living animals. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(21):12044-12049.
- Tavittian, B., Terrazzino, S., Kuhnast, B., Marzabal, S., Stettler, O., Dolle, F., Deverre, J. R., Jobert, A., Hinnen, F., Bendriem, B., Crouzel, C., and Di Giamberardino, L. *In vivo* imaging of oligonucleotides with positron emission tomography. *Nature Medicine* **1998**; 4(4):467-471.
- Ter-Pogossian, M. M., Phelps, M. E., Hoffman, E. J., and Mullani, N. A. A positron-emission transaxial tomograph for nuclear imaging (PETT). *Radiology* **1975**; 114(1):89-98.

- Thierry, A. R., and Dritschilo, A. Intracellular availability of unmodified, phosphorothioated and liposomally encapsulated oligodeoxynucleotides for antisense activity. *Nucleic Acids Research* **1992**; 20(21):5691-5698.
- Thorstensen, K., and Romslo, I. The transferin receptor: its diagnostic value and its potential as therapeutic target. *Scandinavian Journal of Clinical and Laboratory Investigation* **1993**; 215:113-120.
- Tjuvajev, J. G., Stockhammer, G., Desai, R., Uehara, H., Watanabe, K., Gansbacher, B., and Blasberg, R. G. Imaging the expression of transfected genes *in vivo*. *Cancer Research* **1995**; 55(24):6126-6132.
- Tung, C. H., Bredow, S., Mahmood, U., and Weissleder, R. Preparation of a cathepsin D sensitive near-infrared fluorescence probe for imaging. *Bioconjugate Chemistry* **1999**; 10(5):892-896.
- Tung, C., Mahmood, U., Bredow, S., and Weissleder, R. *In vivo* imaging of proteolytic activity using a novel molecular reporter. *Cancer Research* **2000**; 60(17):4953-4958.
- Tyagi, S., Marras, S. A., and Kramer, F. R. Wavelength-shifting molecular beacons. *Nature Biotechnology* **2000**; 18(11):1191-1196.
- Vallabhajosula, S., Moyer, B. R., Lister-Jones, J., McBride, B. J., Lipszyc, H., Lee, H., Bastidas, D., and Dean, R. T. Preclinical evaluation of technetium-99m-labeled somatostatin receptor-binding peptides. *Journal of Nuclear Medicine* **1996**; 37(6):1016-1022.
- Van der Loos, C. M. Immunoenzyme multiple staining methods. *Royal Microscopical Society microscopy handbooks*. BIOS Scientific Publishers Ltd. in association with the Royal Microscopical Society, Oxford, United Kingdom **1999**.
- Von Hevesy, G. K. Some applications of isotopic indicators. Nobel Lecture, December 12, 1944. In: Nobel lectures in chemistry, 1942-1962. Elsevier Publishing Company, Amsterdam, The Netherlands **1964**.
- Wang, J., Cao, L.-G., Wang, Y.-L., and Pederson, T. Localisation of pre-messenger RNA at discrete nuclear sites. *Proceedings of the National Academy of Sciences of the USA* **1991**; 88(16):7391-7395.
- Wansink, D. G., Schul, W., Van der Kraan, I., Van Steensel, B., Van Driel, R., and De Jong, L. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *Journal of Cellular Biology* **1993**; 122(2): 283-293.
- Walter, G., Barton, E. R., and Sweeny, H. L. Noninvasive measurement of gene expression in skeletal muscle. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(10):5151-5155.
- Weissleder, R., Simonova, M., Bogdanova, A., Bredow, S., Enochs, W. S., and Bogdanov, A., Jr. MR imaging and scintigraphy of gene expression through melanin induction. *Radiology* **1997**; 204(2):425-429.
- Weissleder, R., Tung, C. H., Mahmood, U., and Bogdanov, A., Jr. *In vivo* imaging of tumors with protease-activated near-infrared fluorescent probes. *Nature Biotechnology* **1999**; 17(4):375-378.
- Weissleder, R., Moore, A., Mahmood, U., Bhorade, R., Benveniste, H., Chiocca, E. A., and Basilion, J. P. *In vivo* magnetic resonance imaging of transgene expression. *Nature Medicine* **2000**; 6(3):351-355.
- Weissleder R. A clearer vision for *in vivo* imaging. *Nature Biotechnology* **2001**; 19(4):316-317.
- Weissleder, R., and Mahmood, U. Molecular imaging. *Radiology* **2001**; 219(2):316-333.
- Wiebe, L. I., Knaus, E. E., and Morin, K. W. Radiolabelled pyrimidine nucleosides to monitor the expression of HSV-1 thymidine kinase in gene therapy. *Nucleosides and Nucleotides* **1999**; 18(4-5):1065-1066.
- Wirrwar, A., Schramm, N., Vosberg, H., and Muller-Gartner, H. W. High resolution SPECT in small animal research. *Reviews in the Neurosciences* **2001**; 12(2):187-193.
- Yang, M., Baranov, E., Jiang, P., Sun, F. X., Li, X. M., Li, L., Hasegawa, S., Bouvet, M., Al-Tuwaijri, M., Chishima, T., Shimada, H., Moossa, A. R., Penman, S., and Hoffman, R. M. Whole-body optical imaging of green fluorescent protein-expressing tumors and metastases. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(3):1206-1211.
- Yokoe, H., and Meyer, T. Spatial dynamics of GFP-tagged proteins investigated by local fluorescence enhancement. *Nature Biotechnology* **1996**; 14(10):1252-1256.
- Yokomori, H., Oda, M., Yasogawa, Y., Nishi, Y., Ishii, H. Signal detection of endothelin receptor subtypes in human cirrhotic liver by a new *in situ* hybridisation method. *Medical Electron Microscopy* **2000**; 33:207-216.

- Yu, Y., Annala, A. J., Barrio, J. R., Toyokuni, T., Satyamurthy, N., Namavari, M., Cherry, S. R., Phelps, M. E., Herschman, H. R., and Gambhir, S. S. Quantification of target gene expression by imaging reporter gene expression in living animals. *Nature Medicine* **2000**; 6(8):933-937.
- Zaidi, A. U., Enomoto, H., Milbrandt, J., and Roth, K. A. Dual fluorescent *in situ* hybridization and immunohistochemical detection with tyramide signal amplification. *Journal of Histochemistry and Cytochemistry* **2000**; 48(10):1369-1375.
- Zhang, W., Contag, P. R., Madan, A., Stevenson, D. K., and Contag, C. H. Bioluminescence for biological sensing in living mammals. *Advances in Experimental Medicine and Biology* **1999**; 471:775-784.
- Zinn, K. R., Buchsbaum, D. J., Chaudhuri, T. R., Mountz, J. M., Grizzle, W. E., and Rogers, B. E. Noninvasive monitoring of gene transfer using a reporter receptor imaged with a high-affinity peptide radiolabeled with ^{99m}Tc or ^{188}Re . *Journal of Nuclear Medicine* **2000**; 41(5):887-895.

7

Computational methods and bioinformatic tools

7.1

Introduction

The contributions in this book graphically demonstrate that a wide variety of powerful techniques exists for gene sequencing and for the massively parallel investigation of gene expression, both at the mRNA and at the protein level. In addition, much effort is being devoted to the generation of well annotated gene expression databases such as the resources at National Cancer Institute and the National Center for Biotechnology Information (Bethesda, Maryland, USA). The almost industrialised approaches have provided us within a few years with the more or less complete sequence of the human genome and of the genomes of other important organisms such as the fruit fly, the nematode worm and a wide variety of micro-organisms. A recent highlight was the completion of the sequencing of the rice genome, and the complete sequence of the mouse genome may be expected in the near future.

In sheer volume, this data presents a qualitatively new challenge. In fact, a variety of new scientific specialities have developed to convert this data, first into meaningful information, and then into valuable knowledge.

It goes without saying that these developments would not have been possible without parallel advances in computing power. At present, there is no reason to suspect that the pace of development in either field will slacken in the future. In fact, Moore's law, the proposition that computing power per unit costs doubles about every eighteen months, might equally be applied to units of sequencing or gene expression information.

The general field which has sprung up to mine and manage data may be subsumed under the heading of bioinformatics. This chapter provides an introduction to this topic and an overview of the methods currently being used or developed. In addition, a description of the principle databases available for public access at the time of writing is provided.

7.2

Comparative expressed sequence tag analysis

7.2.1

Introduction

One of the best approaches for fast, large-scale gene identification is the analysis of cDNA libraries. cDNAs are made by reverse transcription of messenger RNA isolated from a given tissue, developmental stage or cell type. By sequencing a large number of cDNA clones, one may obtain an overview of the variety and expression level of the transcribed genes in the starting material. In 1991, it was realised that it is possible to survey the genes present in a cDNA library by generating only partial sequences from either the 3' or the 5'-end of the cDNA clones (Adams *et al.*, 1991). The resulting sequences, termed expressed sequence tags (ESTs), have now become the cheapest and most efficient way to

1. systematically sample the transcribed portion of the genome with the possible advantage of obtaining information on the temporal expression and abundance of each transcript (The latter depends on whether primary libraries reflecting the initial mRNA population or subtracted libraries are used. The disadvantage of primary libraries is that very common transcripts can make up to 50 percent of the clones of such a cDNA library, complicating the identification of rare transcripts),
2. identify single nucleotide polymorphisms (SNPs),
3. identify alternative splice variants,
4. assist the gene prediction in genome sequencing projects,
5. construct arrays of non-redundant transcripts that are suitable for differential regulation studies,
6. perform comparative genomics studies, and
7. to identify differentially expressed transcripts *in silico*.

Nevertheless, ESTs have several disadvantages which render them inappropriate for certain applications. Since they are single pass reads, their quality is low, thus the percentage of ESTs with a high score match in similarity searches can be very low. Depending on the type of the cDNA library (primary or normalised) used to generate the ESTs, the redundancy can be very high, thus reducing the chance of identifying rare transcripts (table 7.1). Furthermore, ESTs cannot be used in studies where the sequence accuracy is important such as in phylogenetic investigations. The clustering of ESTs derived from the same transcript and the extraction of a consensus sequence can help to solve this problem. Finally, ESTs reflect the quality of the cDNA library used. Thus, in the case cDNAs primed with oligo(dT) of less than one kilobase insert size are used, the resulting ESTs will mainly cover the 3'-untranslated regions of the transcripts, thus making them unsuitable for gene identification. By contrast, ESTs enriched in 3'-untranslated regions are useful for mapping projects. ESTs may also be derived from intronic or intergenic sequences (Wolfsberg & Landsman, 1997).

Table 7.1 Clustered expressed sequence tag (EST) sets for four different species represented in UniGene (National Center for Biotechnology Information, NCBI; Bethesda, Maryland, USA; state: January, 2002). The percentage of all tags falling into clusters of a specific size in each species is shown. Cluster size 1 means singleton. Note the high redundancy as reflected by the maxima for large clusters.

Cluster size Sequence tag	Percentage of all ESTs falling into cluster size <i>X</i>			
	Human	Mouse	Rat	Zebrafish
1	4.6	1.6	8.6	4.6
2	4.5	0.6	5.6	4.5
3	4.0	1.0	5.7	4.0
4	4.2	0.8	4.4	4.2
5 to 10	14.8	3.1	20.3	14.8
11 to 20	13.1	3.0	22.2	13.1
21 to 50	15.7	7.6	20.9	15.7
51 to 100	9.0	13.7	6.9	9.0
> 100	30.0	68.5	5.4	30.0
Total numbers of ESTs and mRNAs	3,041,000	2,233,000	298,000	99,000

A standard procedure for processing ESTs prior to analysing their content by various methods is outlined here. Finally, computational tools available for performing digital expression profiling and comparative genomics using ESTs are also described.

7.2.2

Processing expressed sequence tags prior to content analysis

Large scale EST projects require automated tools for processing the raw data, termed traces. While a single read from a capillary sequencer normally exceeds 1,000 nucleotide base pairs in length, usually only the first 400 to 500 base pairs are of high quality. In addition, depending on the sequencing protocol, parts of the sequencing vector may be present in the output sequence. These must be removed prior to subsequent analysis steps. A typical automated routine for extraction of high quality sequence reads from raw trace data comprises of the following steps (see also table 7.2): (i) Quality measurement using the Phred programme (Ewing & Green, 1998) and extraction of high quality sequence. (ii) Detection and clipping of remaining vector segments using the tool *cross_match*, which like Phred is part of the Phrap package. (iii) Masking of simple repeats and regions of low complexity (e.g., adenine-rich stretches), which would bias subsequent database searches using the tool RepeatMasker or its recently published faster version MaskerAid (Bedell *et al.*, 2000).

As mentioned above, ESTs are highly redundant, partial and error-prone sequences. Much effort has been invested by many groups in order to assemble ESTs into a catalogue of unique genes. Two main approaches are used. One is based on pairwise sequence comparisons at different levels of stringency, while the other relies on assembly algorithms (Bouck *et al.*, 1999). The UniGene index (Boguski & Schuler, 1995; <http://www.ncbi.nlm.nih.gov/UniGene/>) is built using pairwise comparison normally performed by different versions of the basic local alignment sequence tool (BLAST) algorithm (Altschul *et al.*, 1997) while GeneNest (Haas *et al.*, 2000) uses a pre-clustering step based on pairwise comparison followed by assembly. The most common assembly programmes used are CAP3 (Huang & Madan, 1999), Phrap (<http://www.phrap.org/>) and the TIGR assembler (Quackenbush *et al.*, 2000) (see also table 7.2 for further information). These assembly programmes were developed for genomic sequences. Their performance in clustering ESTs is of variable quality because they assume a quality of sequence data that may not be available for all of these short tags. Of the three assembly programmes, CAP3 was found to have the best performance for expressed sequence tags (Quackenbush *et al.*, 2000). The TIGR (The Institute of Genomic Research, Rockville, Maryland, USA) gene indices (Quackenbush *et al.*, 2001; <http://www.tigr.org/tdb/tgi.shtml>) and STACK (Sequence Tag Alignment and Consensus Knowledgebase; South African National Bioinformatics Institute, University of Western Cape, Bellville, South Africa) (Christoffels *et al.*, 2001; <http://www.sanbi.ac.za/Databases.html>) are constructed using either one of these assembly programmes. All these clustering programmes are effective to a different extent when dealing with the inherent problems of ESTs such as alternatively spliced products, chimeric clones and contamination by intronic sequences (Bouck *et al.*, 1999). Consed (Gordon *et al.*, 1998) is an editor that can be used for viewing Phrap assemblies.

7.2.3

Gene content and annotation of expressed sequence tags

The first step in annotating ESTs is to search for similarity in protein and nucleic acid databases using either the BLAST (Altschul *et al.*, 1997; <http://www.ncbi.nlm.nih.gov/BLAST/>) or the FASTA (Pearson, 1994) programmes. Additionally, the protein matches from the above similarity searches can be used to construct alignment profiles for use in the search for further homologues using the SAM (Sequence Alignment and Modeling Software System) (Karplus *et al.*, 1998), the estwisedb (Wise2 package) (Birney *et al.*, 1996) or the HMMER (hidden Markov model profiler) (Eddy, 1996) programmes. Normally, the homology searches generate large output files. These must be processed in order to select and evaluate the most significant hits. Such filtering is carried out by measuring the level of similarity and the length of the matching region. See table 7.2 for the internet links at which the tools mentioned above are available.

A bottleneck of annotation by homology is that current protein and nucleotide databases mainly consist of mammalian data, and therefore reflect the gene repertoire of only a few species (table 7.3). Database searches with evolutionary distant

Table 7.2 Standard workflow of an expressed sequence tag (EST) analysis procedure. Internet links marked with * are free for academic users, links marked with § are available on request, whereas links marked with # are available by licence only.

Step	Program	Availability
(a) EST processing	Phred	http://www.phrap.com *
	cross-match	http://www.phrap.com *
	RepeatMasker	http://ftp.genome.washington.edu *
	MaskerAid	http://sapiens.wustl.edu/maskeraid *
(b) EST grouping	Pairwise searches (using Wu-BLAST or gapped BLAST)	ftp://ftp.ncbi.nih.gov/blast/ *
	CAP3	huang@mtu.edu §
	TIGR assembler	license@tigr.org #
	Phrap	http://www.phrap.com *
	Consed	http://www.phrap.org/consed *
(c) Homology searches	BLAST	ftp://ftp.ncbi.nih.gov/blast/ *
	FASTA	ftp://ftp.virginia.edu/pub/fasta *
	SAM	http://www.cse.ucsc.edu/research/compbio/sam.html *
	estwisedb	http://www.sanger.ac.uk/Software/Wise2 *
	HMMER	http://hmmer.wustl.edu/ *
(d) Prediction of open reading frames	ESTscan	http://www.ch.embnet.org/software/ESTScan.html *
	Framefinder	http://fe.hgmp.mrc.ac.uk/gslater/estate.tar.gz *
(e) Domain detection	InterPro	http://www.ebi.ac.uk/interpro *
(f) Functional classification	Gene Ontology	http://www.geneontology.org *

Table 7.3 The dominant species in dbEST. Around 70 percent of all expressed sequence tags (ESTs) derive from mammals, more than 40 percent from humans alone (actual state: January 11th, 2002).

Domain	Species	Number of ESTs
Vertebrates	<i>Homo sapiens</i> (human)	3,927,122
	<i>Mus musculus</i> and <i>Mus domesticus</i> (mouse)	2,514,136
	<i>Rattus norvegicus</i> and other <i>Rattus</i> species (rat)	317,151
	<i>Bos taurus</i> (cow)	213,787
	<i>Danio rerio</i> (zebra fish)	185,251
	<i>Xenopus laevis</i> (African clawed frog)	198,118
Invertebrates	<i>Drosophila melanogaster</i> (fruit fly)	255,456
	<i>Caenorhabditis elegans</i> (nematode worm)	135,203
	<i>Ciona intestinalis</i> (ascidian)	82,071
Plants	<i>Glycine max</i> (soya bean)	223,351
	<i>Lycopersicon esculentum</i> (tomato)	141,735
	<i>Arabidopsis thaliana</i> (thale cress)	113,330
Total number of ESTs		10,032,616

organisms will therefore fail in many cases. For instance, the analysis of the invertebrate genomes such as the *Drosophila* genome (Rubin *et al.*, 2000b) revealed that only 50 percent of the *Drosophila* genes have homologues among the mammalian sequences. This means that around half of the invertebrate genes cannot be annotated through simple homology searches.

Since protein domains are broadly conserved among eukaryotes, they provide an additional way to functionally classify expressed sequence tags (Lander *et al.*, 2001). Furthermore, as protein domains are short, the probability of them being observed in partial sequences is high. Therefore, more ESTs can be annotated through the protein domains they contain rather than through database homology searches. The most comprehensive database of protein domains and functional sites is InterPro (Apweiler *et al.*, 2000; www.ebi.ac.uk/interpro). However, ESTs must first be translated into amino acid sequence before they can be used for this purpose. The ESTscan (Iseli *et al.*, 1999) and Framefinder (<http://fe.hgmp.mrc.ac.uk/gslater>) programmes show good performance in translating ESTs since they can tolerate frameshifts caused by sequencing errors.

Classification of ESTs into functional categories provides information on the type of processes that are carried out in the tissue used for generating the tags while also allowing the comparison between two different subsets of tags. The most widely accepted nomenclature for the description of functional classes of gene products is provided by the Gene Ontology¹⁾ (GO) database (Ashburner, 2000). ESTs can be classified into functional classes defined by the Gene Ontology using either the keywords of their best matches in the SWISS-PROT database (<http://www.expasy.org/sprot/>) or by means of the protein domain they contain. In many cases, the InterPro domains and the keywords of SWISS-PROT entries are linked to Gene Ontology terms.

7.2.4

Expressed sequence tags in comparative genomics

During the past five years the number of genomes that have been completely sequenced has increased dramatically. Since 1995, the year when the first genomes of two bacteria *Haemophilus influenzae* and *Mycoplasma genitalium* were finished, the complete genomes of an additional 69 microbial genomes (twelve archaea and 57 bacteria species) and five eukaryotes (*Saccharomyces cerevisiae*, *Schizosaccha-*

1) An **ontology** is an **explicit specification of a conceptualisation**, which in turn is an abstract, simplified view of the world based on the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. The term is borrowed from philosophy, where ontology is defined as: "The science or study of being; that department of metaphysics which relates to the being or essence of things, or to being in the abstract." (Oxford English Dictionary Online; <http://dictionary.oed.com/>

[entrance.dtl](http://dictionary.oed.com/entrance.dtl)). Ontologies are often equated with taxonomic hierarchies of classes, but differ from these in not being limited to definitions that introduce terminology without adding knowledge about the world. In contrast to a class, a conceptualisation is specified by stating **axioms** that constrain the possible interpretations for the defined terms. Adapted from Gruber (1993; <http://www.ksl.stanford.edu/kst/what-is-ontology.html#1>).

romyces pombe, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Homo sapiens*) have become available. This profusion of genomes has sparked a series of whole-genome comparisons in order to assist the functional annotation of newly sequenced genomes, find a core proteome shared by the genomes (Makarova *et al.*, 1999) and estimate the extent of gene loss or expansion in specific lineages (Aravind *et al.*, 2000).

In the first instance, functional information concerning proteins from an already-studied organism is transferred to its orthologue in another organism. The term orthology describes the relationship between homologous characters whose common ancestor lies in the cenancestor of the taxa from which the sequences were derived (Fitch, 1970). However, the functional annotation based on this method should be regarded as preliminary since sharing of ancestry between two genes does not automatically imply common function.

There are two main approaches for comparisons of whole genomes: (i) BLAST search for similarities between the gene complement of one organism and that of another followed by the selection of matches above a certain threshold as applied in the comparison of yeast and worm (Chervitz *et al.*, 1998) or between *Drosophila melanogaster*, *Caenorhabditis elegans* and yeast (Rubin *et al.*, 2000b) genomes. (ii) Pairwise searches of the predicted proteins from the participating genomes. The best hit(s) for each protein in each of the other genomes is identified as its orthologue (Best Hit or BeT method) (Tatusov *et al.*, 1997). The main difference between this method and the previous one is that a threshold is not needed. The Cluster of Orthologous Groups (COG) database of the National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov/COG/>) constructed using the BeT method currently contains the orthologues found in 44 genomes of bacteria, archaea and three eukaryotes, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster* (Tatusov *et al.*, 1997 and 2001).

Although the speed of sequencing entire genomes has dramatically increased as sequencing methods and strategies have been improved, sequencing of a multi-cellular eukaryotic genome is still a time-consuming endeavour. However, ESTs are available even for organisms for which genome sequencing is not planned. Furthermore, there is over-sampling of the expressed sequence tags of organisms whose genomes are undergoing sequencing, as this will aid later annotation of their genome.

Therefore, an important use of ESTs would be in comparing proteomes of organisms in a manner similar to the comparison of entire genomes. This has been applied in the HomoloGene database (Wheeler *et al.*, 2002; <http://www.ncbi.nlm.nih.gov/HomoloGene/>) which includes orthologous gene pairs selected by reciprocal nucleotide comparison of clustered ESTs that exceed a certain cut-off in two organisms.

Assuming that the ESTs have been processed, clustered and translated using the methods described above, the following conditions should apply to the comparison of genomes. The first step is to decide on the method of comparison. As ESTs are partial sequences, the BeT method is likely to be more sensitive. Depending on the information that is intended to be extracted by reciprocal comparisons, the percen-

tage of genome coverage that the ESTs offer might prove critical. As an example, the identification of a core proteome set presupposes that the compared gene sets are more or less complete. If the orthologous groups are to be used for genome annotation then the completeness of the compared gene sets might not be critical as long as they offer a sufficient coverage of the genome. The percentage of an organisms' genome covered by its ESTs depends not only on the number of tags generated, but also on whether the project is based on a pre-normalised cDNA library. As an example, 40 percent of *Drosophila melanogaster* genes were covered by the 80,000 randomly selected sequence tags from the 5'-region of genes in a library that had not been normalised. These were available at the time of completion of sequencing of the *Drosophila* genome (Rubin *et al.*, 2000a). Estimates of the human gene number prior to the release of the complete genome sequence revealed that approximately 80 percent of human genes were covered by the two million ESTs available at that time (Ewing & Green, 2000).

If one attempts to identify gene families between organisms with the purpose of using them in phylogenetic studies, alternative splicing of ESTs would result in including the same gene many times. However, as alternative splicing affects 3'-untranslated regions less than coding regions (Kan *et al.*, 2001), clustering of 5' and 3'-expressed sequence tags may reduce the redundancy of tags within the EST assembly and thus partially eliminate the above problem.

7.2.5

***In silico* subtraction using clustered sets of expressed sequence tags**

The basic idea in using EST data for the analysis of gene expression is to compare relative abundance of specific cDNAs in libraries of different origins. The best tools available are the Digital Gene Expression Displayer (DGED; <http://cgap.nci.nih.gov/Tissues/GXS>) and the cDNA xProfiler at the Cancer Genome Anatomy Project (CGAP) website (<http://cgap.nci.nih.gov/Tissues/xProfiler>). The cDNA xProfiler can identify differences in gene expression between two sets of cDNA libraries. The selection of groups of libraries from healthy and diseased versions of the same tissue or from different tissues allows the differences or similarities in gene expression patterns to be analysed. The cDNA xProfiler displays the genes found in both sets of libraries, as well as genes expressed in only one of the selected library sets. For example, if one compares the genes expressed in libraries prepared from lung tumours with libraries derived from normal lung tissue, the display indicates the genes found in both tissue or in normal or cancer tissue only. The Digital Gene Expression Displayer displays only statistically significant differences. The investigator can perform complex analyses based on the inclusion or exclusion of libraries derived from multiple organs. For example, all blood cell-derived libraries could be grouped into one set for analysis. The Digital Gene Expression Display provides a Virtual Northern informatics tool for the analysis of gene expression *in silico*. These tools are helpful if an investigator has identified a particular gene of interest that is, for example, highly expressed in breast cancer but not in a normal mammary gland, and would like to know if this gene is also

expressed in other tissues. By employing the Virtual Northern tool, the researcher can view information about the expression of that gene within all of the cDNA and serial analysis of gene expression (SAGE) libraries in the Cancer Genome Anatomy Project collection. In a similar manner, SAGEmap (<http://www.ncbi.nlm.nih.gov/SAGE/>) allows the comparison of gene expression among selected SAGE libraries and identification of statistically significant differences in expression levels. While the analysis tools are very powerful, the resulting analysis is only as informative as the care given to the design of the query. At the Cancer Genome Anatomy Project website, the user can specify the biological source of the tissue and the methods used to prepare the tissue and construct the cDNA library. The methods used to prepare the tissue and the library greatly affect the profile of gene expression in that library. For example, comparing the gene expression profiles in a standard library constructed from normal brain tissue with those of a normalised library made from cancerous brain tissue would not produce meaningful results because the complexity and sequence content of the libraries are highly divergent. The depth to which cDNA libraries are sequenced varies from fewer than 1,000 to 50,000 ESTs. This may affect the significance of an *in silico* analysis. Libraries selected for use in an analysis should contain a similar number of sequences. A complete description of each CGAP library and all programmes can be found at <http://cgap.nci.nih.gov/Tools>.

7.2.6

Expressed sequence tag data repositories and cDNA clone distribution centres

Most expressed sequence tags are collected in the database of expressed sequence tags (dbEST) at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). Currently, dbEST contains more than 10,000,000 single ESTs from over 350 different species spanning all animal phyla with the human and mouse section comprising 60 percent of the whole database (table 7.1).

Assemblies of publicly available ESTs have been constructed from multiple sources. The National Center for Biotechnology Information offers species specific UniGene sets though without providing the respective consensus sequences. Consensus ESTs are provided through the TIGR gene indices (Quackenbush *et al.*, 2001), STACK (Christoffels *et al.*, 2001) and GeneNest (Haas *et al.*, 2000). Multiple sites also offer online BLAST similarity search options *versus* the consensus or the whole clusters of assembled ESTs such as the US American National Center for Biotechnology Information (NCBI), the German Resource Center Primary Database (RZPD; <http://www.rzpd.de>), GeneNest and the Italian Telethon Institute of Genetics and Medicine (TIGEM; <http://www.tigem.it/>).

Distribution of the libraries and individual clones used to generate the tags currently available at dbEST is carried out by the IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium (Lennon *et al.*, 1996; <http://image.llnl.gov/>), the German Resource Center Primary Database, the American Type Culture Collection (ATCC; Maglott & Nierman, 1990), Research Genetics (<http://www.resgen.com>) and the Human Genome Mapping Project

(HGMP; <http://www.hgmp.mrc.ac.uk/>). As the mixing of clone coordinates during EST sequencing can be high, care should always be taken when a clone is ordered based on the sequence dbEST indicates to correspond with this clone. Some distribution centres such as the German Resource Center Primary Database and Research Genetics supply clones which have been verified by re-sequencing.

7.3

Data management and data mining

7.3.1

Introduction

This article does not aim to provide a step by step guide on how to do interdisciplinary bioinformatics. Rather, its purpose is to focus on the philosophy of data analysis and bioinformatics, enriched by some hints, links and descriptions which may prove useful to the reader.

7.3.1.1 Current situation

The vast amount of biological data contained in the human genome, transcriptome and proteome, the complexity of differences between groups of individuals, and the number of aberrations with consequences for health make it necessary to develop efficient strategies for the management, processing, integration, retrieval and representation of data.

Research and routine work have different demands, but the technological base is the same for both applications. Thus, the following discussion applies to both situations.

Many different databases exist that were developed by different operators using different means in different legal situations. Some have become quasi standards which will give leadership to others (European Bioinformatics Institute, EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, available at <http://www.ebi.ac.uk>; GenomeNet, Bioinformatics Centre at the Institute for Chemical Research of the Kyoto University, Kyoto, Japan, available at <http://www.genome.ad.jp>; National Center of Biotechnology Information, NCBI; Bethesda, Maryland, USA, available at <http://www.ncbi.nlm.nih.gov>), but by and large no two databases are the same (*e.g.*, Human Mitochondrial Genome Database, MITOMAP; Center for Molecular Medicine, Emory University, Atlanta, Georgia, USA, available at <http://www.gen.emory.edu/mitomap.html>).

These more or less systematic databases are sometimes joined by a common graphical user interface (GUI) allowing biologically or healthcare-orientated researchers to carry out searches and analyses in a joint environment (see the WWW Interface to Sequence Analysis Software Tools, W²H at the German Cancer Research Center, DKFZ; Heidelberg, Germany, available at <http://industry.ebi.ac.uk/w2h/>). In most cases, search queries are manually constructed and the results are evalu-

ated manually. Some DNA microarray-related software automatically searches annotations to existing gene lists. Larger facilities may also have bioinformatic specialists to develop automated solutions for high-throughput experiments. However, there is still no substitute for evaluation of the end-result by hand.

So far, all established solutions have three major shortcomings: (i) They provide no information on how to accurately model the data to give it meaning in the sense of unrevealing rules for the complex biological interactions. (ii) They are not self-explanatory and fail to integrate many aspects of a species, not to mention the complexity which exist with each member of that species. (iii) Quality control of results is poor.

7.3.1.2 Future development

The technological challenge is to establish solutions which are independent of the operating system and the specialised local environment.

The political challenge is to establish trans-national scientific standards and software standards that can provide a basis for complex bioinformatic solutions. First steps in this direction are visible in the activities of such non-commercial institutions as the European Bioinformatics Institute, the National Center of Biotechnology Information and commercial enterprises such as the Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom; available at <http://www.sanger.ac.uk>).

7.3.1.3 Taking part in bioinformatics

The first step in setting up a bioinformatic project is to carefully frame the biological question, to define the role of the collected or modelled biological data and to check if better projects in the same field are under way elsewhere.

The second step is ensuring that adequate local resources are available. Interdisciplinary teams including biologists, statisticians, computer experts, biochemists, physicists, and physicians are an important prerequisite. Often this is not available in one institution, so cooperation needs to be established with outside groups.

In practical terms, it is important to develop a common language and mode of thought to allow the different professions to communicate effectively. Finally, contingency plans should be made to take account of the charge that are likely to occur in the future. In particular, the project should be able to adapt to future changes without large additional costs in resources and engineering. A step-by-step analysis of a putative informational system to store, retrieve and analyse biological data is shown in figure 7.1.

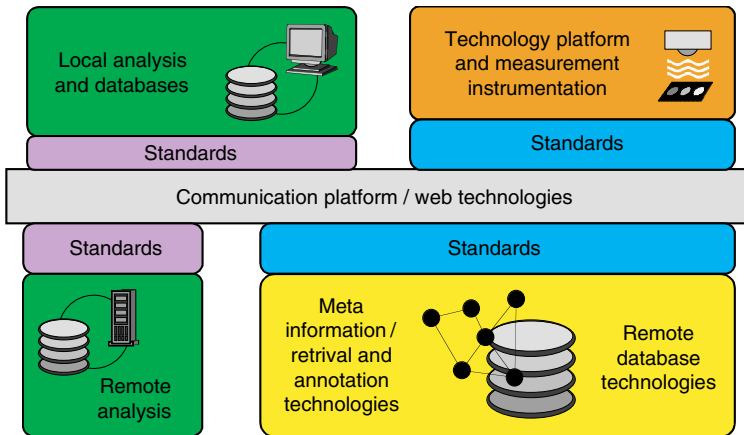


Figure 7.1 Compartments of information technologies. This scheme shows an abstract overview of the important modules for generating, hosting, transmitting and processing in-

formation in a structured bioinformatic environment. The purpose of the communication platforms is to standardise information exchange.

7.3.1.4 Hardware and software demands

Hardware is getting faster and cheaper. At the time of writing most projects are adequately served with workstations or small servers in the range of € 5,000.-. A network with a high bandwidth is more crucial. At the present time, a ten to one-hundred megabit connection to the desktop and a fast backbone net are required.

The choice of a operating system is more problematic. The LINUX and Microsoft® Windows operating systems (OS) are competitors in this area, whereas the Macintosh operating system shows signs of drifting towards the LINUX operating system. One answer to this heterogeneity is to use web-based applications (see information for software developers from Developer Shed, Inc., St. George, Utah, USA, available at <http://www.devshed.com>) or applications which run in all environments, a technically very demanding task. At the present time, LINUX appears to offer advantages in terms of lower licensing costs and a more skilled community of users.

7.3.1.5 Data types, structures and processing

Having chosen the operating system and database platform, we now need to build the informational structure and organisation of data, which implicitly imparts a basic form to later data modelling.

Generally, data is represented in relational databases in 'fields' and 'records', also called lines. Many records form a database table. Fields symbolises the properties while records comprises samples, cases or similar groups of observations. Rela-

tional means that some or all properties may have more than one value and thus must to be stored in sub-tables.

To search in a database a 'query' is generated, mostly in a Boolean form. The results are represented as a table, where the column headers are named with the appropriate field names and lines represent the set of observations belonging together.

The next stage is to integrate the internal data collection with the analysis and modelling platforms and external data collections. This raises the questions of interface structures, data formats and data exchange protocols. On one operating system platform, these steps can normally be done intuitively because all applications have ports for other important applications. In more complex environments with internet connections and different operating systems, solutions may be more difficult to program. The simplest and most robust data format in most cases is a simple American Standard Code for Information Interchange (ASCII) file.

Some common database interfaces are the Open Database Connectivity (ODBC) (see the Visual Basic Bookmark at <http://www.vb-bookmark.com>) and the Java Database Connectivity (JDBC) (see The Source for Java™ Technology at <http://java.sun.com>) standards in conjunction with Java, Professional Homepages (PHP) (PHP group at <http://www.php.net>), Perl (Comprehensive Perl Archive Network, CPAN; available at <http://www.cpan.org>) and other internet-enabled languages. New in this context is .Net with the language C#, a language which is currently under development.

7.3.1.6 Communication structures

The easiest way to share scientific information is to enter the data into existing public databases, where the publication service is provided by other organisations.

If results should be made available to the scientific community using local means, perhaps because there are no public data resources fitting this type of data, than the user is faced with many technological problems. The first task is to implement a universal technological and logical interface. The technological interface may be provided by internet-based technologies such as Hypertext Markup Language (HTML) (World Wide Web Consortium, W3C, at <http://www.w3.org>), Extensible Markup Language (XML) (W3C at <http://www.w3.org/XML>) and PHP. To provide the logical interfaces, the form in which the data should be presented to the public must be defined first. At present, problems arise if very large amounts of data must be transmitted in a single session. The existing software solutions for the internet do not deal very well with this situation.

7.3.2

Building a test scenario

7.3.2.1 Microarray experiments

The following scenario is based on a small microarray facility in a university institute with several external guest users.

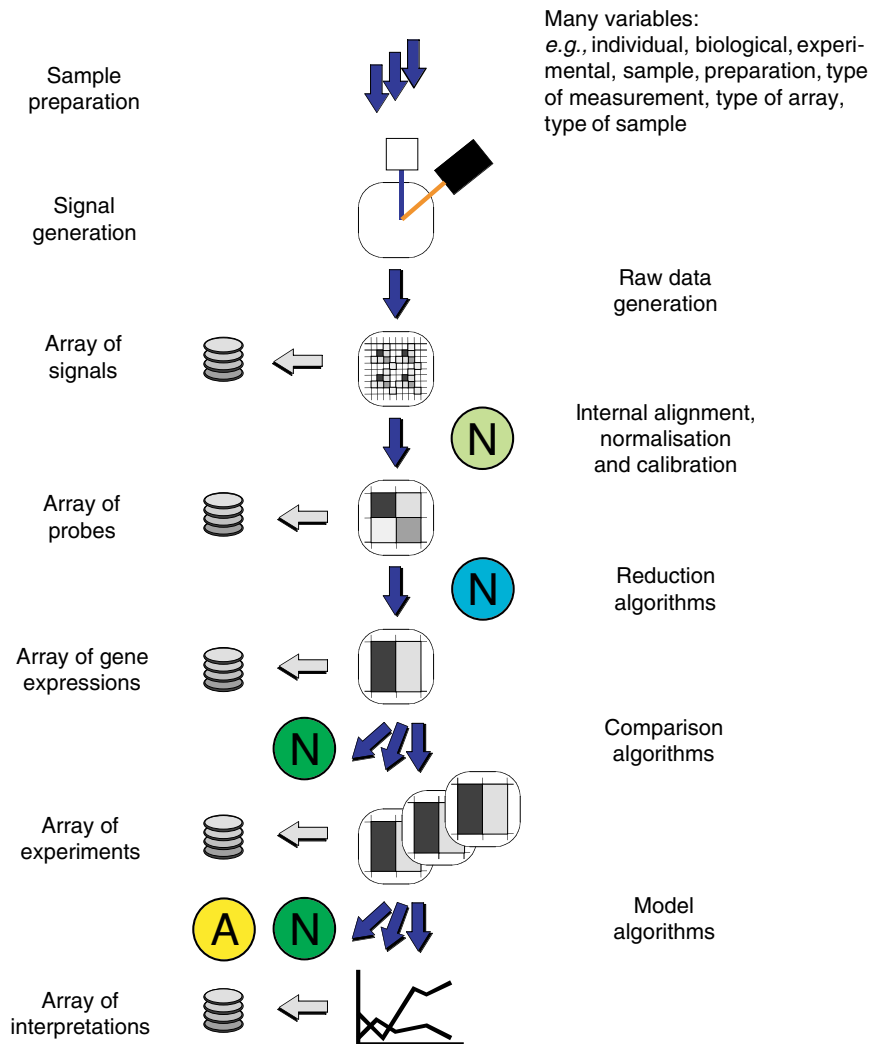


Figure 7.2 DNA microarray experiments – evaluation process. Typical procedures in array experiments. After sample preparation, the array is scanned and the raw data is normalised and calibrated. The probes are combined to form a generalised measure for one gene. Genes are compared across sets of arrays. The evaluation procedures and the interpretation schemes may be varied in order to produce an optimised result. Note that normalisation procedures may critically affect results. ‘N’ denotes

different normalisation operations, ‘A’ denotes annotations of available secondary information that is linked to the generated data. In every step after generation of the raw data, additional data is generated as symbolised by grey arrows and a grey disk array. In some cases (three blue arrows), there may be more than one way to evaluate the data. The direction of the evaluation is directed towards higher levels of abstraction.

Figure 7.3 Real world DNA array.

In the upper part of the picture the raw data of an entire array is shown. Below all the probe pairs belonging to one gene are assembled together. To reduce the possibility that local abnormalities may eliminate all the ten to twenty probe pairs (depends on the array type) representing a gene, the probe pairs for one gene are randomly distributed over the array area. Each probe pair is build up from a perfect match (PM) and a mismatch (MM) probe. The part of the picture below shows that it is not simple to calculate the signal strength of gene expression.



In this case, the database will contain several gigabyte of raw mRNA expression data (figure 7.2). This raw data must first be processed using a standardised system-specific normalisation procedure. Normally, no improvement in this step is possible on site.

In the next step, this pre-processed data is used to analyse properties from different experiments. At this point, the comparability of different experiments must be established (see, for example, Zhang *et al.*, 2000). Many different approaches exist, based either on average signal levels or on internal standards (Callow *et al.*, 2000). It is very important to establish a clear procedure for comparison, because this step has a great influence on experimental results. At this point, weakness in the experimental design will also become apparent.

7.3.2.2 Analysing the workflow – getting things done

In rare cases, a single person may be skilled enough and has enough time to do all the necessary molecular biology, informatics and statistics procedures. In practice, extensive cooperation, discussion and documentation is necessary. It is very important to use a precise documentation system to control the workflow (figure 7.3). Based on this documentation, the bioinformatics expert and statistician may be able to build tools to accelerate the evaluation procedures (Altmann *et al.*, 2001).

7.3.2.3 Designing the question and choosing the right tools for the answer

In most cases, it will be necessary to reduce the complex real world problem to manageable dimensions. It must therefore be ensured that the few measurable snapshots of gene expression will be able to answer the biological question (see, for example, Mills *et al.*, 2001).

Once the data is generated, instrument-specific pre-processing is carried out, usually by means of commercially available software. Some solutions, such as the GeneData programme package The Expressionist® (GeneData AG, Basel, Switzerland), are written in Java and generally available on every internet-enabled system. A major focus of GeneData's software is quality control. Other programme packages, such as GeneSpring™ (Silicon Genetics, Redwood City, California, USA), are aimed at Microsoft® Windows users. The most advanced solutions are available under LINUX, but are generally only practicable for experienced users. It is usually worthwhile to perform a literature and internet search before analysing the data because of the rapid pace of development and the open nature of the bioinformatics community (see, for example, Bioinformatics.Org, USA, at <http://bioinformatics.org> and Terry Speed's Microarray Data Analysis Group website, Department of Statistics, University of California, Berkeley, California, USA, at <http://www.stat.berkeley.edu/users/terry/zarray/Html/links.html>). In many cases, it may be sufficient to use standard software to perform a standard analysis (*e.g.*, so-called twofold changes in expression levels).

7.3.2.4 Scaling up

Generally, at the present time, many molecular biologists and biochemists, but too few bioinformaticians and statisticians are available to evaluate the enormous amount of data created. This situation will improve as analysis techniques are standardised and automated. But for the foreseeable future, technicians and scientists must be specifically recruited to administrate the workstations and servers, to oversee the data management technology, to establish analysis tools, and to evaluate ongoing projects. Ideally, all operatives should have some experience in the other fields, and should at least be able to speak a common language.

Based on daily experience, useful microarray analysis with a standard commercial system, such as the Affymetrix GeneChip® analysis system, requires at least two technicians and three bioinformaticians. Use of customised arrays requires more quality control and more laboratory staff.

7.3.3

Strategies of data mining

7.3.3.1 Data evaluation and representation

The data should be logically organised in tabular form with sub-tables if required. Every signal value should be measured at least three times to improve normalisation (so-called threefold experiments). This is also the minimum for standard statistical analysis. In addition to standard statistics (*e.g.*, Wilcoxon), novel

approaches such as Bayesian logic may be used (Baldi & Long, 2001). Typically, a line plot – also called a profile plot – is used to visualise the expression behaviour of a single gene or gene sets in a number of different microarray experiments.

In order to reveal complex biochemical pathways, the raw output of the experiment must be linked to internet repositories such as GenBank (available at <http://www.ncbi.nlm.nih.gov>) and others (see, for example, links of the ExPASy server at <http://www.expasy.org> and links to database services of the European Bioinformatics Institute server at <http://www.ebi.ac.uk/Databases/index.html>). The software should be able to integrate this information in an automated user-friendly fashion into a local database. The linking with the experimental data must be completely transparent for the user.

If repeated experiments are performed, it is often useful to write scripts in order to automate procedures. Many software packages offer scripting tools, though these are often limited in scope.

7.3.3.2 Principles of query languages

As the size of the database grows, efficient querying routines are required. Most of the software packages provided with the hardware include simple querying routines. While simple search routines or filter systems are based on simple Boolean 'AND', 'OR' and 'NOT' operations, more advanced systems utilise the full range of the Structured Query Language (SQL). See the SQL tutorial of Refsnes Data (Norway) at <http://www.w3schools.com/sql> for detailed information on SQL. Simple statements are assembled by varying the number and value of field parameters (figures 7.4 and 7.5).

7.3.3.3 Data mining

Data mining refers to a structured, multi-step approach to extract biologically meaningful information from databases. For example, a sophisticated query may be sent to one server, and the result may form the basis of other queries and third party analysis, which ultimately leads to information on a complex structure such as a regulatory pathway.

	Probe Set Name	Signal	Detection	Detection p-value	Stat Pairs	Stat Pairs Used
Sort:	Ascending					
Line 1:		>50	= 'P'			
Line 2:			= 'P'	<0.1		
Line 3:						

Figure 7.4 Query area of a typical data mining tool. The query is based on signal properties. Each column represents a property which should appear in the result set. Restrictions are directly placed into the appropriate cells. Different lines represent a combination of restriction criteria by logical 'OR'. Restrictions in the same line are combined by a logical 'AND'. The structured query language (SQL) basis of the query is hidden from the user.

```

SELECT      ITEM_NAME, SIGNAL, NAME, DETECTION_PVALUE, PAIRS, PAIRS_USED
FROM        ABS_GENE_EXPR_RESULT, ABS_GENE_EXPR_RESULT_TYPE
            ABS_GENE_EXPR_RES_STAT, BIOLOGICAL_ITEM
WHERE       ((SIGNAL>50 AND NAME like 'P') OR
            (NAME like 'P' AND DETECTION_PVALUE<0.1)) AND
            ABS_GENE_EXPR_RESULT.ITEM_ID=BIOLOGICAL_ITEM.ID AND
            ABS_GENE_EXPR_RES_STAT.ITEM_ID=BIOLOGICAL_ITEM.ID AND
            ABS_GENE_EXPR_RESULT.TYPE_ID=ABS_GENE_EXPR_RESULT_TYPE.ID
            ABS_GENE_EXPR_RES_STAT.TYPE_ID=ABS_GENE_EXPR_RESULT_TYPE.ID
ORDER BY    ITEM_NAME

```

Figure 7.5 Principles of the structured query language (SQL). This sample shows the underlying SQL statement of figure 7.4. The query is based on typical SQL syntax. The SELECT statement returns the field values of interest. The FROM statement include all the affected tables of the database. The WHERE statement contain the query specific and SQL specific restrictions. The ORDER BY statement sorts using certain field values. Depending on the SQL server, there may be some minor differences in the syntax.

Most commercial user interfaces can perform queries enriched with annotations and may also be capable of simple re-evaluations. Thus, they are all simple data mining tools.

Data mining has its origin in the world of business and commerce, where it is more advanced than in the life sciences. However, data structures in business are well-described and much less complex than in biological systems.

At the present time, data modelling is generally performed by hand on platforms such as R (for information on R see The R Project for Statistical Computing at <http://www.r-project.org/> or the R resources of the Comprehensive R Archive Network (CRAN) at <http://cran.r-project.org/> or S-Plus® (Human Resources, Insightful Corporation, Seattle, Washington, USA), sometimes comprising functions programmed in the FORTRAN (Formular Transistor) programming language that allows the use of fast numerical algorithms. So-called ‘meta information’ and biological annotations must be assembled by hand because of the lack of a defined universal biological information structure. Many commercial and scientific solutions also exist (see, for example, the online tools of the Swiss Institute of Bioinformatics, Geneva, Switzerland, at <http://www.expasy.org/tools/> and the Bioinformatics Template Library at <http://www.cryst.bbk.ac.uk/classlib/bioinf/BTL.html>), but none may completely satisfy a particular need.

7.3.3.4 Custom solutions

In the following paragraphs some ways to custom solutions are mentioned.

- (a) *The experiment is straight forward and only core results are required.*
Use of a commercial technology platform including the basic software is recommended.
- (b) *The experiments revealed a very large amount of data.*
Use of a commercial data-management tool or the construction of a comparable tool is recommended. The latter is more time-consuming and requires comput-

ing skills. Then the user has to decide if he requires approach (a) or (c) for analysis of his data.

(c) *Additional evaluation is required.*

The theoretical basis of the application has to be established and over-interpretation of the results must be avoided. Most procedures require compliance with a stringent set of conditions concerning the data, which are often not fulfilled in practice. This is important because no amount of manipulation can produce useful information from a poorly designed experiment.

(d) *Establishment of a bioinformatics facility.*

Customise all the things you need on a very flexible basis using a tool such as R or S-Plus®. Enter into cooperation and share your results with the community and be patient.

7.3.4

Summary

Bioinformatics requires first and foremost close collaborations between experimentalists and theoreticians. The scientist has to think carefully about the results he expects from his experiments. Refinement of the study design in close discussion with the theoreticians is a prerequisite and the internet is a helpful and comprehensive resource and library, both for computational solutions and for publication of experimental data.

7.4

Integration of heterogeneous high-throughput gene expression data

7.4.1

Introduction

Following the initial sequencing of the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001), requests for comprehensive databases of gene function information have been voiced by many scientists. High-throughput gene expression profiling is one of the most prominent genome-wide approaches to produce functional data – snapshots of simultaneous transcriptional activity for thousands of genes at a time. Not surprisingly, as these experimental techniques become more affordable and robust, the rate of expression data accumulation in the public domain has been increasing dramatically. While the technology behind these studies is still developing rapidly, uniform standards for storage and exchange of gene expression data are already in the making, as are general guidelines for the documentation of data generated in expression profiling experiments. However, it is still difficult to make comparisons between these large data sets in a biologically meaningful way. These problems constitute a factor that is significantly limiting the wider use of published large-scale gene expression observations by the general scientific community.

7.4.2

Steps towards data integration

The levels at which data integration needs to be addressed include, in order of increasing complexity:

- (i) Integration between independent 'batches' or 'prints' of arrays produced and processed by the same group. It is not uncommon for such batches to be analysed separately, which means that only experiments within the same batch can be compared to each other, while data integration between batches is limited to the comparison of trends in expression profiles for a subset of genes.
- (ii) Integration of experiments performed by different groups using the same arraying platform, or differently designed studies within the same group. In addition to the obvious divergence in experimental protocols to include different sets of cDNA or oligonucleotide probes selected for the array, differences in experimental design become particularly important. For example, in the studies that utilise the popular two-channel fluorescent detection on cDNA microarrays, it is important to distinguish between 'type I' experiments, which use an internal reference sample relevant to a given study such as a zero time-point, and 'type II' experiments, which survey a large number of samples, comparing them to the same, often artificially constructed, RNA pool. Clearly, even if performed on the same samples, data sets obtained with different reference pools require processing before they can be reliably integrated. The same applies to the comparison of type I and type II data sets.
- (iii) Integration of data obtained with different expression profiling technologies, such as arrays of oligonucleotides or complementary DNA molecules, or SAGE. Often, the information provided by the researchers or the array providers on the precise nature of the probes attached to the array is insufficient to allow reliable conclusion to be drawn regarding the identity of the mRNA-derived molecules that hybridise to these probes.
- (iv) As a long-term objective, it would be an advantage to meaningfully integrate gene expression profiles across species. However, this type of comparison is very hard to interpret, in part, because of the difficulty in assigning orthologous gene sets between species. More importantly, interpreting high-throughput cross-species expression profiling data will require caution, because even proven orthologues may have adopted different functions since the species diverged, and thus there may be no biological reason for their expression profiles to be parallel.

7.4.3

Initial steps in realising data integration

Several efforts that address the issue of standards in handling high-throughput expression data are currently under way. Arguably the most prominent grass-roots movement is the Microarray Gene Expression Database (MGED) group (<http://>

www.mged.org), which aims “to facilitate the establishment of gene expression databases, comparability of microarray data from different sources, interoperability of different functional genomics databases and data analysis software”. The working groups within the MGED currently include the Minimum Information about a Microarray Experiment (MIAME) group (<http://www.mged.org/Annotations-wg/index.html>), the Ontology group (http://www.cbil.upenn.edu/Ontology/MGED_ontology.html), the Microarray Gene Expression Data Object Model (MAGE-OM) group and the Microarray Gene Expression Data Markup Language (MAGE-ML) group (<http://sourceforge.net/projects/mged/>) that develops the microarray data XML exchange format and the XML-based microarray gene expression markup language, and the Normalisation working group (<http://www.dnachip.org/mged/normalisation.html>).

The data standard recently proposed by MIAME (Brazma *et al.*, 2001) outlines the content, not the format, of the minimum information required for accurate interpretation of a microarray experiment. Six main components exist, the descriptions of which are currently formalised in MIAME: (i) the experimental design for a set of hybridisations; (ii) the array design; (iii) the samples used in the experiment, including all treatment procedures; (iv) the hybridisation protocols; (v) the data itself, from images to quantification and gene expression matrices; and, finally, (vi) detailed information on normalisation controls, which could involve a variety of normalisation schemes performed on different subsets of genes under study (full gene set, exogenous ‘spiked’ RNA, *etc.*). MIAME also offers a number of suggestions for the use of controlled vocabularies, intended to facilitate further querying and automated analysis of gene expression data. The particulars of these standards are likely to continue evolving for some time in order to accommodate both technological development and feedback from the scientific community. One exciting development of future MIAME standards is expected to be the introduction of standards for high-throughput arraying data of non-expression profiling nature, such as comparative genomic hybridisation (Brazma *et al.*, 2001).

7.4.4

Conclusions

Uniform standards for microarray data annotation, storage and processing are essential to allow development of gene expression databases that can be queried and to produce tools for data mining and automated data analysis. The introduction of standard experimental controls and guidelines for statistical analyses will increase the ability to compare data from a variety of sources and will increase the confidence in the validity of conclusions based on numeric data derived from a large number of high-throughput expression profiling experiments.

7.5

Cluster analysis of gene expression profiles

7.5.1

Introduction

Gene expression profiling is a widely used technique in the analysis of macro and microarray data. The transcription level of a large amount of genes is measured simultaneously in a single hybridisation experiment: Representatives of genes, the probes, are immobilised on a solid support (either glass slides or nylon filter membranes) and are hybridised with a pool of labelled cDNAs (either radioactively or fluorescently labelled), the targets (DeRisi *et al.*, 1997; Dickmeis *et al.*, 2001; Eickhoff *et al.*, 2000). The target material depends on the technology in use, either there are single-stranded cDNA sequences or a set of short oligonucleotide sequences (typically eleven to twenty 23-mers). The central assumption in this approach is that the transcription level of an individual gene in the cells of interest can be measured in a linear way by the intensity of the bound fluorescent or radioactive target of the corresponding spot on the support (see chapter 4).

The hybridisation experiment is repeated with a series of cDNA pools, for example, derived from different time points, different tissue types or different cell lines and each gene representative is assigned an *expression profile*, a series of numerical values that describes its transcription state with respect to the tissues under analysis (figure 7.6).

7.5.2

Information content of gene expression clusters

The main assumption in the data analysis of expression profiles is that genes that have similar expression profiles are commonly regulated and are likely to be related functionally (guilt-by-association concept).

The purpose of cluster analysis is at least three-fold. (i) Gene expression clusters help characterising unknown genes assigned to the cluster by those genes that have a known function. (ii) Gene expression clusters are the basis for identifying common upstream regulatory sequence elements (Brazma *et al.*, 2000; Tavazoie *et al.*, 1999). (iii) Gene expression clusters are used as an input to genetic network analyses in order to average individual bias and error of a single gene's profile.

Clustering of expression profiles and functional categorisation is particularly interesting in cases where the complete set of genes is known. Therefore, yeast is the most studied organism in this field. For example, DeRisi *et al.* (1997) studied the time course of the so-called diauxic shift (the switch from one carbon source to another) and Spellman *et al.* (1998) and Cho *et al.* (1998) studied the transcriptional programme of the cell-cycle. These and other studies yield a large publicly available data set including about 80 experimental conditions included in the Stanford Yeast Database at <http://cellcycle-www.stanford.edu/>.

A variety of applications is devoted to the molecular classification of diseases based on gene expression profiling and clustering. For example, leukemias have been classified into two subtypes (acute lymphoblastic and acute myeloid leukemia) based on 50 genes selected from an array carrying about 7,000 gene representatives. These genes included those previously known to differ between acute lymphoblastic and acute myeloid leukemia, as well as new markers (Golub *et al.*, 1999). Another study classified B-cell lymphomas into two molecular distinct

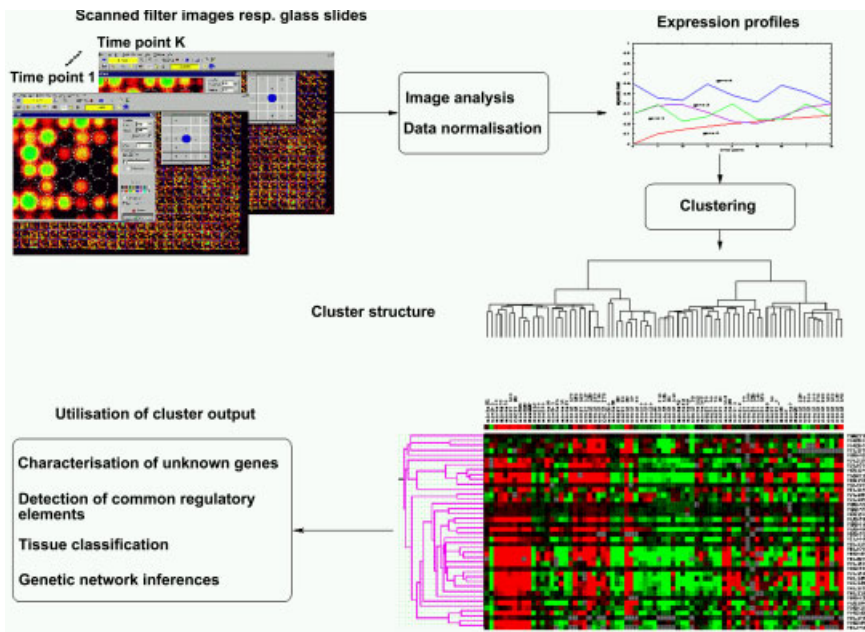


Figure 7.6 Scheme of a gene expression profiling experiment. The relative abundance of a gene is measured by hybridisation experiments of cDNA pools from different conditions, time series, tissue types, *etc.* Through signal acquisition and quantification processes such as confocal scanning, phosphor imaging and image analysis, each gene is assigned a vector of numerical values — the gene expression profile — that describes the gene's abundance in the samples under analysis. Data normalisation steps transform the raw intensity values to normalised values. This is done in order to eliminate factors that do not arise from the probe/target interaction but rather from experimental artefacts such as differences in labelling, transferred material on the support, hybridisation performance, *etc.* The normalised expression profiles of thousands of genes are clustered in order to find groups of commonly

regulated genes. In this example, a cluster of 40 yeast genes is displayed that were immobilised on a glass slide carrying about 7,000 yeast genes. Transcription levels of these genes were measured under 77 conditions during the cell-cycle process (<http://cellcycle-www.stanford.edu/>). Clustering was performed with a hierarchical clustering algorithm (Eisen *et al.*, 1998), using Pearson correlation as a pairwise similarity measure and the average-linkage rule for joining of two clusters (<http://rana.lbl.gov/EisenSoftware.htm>). The clustering output is used to characterise the function of unknown genes using information on known genes, to identify putative regulatory elements in the upstream regions of the genes in the same clusters and reduce redundancy and generate averaging profiles in the context of regulatory networks.

classes that reflect different stages of B cell differentiation (Alizadeh *et al.*, 2000). Here, authors used a customised DNA microarray carrying about 18,000 gene representatives.

These and other studies proof the high potential of microarray bioinformatics such as clustering and classification techniques for scientific and industrial research.

7.5.3

Similarity matrices and gene expression matrices

Given M probes (cDNA clones, oligonucleotides) on the array that were measured by hybridisation experiments with N samples (time points, tissue types, cell lines, *etc.*), clustering algorithms take as input either an $M \times N$ data matrix (the *gene expression matrix*), where the M rows describe the feature vectors of the M data points (probes) across the experiments and the N columns describe the feature vectors of the experiments across the data points or, alternatively, an $M \times M$ matrix (the *similarity matrix*) derived by calculating all M^2 pairwise similarities of the possible pairs of data points according to some pairwise similarity measure.

Commonly used pairwise similarity measures are mathematical functions based on the feature vectors of two data points that take on high values if the data points are similar and low values if they are different. Alternatively, dissimilarity measures (such as metrics) take on low values if two data points are similar and high values if they are dissimilar. A further practical requirement is the handling of missing values since experimental data often contains cases where measurements are extremely noisy so that some values can not be reliably detected. Reasonable pairwise similarity and dissimilarity measures are, for example, Pearson correlation, Spearman's rank correlation, Hamming distance, Euclidean distance and mutual information (Jain & Dubes, 1988; Mirkin, 1996). Each measure has its specific advantages and disadvantages. Furthermore, each measure introduces an element of interpretation into the data structure. For example, Euclidean distance has a geometric meaning in the sense that two points are similar if they are geometrically close to each other; in the case of Pearson correlation two data points are similar if the linear correlation between the feature values is strong; mutual information is an information theoretic concept that quantifies the information about a data point contained in the other data point and *vice versa*. Therefore, there is no 'best' overall choice of similarity measure, it should rather be chosen dependent on the specific application.

7.5.4

Clustering algorithms

In general, a cluster algorithm tries to group together data points that are similar and tries to separate those that are dissimilar where similarity is based on one of the pairwise similarity measures described in the previous paragraph. Two classes of clustering algorithms are commonly distinguished in *hierarchical* and *partition-*

ing methods (Jain & Dubes, 1988; Mirkin, 1996). In contrast to partitioning methods that try to find the ‘best’ partition given a fixed number of clusters, hierarchical methods calculate a full series of partitions starting from M clusters each of which contains one single data point and ending with one cluster that contains all M data points (or *vice versa*); in each step of the procedure two clusters are merged according to a pre-specified rule. In the following, the classical hierarchical algorithm and two commonly used partitioning methods (self-organising maps and K-means) are described. Additionally, gene shaving – a recently developed method that is based on principal component analysis – is also described.

Not surprisingly, there are many other cluster algorithms used in the context of gene expression profiling, for example, algorithms based on concepts of graph theory (cluster affinity search technique, CAST, described by Ben-Dor *et al.*, 2000; cluster identification *via* connectivity kernels, CLICK, described by Sharan & Shamir, 2000), percolation clustering (Sasik *et al.*, 2001), algorithms that are used in supervised classification approaches such as the support vector machines (Brown *et al.*, 1999) and tree harvesting techniques (Hastie *et al.*, 2001). Besides developing new algorithms, a current effort in research is focused on the comparison and integration of different methods under a unifying concept.

7.5.4.1 Hierarchical clustering

Let $x_1, \dots, x_M$ be the data points (expression profiles of M gene representatives across the experiments). The process of hierarchical algorithms requires a dissimilarity measure, d , between pairs of clusters (related to a dissimilarity measure, $\tilde{d}$, between pairs of data points) and an update procedure for recalculation of the merged clusters. It then has the scheme given in figure 7.7.

Several cluster dissimilarity measures are in use:

$$\begin{aligned} \text{Single-linkage} \quad d(C_k^{(v)}, C_l^{(v)}) &= \min_{x_i \in C_k^{(v)}, x_j \in C_l^{(v)}} \tilde{d}(x_i, x_j) \\ \text{Complete linkage} \quad d(C_k^{(v)}, C_l^{(v)}) &= \max_{x_i \in C_k^{(v)}, x_j \in C_l^{(v)}} \tilde{d}(x_i, x_j) \\ \text{Average linkage} \quad d(C_k^{(v)}, C_l^{(v)}) &= \frac{1}{|C_k^{(v)}| |C_l^{(v)}|} \sum_{x_i \in C_k^{(v)}, x_j \in C_l^{(v)}} \tilde{d}(x_i, x_j) \end{aligned}$$

Here, $C_i^{(v)}$, denotes the i^{th} cluster at the v^{th} iteration step ($i = k, l$). In the single-linkage procedure the distance of two clusters is given by the minimal pairwise distance of the members of the one and the members of the other, in the complete-linkage the distance of two clusters is given by the maximal pairwise distance of the members of the one and the members of the other and in the average-linkage

Figure 7.7
Hierarchical clustering algorithm.
See text for details.

1. For $v = M$ start with the finest possible partition.
2. Calculate a new partition by joining two clusters that minimise d .
3. Update the distances of the remaining clusters and the joined cluster.
4. Stop, if $v = 1$, i.e. all data points are in one cluster.

Table 7.4 Update parameters for hierarchical clustering methods.

Method	a_i ($i = k, l$)	β	γ
Single linkage	0.5	0	-0.5
Complete linkage	0.5	0	0.5
Average linkage	$\frac{ C_i^{(v)} }{ C_k^{(v)} + C_l^{(v)} }$	0	0

The parameter β is zero in these examples but there exist other update methods (for example, the centroid and the Ward method) that incorporate a positive β .

procedure the distance of two clusters is given by the pairwise distance of the arithmetic means of the clusters. In all three procedures those two clusters are merged that minimise the cluster distance over all possible pairs of clusters.

Once two clusters were merged to a new cluster the distances to all other clusters must be recomputed. This is usually implemented using the following recursive formula:

$$d(C_m^{(v-1)}, C_k^{(v)} \cup C_l^{(v)}) = a_k d(C_m^{(v)}, C_k^{(v)}) + a_l d(C_m^{(v)}, C_l^{(v)}) + \beta d(C_k^{(v)}, C_l^{(v)}) + \gamma |d(C_m^{(v)}, C_l^{(v)}) - d(C_m^{(v)}, C_k^{(v)})|$$

where the parameters depend on the cluster distance measure. The values of the update parameters are summarised in table 7.4.

Hierarchical methods have been applied in the context of clustering gene expression profiles (Alon *et al.*, 1999; Eisen *et al.*, 1998; Wen *et al.*, 1998). They are memory intensive with increasing data size because all pairwise distances must be calculated and stored. Hierarchical methods suffer from the fact that they do not ‘repair’ false joining of data points from previous steps, indeed they follow a determined path for a given rule.

Another problem with hierarchical clustering methods is that it may be difficult to decide on a representative member for each cluster, especially when using the single-linkage algorithm. In contrast when a partitioning algorithm is used, the centre of each cluster is a natural representation of the cluster’s feature.

7.5.4.2 Self-organising maps

Clustering methods are implicitly used in the construction of self-organising maps (SOMs), a method in the neural network framework introduced by Kohonen (1997).

Teuvo Kohonen’s algorithm tries to find an illustrative display of N -dimensional data points in a given lattice, L , of points, usually in two or three dimensions such that the high dimensional data structure (neighbourhoods, topological ordering, clusters) are preserved and can be detected in this low dimensional structure. The points $r_j \in L$ are called nodes (neurons). Each node r_j has a representation in the N -dimensional space of the data points; this representation is called refer-

1. Randomly initialise the reference vector, $c_j^{(1)}$, for each node.
2. For each iteration step $v + 1$ do:
 - a) Randomly pick an input data vector, x_{v+1} . Denote by $c_j^{(v)}$ the weight vector of the j^{th} node at iteration v . The matching node is defined by $c_{j_0}^{(v)} \in \arg \min \{d_2(x_{v+1}, c_j^{(v)}); j\}$ where $d_2(x, y)^2 = \sum_{n=1}^N (x_n - y_n)^2$ for two N -dimensional data points x and y .
 - b) Update the reference vector of the matching node and its neighbours by the formula: $c_{j_0}^{(v+1)} = c_{j_0}^{(v)} + \eta^{(v)} h_{j_0 j}^{(v)} (x_{v+1} - c_{j_0}^{(v)})$.
3. Assign each data vector to the cluster with the most similar reference vector.

Figure 7.8 Self-organising map clustering algorithm. $0 < \eta^{(v)} < 1$ is called the learning function, that monotonically decreases with the number of iterations; $0 < h_{j_0 j}^{(v)} < 1$ is called the neighbourhood function, that decreases monotonically with the distance of the nodes.

ence vector, c_j , (or weight vector of the neuron). Basically, there are two main steps which are repeated for each data vector for a number of iterations, in the order of tens of thousands iterations (figure 7.8).

The main task of the neighbourhood function is to provide learning, *i.e.* updating of the weights, not only for the best matching node, but also for its neighbours. The task of the learning function is to shrink learning in time as iterations increase.

The result of Teuvo Kohonen's algorithm is that units that are spatially close tend to develop similar weight vectors. Of course, the rate at which the neighbourhood shrinks is critical. If the neighbourhood is large and it shrinks slowly, the cluster centres will tend to stick close to the overall mean of all of the samples.

Commonly used neighbourhood functions are

$$h_{j_0 j}^{(v)} = e^{-\frac{d_2(r_{j_0}, r_j)^2}{2\sigma^2(v)}} \quad \text{and} \quad h_{j_0 j}^{(v)} = \begin{cases} 1, & d_2(r_{j_0}, r_j) < \sigma(v) \\ 0, & d_2(r_{j_0}, r_j) \geq \sigma(v) \end{cases}$$

where r_{j_0} is the matching node, r_j is the adapted node whose reference vector is updated and $\sigma^2(v)$ is the neighbourhood radius that decreases as the number of iterations increases. Self-organising maps have been used in the context of clustering gene expression profiles in Tamayo *et al.* (1999) and Törönen *et al.* (1999).

7.5.4.3 K-means

K-means algorithms are a fast clustering method that is suitable for large-scale applications. The main idea behind these techniques is the optimisation of an objective function usually taken up as a function of the deviates between all patterns of the data points from their respective cluster centres. The most commonly used optimisation is the minimisation of the within-cluster sum of squared Euclidean distances utilising an iterative scheme which starts with a random initialisation of the cluster centres, then alters the clustering of the data to obtain a better value of the objective function.

1. Start with an initial partition of the data points in K clusters with cluster centres $c_1^{(1)}, \dots, c_K^{(1)}$ and let $W^{(1)}$ be the value of the initial objective function.
2. At the v^{th} step of the iteration, assign each data point to the cluster with the lowest pairwise distance.
3. Recompute the cluster centres $c_1^{(v+1)}, \dots, c_K^{(v+1)}$ by minimising $W^{(v+1)}$.
4. If for all k , $c_k^{(v+1)} = c_k^{(v)}$, stop; otherwise goto step 2.

Figure 7.9 K-means algorithm. See text for details.

K-means algorithms alternate between two steps until a stopping criterion is satisfied. These steps are a pairwise distance measure of the data vectors and the cluster centres related to the optimisation criterion and an update procedure for the cluster centres.

Euclidean distance has been used in most cases as pairwise similarity measure because of its computational simplicity: The cluster centre at each iteration can be calculated in a straightforward manner by the arithmetic mean of the data vectors currently assigned to the cluster which is known to minimise the within-cluster sum of squared Euclidean distances. The original K-means algorithm is shown in figure 7.9.

If the pairwise distance is defined as the Euclidean distance, the algorithm minimises the within-cluster sum of squares of the K clusters if the cluster centres at each iteration are recomputed as the arithmetic means of the respective data points. Other pairwise distance measures found in the literature (Jain & Dubes, 1988) are the l_1 -metric (K-median clustering) and the l_∞ -metric (K-midranges clustering). Common criticism on K-means algorithms focuses on the fact that the number of centres has to be fixed from beginning of the procedure. Furthermore, the results are highly dependent on the initialised set of centres. Alternative algorithms have been published that do not require to determine the number of clusters in advance and thus overcome this criticism (Tavazoie *et al.*, 1999; Herwig *et al.*, 1999).

7.5.4.4 Gene shaving

The algorithms described above have a geometric interpretation in that clusters are defined in topological terms such as neighbourhood. The criterion of joining two data points to a cluster is based on geometric distance and N -dimensional balls that are centred at the average expression vector of the data points assigned to that cluster. Another way of interpreting a cluster is the concept of variance. Given a sample of M data points, a cluster can be defined as a subset of data points explaining a reasonable amount of variance in the data incorporating principal component analysis. The main algorithm has the outline shown in figure 7.10.

The shaving process requires repeated computation of the largest principal component of the remaining data points from the previous iteration steps. It removes then those genes with smallest projections on these components. The proposed shaving method searches for clusters of genes showing both high variation across the samples and correlation across the genes. Gene-shaving requires the decision

1. Center the rows of the $M \times N$ gene expression matrix, X , to have zero mean.
2. Compute the leading principal component of the rows of X .
3. Shave off the proportion α (about ten percent) of the rows having smallest dot-product with the leading principal component.
4. Repeat steps 2 and 3 until one gene remains.
5. From the sequence of nested gene clusters $S_M \supset S_k \supset \dots \supset S_1$, here S_k is a cluster of k genes, estimate the optimal cluster size, $\hat{k}$, using a statistical measure.
6. Orthogonalise the rows of X with respect to the average vector of the cluster $S_{\hat{k}}$.
7. Repeat steps 1 to 5 with the orthogonalised data until a maximum prefixed number of clusters are found.

Figure 7.10 Gene shaving algorithm. See text for details.

which set of genes to use for choosing the cluster in step five of the algorithm. This can be done using the Gap-statistic (Hastie *et al.*, 2000) that is based on classical analysis of variance arguments (ANOVA). The authors applied this algorithm to the B-cell lymphoma data of Alizadeh *et al.* (2000) where they find groups of genes that are highly discriminative for the two molecular subtypes.

7.5.5

Evaluation of gene expression clusters

Many clustering algorithms are in use, each of which claims special merits and has some interpretation that makes it suitable for a class of applications. However, in general, one wishes to compare the output of cluster algorithms in order to decide which one gives best results for the problem under consideration. Here, cluster validation measures can be used. In principle, two groups of measures, external and internal, can be separated (Jain & Dubes, 1988). External validation measures incorporate *a priori* knowledge on the clustering structure of the data, for example, in simulation experiments when the true partition of the data is known, or in real experiments when specific gene clusters are known. Typically, an external cluster validation measure is a numerical function on the corresponding cluster partitioning matrices that are derived from the true and the calculated partitions of the data. Examples are the Jaccard coefficient, Hubert's F statistic or the goodness-of-fit statistic (Jain & Dubes, 1988; Mirkin, 1996).

Internal measures compare the quality of the calculated clusters solely by use of the data itself. Indices of quality are topological concepts, for example compactness or isolation, that are computed by numerical functions, concepts from information theory that quantify, for example, high informative clusters and variance concepts that quantify the overall variance explained by the cluster. Another approach is to calculate figures of merit (Yeung *et al.*, 2001) for each cluster. Here, the predictive power of the clustering is quantified in a straightforward fashion by omitting one condition from the gene expression matrix, clustering the gene expression profiles

according to the remaining $N-1$ conditions and calculating the mean deviation of the patterns of the data points and the respective cluster centroids according to the condition which has been left out.

Cluster validation is an important topic that has drawn insufficient attention in gene expression analysis. Currently, the user of any software package is left with a number of algorithms that stand fairly loosely in parallel. In the future, however, research will focus on the comparison and integration of different methods in order to reduce the bias of the individual methods.

7.5.6

Conclusion

The research in clustering gene expression profiles has been increasing in recent years in step with the increase in the number of publications on micro and macro-array studies. Currently, a number of algorithms are available, each of which has its own advantages and disadvantages. Future research will focus on the integration and evaluation of these algorithms in order to reduce the individual bias of each method. Future developments will also concentrate to a far greater extent on the biological context of experiments. For example, the increasing amount of information about regulatory pathways and regulatory interactions of genes will be incorporated in gene expression profile clustering and cluster validation. This provides a reasonable amount of training data that can guide data analysis to biologically meaningful results.

7.6

Promoter finding in eukaryotic genomes

7.6.1

Introduction

Genes in higher eukaryotes are regulated by the multiple regulatory transcription factors, acting through specific regulatory binding sites that are usually located in the proximity of the genes when constituting a promoter, or at more remote locations when part of an enhancer. With the availability of genomic sequences and massive though phenomenological gene expression data, the challenge is to understand the regulatory mechanisms of every gene in the genome by computer analysis of the regulatory sequences and by integrating this data with biological knowledge of the signal transduction, metabolic and physiological networks involved in regulation of that gene. Sophisticated computational regulatory sequence analysis tools that employ powerful statistical and machine-learning algorithms driven by the rich databases that collect known biological facts enable us to make profound *in silico* predictions and formulate experimentally testable hypotheses. Such *in silico* driven experiments can greatly speed up the process of understanding of the gene regulatory mechanisms and identification of new target genes. The understanding

of how gene regulation mechanisms are encoded in the genomic regulatory sequences offers a powerful means for deciphering the causes of major human diseases.

7.6.2

Transcription regulation in eukaryotes

Like the other eukaryotic RNA polymerases, RNA polymerase II depends on the presence of numerous additional proteins in order to recognise its promoters and to transcribe the attached genes under well-defined spatio-temporal and circumstantial conditions. Some of these additional proteins, transcription factors and co-factors, are specific DNA-binding proteins, others attach to DNA-bound protein complexes. The total complex thus provides an environment for the polymerase to act in a very specific manner (see chapter 1).

7.6.2.1 Promoter structure

A promoter can be defined as the structural part of a gene that defines its transcription start point and mediates and controls the initiation of transcription. Promoter sequences may include a TATA box, the initiator region (INR), upstream activating elements, and downstream elements. However, not all of these elements are required in every case. This means that the DNA pattern that defines a so-called core promoter (around -50 to +10 base pair around the transcription start site) may vary significantly in different cases and can be divided into four classes (Bajic *et al.*, 2002; Berg & Hoppel, 1987 and 1988). Briefly, these classes comprise: (i) core promoters which consist of a TATA box only that directs transcriptional initiation at a position about 30 base pairs downstream; (ii) core promoters which do not contain any TATA box (and, therefore, are referred to as TATA-less promoters); in these promoters, the exact position of the transcriptional start point is controlled mainly by an initiator element, called Inr (Smale, 1994 and 1997); (iii) composite promoters consisting of both a TATA box and an initiator element; and (iv) so-called Null-promoters, which have neither a TATA box nor an initiator and rely exclusively on upstream and downstream promoter elements (DPE) (Novina & Roy, 1996; Smale, 1994).

Thus, it is clear that a variety of promoters rather than one single type of core promoters exist. This is a major reason why promoter recognition in eukaryotes is so difficult.

7.6.2.2 Transcription factors

In addition to a considerable number of general transcription factors that constitute the basal transcription complex, a multitude of transcriptional activator and repressor proteins are involved in transcriptional regulation (Wingender, 1993). These bind in a sequence-specific manner to DNA sites that are either adjacent to the core promoter or several kilobases upstream or downstream of the transcrip-

tion start site (TSS). Some distal sequences called enhancers may influence transcription without respect to their positioning or orientation (Rippe *et al.*, 1995, Thompson & McKnight, 1992). Transcription factors achieve transcriptional activation by recruitment of the basal transcriptional machinery to the promoter through protein/protein interactions, either directly or through adaptor proteins (Ptashne & Gann, 1997).

The subset of transcription factors bound to the promoter of a gene ensures that this gene is transcribed in the proper cell type(s) at the proper time during development, or in response to an extracellular signals such as, for instance, blood sugar level or viral infection. Transcription factor-binding sites in promoter regions have no obvious patterns with respect to location or orientation. Besides transcription factor-binding sites, other elements or sequence properties exist, only in vertebrates, that are associated with promoter sequences. These include areas of increased cytosine and guanine content, so-called CpG islands (Shago & Giguere, 1996).

7.6.2.3 Combinatorial nature of transcription regulation

In multi-cellular organisms, the number of possible intracellular molecular states is very large. The complexity of gene expression at a transcriptional level is a result of combinatorial regulation. Combinatorial regulation of transcription is organised through binding of a multiplicity of transcription factors to their target sites, termed *cis*-elements, in regulatory regions. These combinations constitute the specificity of the regulation of a certain gene or a group of genes. The combinatorial properties of regulatory sequence signals can be considered in the following hierarchy:

- (i) The lowest level of hierarchy corresponds to the ***cis*-regulatory elements** (five to twenty-five base pairs in size) that bind definite transcription factors. More than 3,000 binding sites for various transcriptional factors are known to date.
- (ii) **Composite regulatory elements** correspond to the second hierarchical level. They are formed by specific combination of individual *cis*-elements that jointly act as one functional unit (Kel *et al.*, 1995a).
- (iii) **Promoters and enhancers** are assigned to the third hierarchical level in the structural organisation of transcription regulation. Enhancers modulate the level of transcription depending on the type of tissue, the developmental stage, the stage of the cell-cycle, induction by hormones or other molecular signals.
- (iv) **Transcription regulatory regions** consist of several elements of the preceding levels (individual *cis*-elements, composite elements, enhancers and promoters). The sum of all the regulatory genomic sequences makes up the system of integral regulation of gene transcription, which is the highest level of hierarchy.

7.6.3

Databases on transcriptional regulation

Information about regulatory sequences of eukaryotic genomes has been collected in the databases on transcriptional regulation. Some of which are described in the following sections.

The **Eukaryotic Promoter Database (EPD)** contains general information about promoters, as they are defined by an experimentally proven transcription start site, and their tissue-specificities (Perier *et al.*, 1998, 1999 and 2000; Praz *et al.*, 2002); <http://www.epd.isb-sib.ch>).

The **Transcription Factor Database (TFD)** – and its successor the Object-oriented Transcription Factors Database (ooTFD) – is a specialised database focusing on transcription factors and their properties (Gosh, 1992, 1993, 1998 and 2000; <http://www.ifti.org/>). This database is aimed at capturing information regarding the polypeptide interactions which comprise and define the properties of transcription factors. ooTFD contains information about transcription factor-binding sites, both eukaryotic and prokaryotic, as well as composite relationships within transcription factors which frequently occur as multi-subunit proteins that form a complex interface to cellular processes outside the transcription machinery through protein/protein interactions.

The **TRANSFAC database** provides information on structure, function, sequence, DNA-binding profiles, and classification of transcription factors. TRANSFAC also contains information on binding sites for transcription factor within genes (Heinemeyer *et al.*, 1998; <http://www.gene-regulation.com/pub/databases.html#transfac>). An entry in the TRANSFAC database corresponds either to a transcription factor or to an individual transcription factor-binding site. In the current release 6.1, TRANSFAC contains information on 4,822 transcription factors and more than 11,900 binding sites, among which are about 5,700 genomic binding sites. Five-hundred and fifty-eight nucleotide distribution matrices of aligned binding sequences are also provided.

The **Transcription Regulatory Region Database (TRRD)** collects information on the hierarchical structure of whole regulatory regions of eukaryotic genes and on gene expression patterns (Kolchanov *et al.*, 1999). Each TRRD entry corresponds to an entire gene, and the genes in TRRD are classified into several functional groups. For instance, one of these functional groups are the genes that regulate the cell-cycle regulatory genes (<http://www.bionet.nsc.ru/trrd/celcyc/>).

COMPEL (or its recent variant TRANSCompel) collects information about composite regulatory elements which are pairs of closely situated binding sites, and thus reflecting the combinatorial nature of transcription regulation (Kel-Margoulis *et al.*, 2001a; <http://www.gene-regulation.com/pub/databases.html#transcompel>). The actual release 6.1 of the COMPEL database contains 322 composite elements.

7.6.4

***In silico* study of gene transcription regulation**

The rapid growth of sequence data for ever more newly discovered genes and their regulatory regions opens up a new source of information. However, the pure sequence is of little value if not accompanied by functional data. For this reason, computational approaches that provide experimentally testable hypotheses about possible mechanisms of transcriptional regulation of genes on the basis of sequence data are urgently needed. Such computationally predicted regulation properties of newly discovered genes may help to localise them within complex gene regulatory networks and thus provide valuable information on the gene network control of cellular developmental processes.

Several thousands transcription factors operate in human cells. More than 800 of these are well studied and characterised in databases such as SWISS-PROT and TRANSFAC. Although TRANSFAC contains information on more than 5,000 genomic sites (about 1,500 of which lie in human genes), it is far from being complete. Assuming that the human genome contains about 33,000 genes and that each gene might have up to 100 functioning sites in all regulatory regions including promoters, enhancers and far upstream regulatory regions, we may expect millions of regulatory sites within the human genome. Knowledge about the number, the sequence and the position of all these sites in the genome should provide a new level of understanding on how genes are regulated during the development and functioning in the organism and on how genes are dis-regulated in disease.

7.6.4.1 Recognition of *cis*-regulatory elements

Computer analysis of genome sequences provide means for prediction of binding sites for different transcription factors. Every transcription factor can be characterised by a specific DNA motif that is common for most of their binding sites. Different methods of motif analysis have been developed for recognition of individual *cis*-regulatory elements.

Pattern search. In the early stages of analysis of binding sites for a new transcription factor, only a few examples of known sites for this factor are available. The first step is to use them as patterns in the search for new potential bindings sites for the transcription factors under consideration. Since information on variability at each base position in the pattern is limited, simple pattern search methods usually allow a certain percentage of mismatches at any position in the pattern. While more examples of the known sites become available, consensus are created that describe the observed variability of nucleotides at different positions of the sites using the ambiguous letter code (*e.g.*, the letter S marks a position where both nucleotides guanine or cytosine are equally observed, W corresponds to either adenine or thymine and so on). Pattern search tools such as SIGNAL SCAN (Prestridge, 1991; Prestridge & Stormo, 1993; <http://bimas.dcrf.nih.gov/molbio/signal/index.html>) or PatSearch (Wingender *et al.*, 1996; <http://transfac.gbf.de/cgi-bin/patSearch/patsearch.pl>), SITE (Solovyev & Rogozin, 1986; <http://www.softberry.com/>) can

use consensus in the search and assign different mismatch penalties to the different letters of the ambiguous code. Pattern search methods are quite useful for finding new potential sites but they are rather inaccurate and characterised by relatively high false-negative and false-positive rates.

Weight matrix method. Weight matrices are used to describe highly degenerate transcription factor-binding sites. The basis of any weight matrix is the counts for the observed nucleotides in the corresponding position of known sites. To build a reliable weight matrix one needs a collection of a few known binding sites for a certain transcription factor (figure 7.11). The most up-to-date collection of weight matrices is presented in the TRANSFAC database which is available at <http://www.biobase.de>. Several weight matrix-based search programmes have been developed, for example, ConsInspector (Frech *et al.*, 1993 and 1997; <http://www.gsf.de/biodv/consinspector.html>), MATRIX SEARCH (Chen *et al.*, 1995; <http://www.cbs.umn.edu/software/sigmatrix.html>) or MatInspector (Quandt *et al.*, 1995; <http://www.genomatix.de/cgi-bin/matinspector/matinspector.pl>), Match (Gössling *et al.*, 2001; <http://www.gene-regulation.com/pub/programs.html#match>) and TRANSPLOERER (<http://www.biobase.de>). Different programmes use different formalism for the calculation of weight matrices from the nucleotide counts based on thermodynamic considerations (Berg & von Hippel, 1987; Berg & von Hippel, 1988), information theory (Stormo, 1998); pseudo-counts (Henikoff & Henikoff, 1996) and forbidden nucleotides (Tronche *et al.*, 1997). The search algorithms are greatly speeded up by using hash tables (Gössling *et al.*, 2001). These tools have been applied intensively in the last years for the analysis of regulatory regions of many different functional classes of genes. These include the globin genes (Hardison *et al.*, 1997), muscle and liver-specific genes (Wasserman & Fickett, 1998) and genes involved in the regulation of cell-cycle (Kel *et al.*, 2001a).

CATGAGTCAGA														
AGTGACTCAGC														
- CTGCCTCAGC														
GATGAGTCAAG														
TCTGACTAAAA														
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓														
A	1	2	0	0	4	0	0	1	5	2	2	Positional weight matrix		
C	1	2	0	0	1	3	0	4	0	0	2			
G	1	1	0	5	0	2	0	0	0	3	1			
T	1	0	5	0	0	0	5	0	0	0	0			
N M T G A S T C A R M												Consensus		

Figure 7.11 Generation of a nucleotide distribution matrix from aligned binding sites. A set of proven transcription factor-binding sites is aligned and the frequency of each nucleotide is counted to form a positional weight matrix (PWM). From such matrices, information pro-

files may be calculated and used for the identification of potential new transcription factor-binding sites. A consensus sequence in the 15-letter IUPAC code is derived that may be used for sequence analysis albeit at a much lower accuracy than with conventional PMW.

Matrix methods have been quite successful for recognition of binding sites of many transcription factors such as cAMP response element-binding protein (CREB), activator protein 1 (AP1), E2F, but have not been very useful in recognising target sites for some other factors. A lot of efforts have therefore been made to improve methods for the recognition of individual *cis*-elements.

Correlation between positions. In the classical weight matrix method, positions are considered as independent and their contribution to binding energy is considered in an additive fashion (Berg & von Hippel, 1987). However, neither of these assumptions is quite correct for the binding sites of transcription factors. To deal with the non-additive effect of different positions, an information measure has been suggested that assigns higher weights to the more conserved positions (Quandt *et al.*, 1995).

To consider the correlation between neighbouring site positions, hidden Markov model (HMM) approaches were applied (Ehret *et al.*, 2001). Hidden Markov models calculate the probability of nucleotides in a certain position in dependence on the preceding nucleotide(s). Another approach is based on dinucleotide and trinucleotide matrices that consist of frequencies of all duplets or triplets occurring at a given position (Kondrakhin *et al.*, 1994). Long distance correlation between nucleotides in non-neighbour positions of sites may also contribute to the recognition of sites (Kel *et al.*, 1995b).

Dinucleotide weight matrices have been applied to the recognition of a vast number of transcription factor-binding sites (Kel *et al.*, 1995b). In most cases, the dinucleotide matrices showed sufficiently better recognition accuracy than mononucleotide matrices constructed under the same conditions. Dinucleotide matrices are able to take into account the correlation among neighbouring site positions, thus providing better recognition.

Flanking sequences. Weight matrices capture position-specific preferences in a short region only, which often corresponds to the most conserved part of the site, but fail to recognise regularities within flanking regions. Therefore, new methods have been developed that take flanking regions of the sites into account. The programme ConsInspector makes a pair-wise alignment of a sequence to the set of sites that includes the core of the sites as well as the flanking regions (Frech *et al.*, 1993). Similarities found in the flanks are used for site recognition. Various artificial intelligence approaches have been used to develop new methods for recognition of *cis*-elements in order to reveal additional features in the flanking regions of sites. For example, an application of the perception method has been developed for the recognition of gene regulatory elements (Seledtsov *et al.*, 1991). This creates a model that includes 30 base pairs of each side of the gene regulatory element sites.

The pattern recognition software SITEVIDEO (Kel *et al.*, 1993) has been used to build recognising programmes for transcription factor-binding sites (Kel *et al.*, 1995b). SITEVIDEO permits analysis of a set of sites including their flanking regions and permits design of high accuracy recognition methods by using a number of standard multi-dimensional discriminating methods and approaches from the field of artificial intelligence including perceptron and neural networks (Bolch &

Huang, 1974; Minsky & Papert, 1969). In addition, a new approach for the visual monitoring of recurrent design of the recognition methods was developed (Kel *et al.*, 1993). Recently, this technique was used for building a new recognition method for binding sites for E2F transcription factors – the key regulators of the cell-cycle. An exhaustive search for contextual motifs was made in the flanking regions of these sites. On the basis of over-represented motifs found a ‘score of context’ was build that is used in addition to the weight matrix search. This new method of recognition of the E2F sites provided five times more accurate predictions than the conventional weight matrix method (Kel *et al.*, 2001a).

Subsets of sites. Consideration of the heterogeneity of functional sites led to the development of an approach in which each site is allocated to a set of typical consensus sequences (Kolchanov *et al.*, 1991). This method provides a possibility to describe sites that serve as targets for heterogeneous sets of transcription factors. This situation is typical for most types of transcription factors when a family of homologous factors can bind to binding sites that are similar but that have slightly different sequence specificities. Thus, the set of binding sites can be divided into overlapping subsets of sites most favoured for binding of a certain transcription factor. A new method based on kernel function estimations was developed for revealing such subsets of binding sites (Tikunov *et al.*, 2000). Every subset is characterised by a specific weight matrix. A similar approach based on a set of examples was successfully applied to the recognition of transcription factor-binding sites in FunSite (Kel *et al.*, 1995b).

Thus, there are many methods for the automated recognition of transcription factor-binding sites. However, the best approaches need to be trained on sample data, a process which has up to now only been possible for the limited number of transcription factors for which many binding sites are known.

Widely used searches for individual transcription factor-binding sites in DNA sequences (for review, see Frech *et al.*, 1997) cannot be applied directly to the characterisation of transcriptional regulation of genes for three main reasons: (i) the poor recognition ability of most programmes currently used; (ii) the incomplete lists of transcription factor-binding sites being searched; (iii) even if individual sites are found correctly, this does not define the specificity of transcriptional regulation of a gene because of the combinatorial nature of this regulation.

7.6.4.2 Recognition of composite regulatory elements

Combinatorial regulation is the basic mechanism by which gene expression is controlled in eukaryotes. The pattern of expression of eukaryotic genes is generally encoded in the structure of their transcription regulatory regions in the form of a combination of transcription factor-binding sites. Over the last few years, several computational approaches addressing the problem of combinatorial regulation of transcription have appeared. Specific transcription factor-binding site combinations were used for the identification of muscle-specific promoters (Frech *et al.*, 1998; Wasserman & Fickett, 1998) for liver-enriched genes (Tronche *et al.*, 1997) and for yeast genes (Brazma *et al.*, 1997).

The first and most important level of this combinatorial regulation is provided by composite regulatory elements – combinations of target sites for two different transcription factors that interact with each other resulting in a particular expression pattern common for genes containing this composite regulatory element (Kel *et al.*, 1995c). A number of known examples of composite elements have been collected in the TRANSCompel database (Kel-Margoulis *et al.*, 2001a).

Search for composite elements. Several tools have been designed to search for composite elements in nucleotide sequences. With the programme FastM (Frech *et al.*, 1998; <http://www.genomatix.de/cgi-bin/fastm2/fastm.pl>), the user can specify a pair of weight matrices and set a preferable distance range between corresponding sites in the composite element. Another tool, Catch (<http://www.biobase.de>), searches for composite elements using a pattern matching approach. Composite elements from the TRANSCompel database are used as patterns for the Catch programme. The user can specify the amount of variability in the distance between two sites and their degree of conservation.

Statistical estimations have been made to reveal pairs of transcription factor-binding sites in nucleotide sequences. Application of χ^2 statistics may reveal pairs of sites that are often found close to each other (Kel *et al.*, 1995c). Using this tool, unknown pairs, such as cAMP response element-binding protein and SP1 (CREB/SP1), GATA-binding protein and nuclear factor κ B (GATA/NF- κ B), have been revealed as potential new composite regulatory elements. Another statistical technique was used more recently for the identification of pairs of motifs that may correspond to new composite elements (GuhaThakurta & Stormo, 2001).

A method called ‘composite score’ was developed for revealing nuclear factor of activated T-cells and activator protein 1 (NFAT/AP1) composite elements (Kel *et al.*, 1999). This uses two matrices for two corresponding transcription factors. The range of allowed distances between matrix matches and their mutual orientation are taken into account, as well as coordinate variation of the matrix scores for these two factors. A low score of one matrix is compensated by a high score of another matrix, thus providing an optimal binding energy for the protein/DNA complex on the composite element.

Recently, it was shown that searching for composite elements is a very effective tool for predicting gene expression patterns. NFAT/AP1 composite elements were found in high concentration in the promoters of genes that are induced upon immune cell activation. Clusters of these composite elements might therefore serve as a landmark for identifying promoters of this type (Kel *et al.*, 1999). Promoters of genes regulated during the cell-cycle were recognised using a combination of E2F-binding sites and additional potential target motifs for other transcription factors (figure 7.12) (Kel *et al.*, 2001a).

Clusters of sites. Transcription factor-binding sites tend to cluster in promoters and in other genomic regulatory regions. Clusters of binding sites for the same transcription factors (so called homo-clusters) are believed to increase the probability of binding of the factors to their target regions in the genome. Statistical estimations made by Karlin & Macken (1991) have been used to reveal statistically significant homo-clusters of transcription factor-binding sites in nucleotide sequences

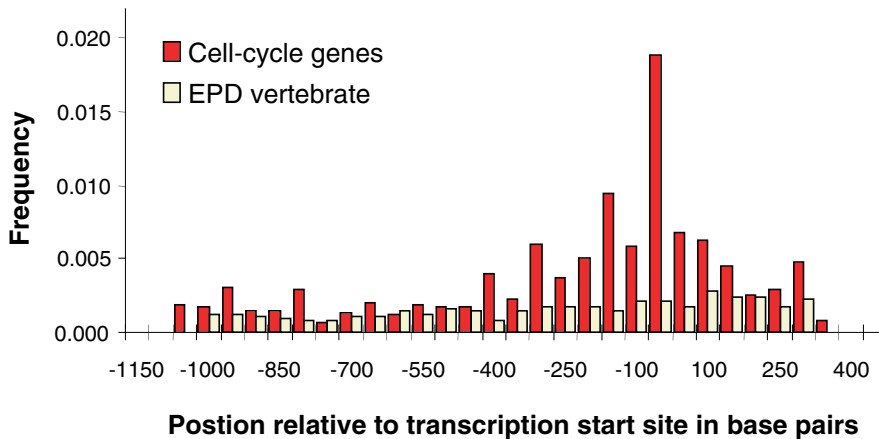


Figure 7.12 Frequencies of putative E2F-binding sites in cell-cycle genes. Potential E2F sites are found in promoters of cell-cycle genes with much higher frequency than in all other promoters stored in the Eukaryotic Promoter Database (EPD). The numbers on the x-axis refer to the interval relative to the transcription start site, for example, -250 refers to the E2F-binding sites found in the window between -200 and -250 relative to the transcription start site.

(Kel *et al.*, 1999). An option has been implemented in the tool Match to reveal such clusters. By applying this tool, clusters of E2F-binding sites have been found in the sequences of human chromosomes 21 and 22 (Kel-Margoulis *et al.*, 2001b).

Recently, a new statistical technique appeared that is suitable for the identification of statistically significant homo-clusters and hetero-clusters (Wagner, 1999). This technique was applied to the transcription factors *Mcm1* and *Stel2*, which are involved in the cell-cycle and mating control of the yeast *Saccharomyces cerevisiae*. Clusters of binding sites for these factors were revealed in the yeast genome (Wagner, 1999).

A tool called Cister (Cis-element Cluster Finder) was developed for revealing hetero-clusters (Frith *et al.*, 2001; <http://zlab.bu.edu/~mfrith/cister.shtml>). This uses a hidden Markov model-based method for searching for clusters of cis-elements. The clusters found are used as landmarks for detecting promoters and other regulatory regions in DNA sequences. According to the estimation given by the authors, the programme achieves a sensitivity of promoter predictions of 67 percent, while making one prediction per 33 kilobases of non-repetitive human genomic sequence. A web interface is available at <http://sullivan.bu.edu/~mfrith/cister.shtml>.

To reveal site clusters, correct cut-offs must be set for the weight matrices used to search for transcription factor-binding sites. It is known that some binding sites that are part of composite elements often differ significantly from the consensus. In such cases the lack of binding energy for a transcription factor is compensated by protein/protein interactions with other factors. The programme ClusterScan was developed to find such cryptic sites as parts of hetero-clusters (also called composite clusters) (Kel *et al.*, 2001b). This programme uses a genetic algorithm

to define optimal parameters to search for composite clusters of transcription factor-binding sites in regulatory genomic sequences. The composition of the identified clusters may reveal a lot about regulatory mechanisms and may provide a means for functional classification of regulatory regions and for their automatic annotation in genomes (Kel-Margoulis *et al.*, 2001b).

7.6.4.3 Recognition of promoters

Computer-assisted prediction of eukaryotic promoters is one of the most straightforward approaches for the analysis of transcription regulation using sequence data (Bucher, 1990).

In 1997, James W. Fickett and Artemis G. Hatzigeorgiou reviewed promoter recognition programmes. The recognition quality of nine different tools was compared using a test set of promoters. Since that time, a lot of new tools for promoter prediction has appeared. We have used the same test set that was used in the review mentioned above and compared results of promoter prediction of the old tools with the newer versions of these tools as well with some new tools.

Basically, two different approaches have been applied for promoter recognition. The first approach is purely sequence-based and makes use of different statistics of nucleotides and oligonucleotides in the promoter sequences without considering known promoter signals and *cis*-elements. Specific features of the distribution of nucleotides and oligonucleotides of different length (dinucleotides, trinucleotides and longer oligonucleotides) are revealed in the training set of known promoters and these features are used then to build recognition procedures. In most of the applications that use the sequence-based approach, the authors did not pay much attention to the biological meaning of the oligonucleotides found or to their distribution in promoter sequences.

By contrast, the signal-based approach uses features and signals such as TATA boxes, CAAT boxes, consensus transcription factor-binding sites, physical and chemical properties of DNA that are known to have high biological significance for promoter structure.

Both approaches have their merits. Usage of known signals makes much more sense than simple oligonucleotide counting since it may capture crucial features of promoter structure that may be missed by the pure oligonucleotide-based approach. On the other hand, oligonucleotide analysis provides a general sketch of all possible binding sites. Some of these may be unknown and therefore missed by procedures that are based on known sites only.

First tools that belong to the sequence-based approach are described in the following sections.

Autogene Promoter. The first programme for recognition of eukaryotic promoters was created in 1992 (Kel *et al.*, 1993b) based on distinctive features of dinucleotide distribution in a sample of promoters derived from the Eukaryotic Promoter Database. This programme does not show a high degree of recognition accuracy, but it was nevertheless included into the Autogene package for gene prediction and annotation of nucleotide sequences (Ptitsyn *et al.*, 1996). The utility of this pro-

motor prediction programme was improved by combining it with other programmes for identification of binding sites as well as recognition of exons.

PromFind is a programme that uses oligonucleotide distributions for promoter prediction (Hutchinson, 1996; <http://iubio.bio.indiana.edu/soft/molbio/mswin/mswin-or-dos/profin11.exe>). The algorithm operates by scoring the sequence using a differential hexamer measure. The author suggests searching for a peak score independent of an absolute threshold because this is more accurate than a search based upon a fixed scoring threshold.

Neural Network Promoter Prediction (NNPP) programme. The Neural Network Promoter Prediction programme (Reese & Eeckman, 1995) applies several neural networks for analysis of nucleotide composition of the core promoter region that includes the transcription start site and TATA box. The programme is available at http://www.fruitfly.org/seq_tools/promoter.html. A rather short region (−40 to +11 around the transcription start site) comprising the core region of promoters only is used for recognition.

A number of promoter prediction algorithms were developed that use different variants of neural networks and similar techniques. Among these the most recent programmes are described in the following sections.

Promoter 2.0. Promoter 2.0 is a new version of a tool described by Fickett & Artemis (1997) as GeneID (<http://www.cbs.dtu.dk/services/promoter>). The Promoter 2.0 recognition algorithm is a combination of neural networks and genetic algorithms. The neural network uses either a DNA sequence, or the output of other neural networks as input. Genetic algorithms are used for optimisation of the weights in the neural networks in order to discriminate maximally between promoters and non-promoters (Knudsen, 1999). The program recognises a region between 200 to 300 base pairs surrounding the transcription start site are entered. Thus, this program characterises the core promoter region only.

McPromoter V3.0. Version 3.0 of Markov Chain Promoter Finder (McPromoter V3.0) (Ohler *et al.*, 2001; <http://promoter.informatik.uni-erlangen.de/>) is a hybrid method to search for eukaryotic RNA polymerase II transcription start sites. The system contains a promoter model which divides a promoter into six consecutive segments. The model is applied to a window of 300 bases. A neural network is used to combine features of nucleotide context, features describing CpG islands and some selected physico-chemical parameters of DNA. The addition of physical properties of DNA improves promoter recognition. The main features used for recognition by this programme is the closeness to the CpG islands.

CorePromoter Finder. CorePromoter Finder (<http://argon.cshl.org/genefinder/CPROMOTER/>) uses a stepwise strategy based on the initial localisation of an extended promoter of one to two kilobases within a genomic DNA sequence of 100 kilobases or more and further localisation of a transcriptional start site of 50 to 100 base pairs (core promoter) covering the interval −60 to +40 around the transcription start site. The method uses a position-dependent quintuple measure that is analysed with the help of a quadratic discriminant analysis technique (Zhang, 1998a). In addition to the standard training set of the Eukaryotic Promoter Data-

base, the novel Lead Exon Database (LEPD) of 673 sequences constructed by the author on the basis of GenBank annotations was also used (Zhang, 1998b).

PromoterInspector. PromoterInspector (Scherf *et al.*, 2000a) focuses on the genetic context of promoters rather than on their exact location (<http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl>). The programme is based purely on libraries of International Union of Pure and Applied Chemistry (IUPAC) words extracted from training sequences by an unsupervised learning approach. The programme was used for gene prediction in human chromosome 22. Results are available online at http://genomatix.gsf.de/software_services/databases/chr22/intro.html.

FirstEF. This is a programme designed to find promoters and first exons (Davuluri *et al.*, 2001; <http://www.cshl.org/mzhanglab/>). It uses a set of discriminant functions that can recognise structural and compositional features such as CpG islands, promoter regions and first splice-donor sites. The core of the algorithm is a decision tree that process results of this discriminant function. The programme was trained on a large verified training set containing more than 2,000 human promoters. FirstEF was used by the authors to predict promoter and first exon regions for all human chromosomes. The results are available on the website shown.

Dragon Promoter Finder (DPF). The Dragon Promoter Finder algorithm (<http://sdmc.krdl.org.sg:8080/promoter/>) identifies transcription start sites using five independent promoter recognition models (Bajic *et al.*, 2002). Each model uses a data window that slides along the DNA sequence. Based on the competition of three sensors, the Dragon Promoter Finder predicts a potential transcription start site for each window. The three sensors recognise the promoter, exon and intron regions. Sensor outputs serve as inputs to an artificial neuron network which predicts the presence of a transcription start site.

Tools that use the signal-based approach appeared very early. The first well-known programme for promoter prediction was **PromoterScan** (Prestridge, 1995; <http://bimas.dcrtn.nih.gov/molbio/proscan/>). To design the programme, promoter sequences were analysed and density of each mammalian transcription factor binding site was compared with its density in non-promoter sequences. Transcription factor-binding sites were located using a pattern matching algorithm based on the Transcription Factor Database (TFD) (Ghosh, 1991). The combined density ratios of all binding sites were then used to build a scoring profile called the promoter recognition profile. This profile, used in combination with a weighted matrix for TATA boxes, was used to predict locations of transcription start sites. PromoterScan programme also outputs the potential sites that may be used for interpretation of the search results.

TSSG and TSSW. The algorithm used in the programs TSSG and TSSW predicts potential transcription start positions using a linear discriminant function that combines functional motifs and the oligonucleotide composition of promoters (Solovyev & Salamov, 1997; <http://www.softberry.com>). TSSW uses selected factor-binding sites from the TRANSFAC database. The TSSG method is based on the Transcription Factor Database. Special attention is paid to the oligonucleotide composition surrounding transcription start sites.

FunSiteP. The programme FunSiteP (Kondrakhin *et al.*, 1995; <http://compel.bionet.nsc.ru/FunSite/fsp.html>) recognises promoters in nucleotide sequences and tentatively identifies the functional class of these promoters using Bucher's specifications (Bucher, 1990). The method is based on the localisation of transcription factor-binding sites. The authors have developed a weight matrix to localise binding sites based on the uneven distribution of transcription factor-binding sites. Using these matrices, an algorithm for promoter recognition was developed. To increase the accuracy of the method, a routine was added that breaks a promoter sample into sub-samples.

7.6.5

Conclusions

Testing of the promoter prediction tools has demonstrated that nearly all of the available tools have a low level of sensitivity and specificity (Fickett & Hatzigeorgiou, 1997). The accuracy of some of more recently developed tools appears to be higher because larger training sets, better binding site databases and newer, more powerful techniques of machine learning and pattern recognition are used. Overall, recognition accuracy, however, remains fairly low so that these programmes cannot be used for direct annotation of the genome. It is now clear that prediction of a general promoter structure is not just a difficult but perhaps even an impossible task, and that each promoter should be described on the basis of its specific composition of regulatory elements. The most promising trend at present is therefore to identify specific promoter structures that are common for a group of functionally related promoters (*e.g.*, tissue-specific promoters). Identification of composite regulatory elements is now very important for the recognition of promoters and the prediction of their structure.

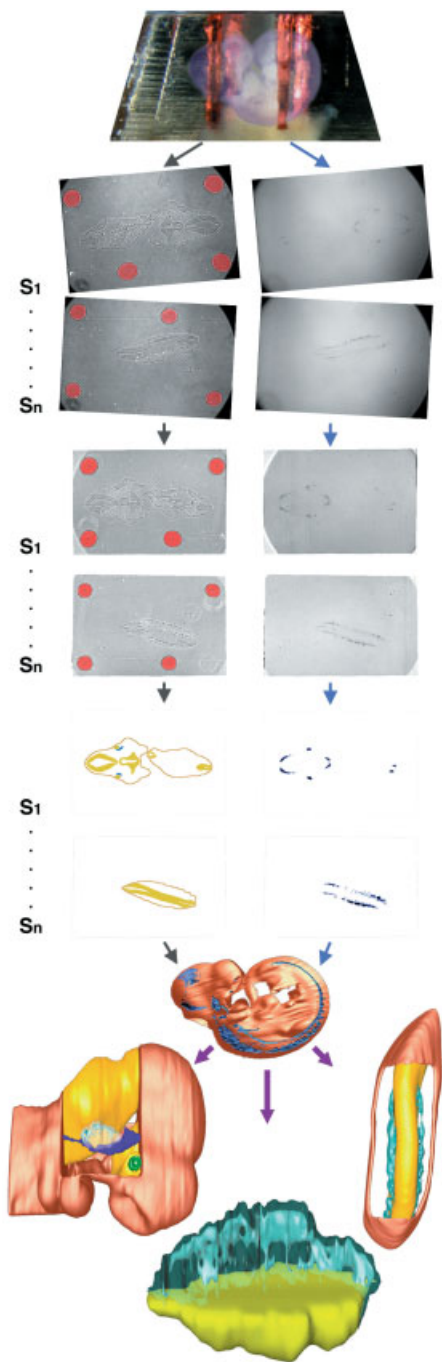
7.7

GeneEMAC – Three-dimensional visualisation of gene expression

7.7.1

Introduction

A central post-genomic question will be how genetic information is translated into the three-dimensional generation of form and function of an organism. Gene expression studies of developing organisms provide snap-shot data on the activation of individual genes at selected locations and single moments. But an understanding of the dynamic activation of multiple genes and their functional role in controlling the three-dimensional processes of collective cell behaviour, pattern formation, and morphogenesis requires special tools for a systematic description of spatio-temporal patterns of gene activation and the ensuing phenotypic effects. Such novel, computational tools will assist in unravelling the functional roles of genes (Skolnick & Fetrow, 2000) and gene networks by providing com-



Specimen preparation

whole mount *in situ* hybridisation
embedding
marker drilling
serial sectioning

Image capturing

phasecontrast image of morphological structures and markers
brightfield image of gene expression signal

External marker based automatic congruencing

Dynamic threshold segmentation

of morphological structures
of gene expression signal

Surface generation & 3D-Visualisation

◀ **Figure 7.13 Schematic illustration of the GeneEMAC protocol.** *Specimen preparation:* The whole-mount *in situ* hybridised specimen (here a Theiler stage 17 mouse embryo embedded and stained for *Scgn 10* gene expression) (Sporle *et al.*, 1998; Stein *et al.*, 1988). Before serial sectioning (478 sections, cut at five micrometres; total size of 2.49 millimetres) four drill holes (red) are introduced perpendicularly to the block surface in a rectangular arrangement encaging the specimen. *Image capturing:* From every serial section (s_1 through s_n) two computer images — one phase-contrast image and one bright field image — are captured *via* a video camera attached to the microscope. *External marker-based automatic congruencing:* All images are automatically realigned and rectified (congruenced) by a computer routine, which is a user-programmed sequence of image processing steps that uses the cross-sectional profiles of the drill holes (shown in red) as orientation markers. The routine computes the actual positions of the drill holes in a section, compares them with a reference image taken prior to sectioning, and restores both images (the phase-contrast image and the bright field image) to their original orientation. *Dynamic threshold segmentation:* The objects to be re-

constructed are segmented from the respective images based on their distinct grey-values. This segmentation procedure results in separate stacks of corresponding images that contain either a morphological structure (e.g., the brain vesicle, yellow, the major head vein, blue) or the gene expression signal. *Surface generation:* Linking of all sectional profiles from a stack forms a continuous three-dimensional surface of the selected object. *Three-dimensional visualisation:* The separate models of gene expression patterns and of morphological structures are reassembled to form a three-dimensional compound model, and selective visualisations of its aspects are generated. Medial view of the left half of the embryo surface and the *Scgn10* gene expression signal (blue) of the left body side (*top*). Reconstruction of the head (*left*). Magnified view of the *Scgn10* gene expression signal (blue) inside the precursor of the trigeminal ganglion (yellow) (*middle*). Dorsal view of the segmental *Scgn10* gene expression pattern in the back (light blue), within the spinal ganglia, alongside the neural tube (yellow) (*right*). For a detailed description of the models see legend of figure 7.14). With permission from *Trends in Biotechnology* (Streicher and Müller, 2001).

prehensive expression maps that can be published in visual, image-based internet databases.

The ideal methodological approach would be high resolution monitoring *in vivo* using multiple stains. Although several attempts have been made to use magnetic resonance imaging (MRI) (Louie *et al.*, 2000) and confocal laser scanning microscopy (CLSM) (Hecksher *et al.*, 2001) in gene expression studies, there is still no routine technology that combines adequate resolution with multiple staining and *in vivo* observation (Streicher & Müller, 2001). This chapter introduces a highly automated, computer-based method to visualise the expression of genes in three-dimensions using serial histological sections taken from whole-mount *in situ* hybridised specimens (figure 7.13).

7.7.2

Principles and basics of the GeneEMAC concept

The method is based on a previously established concept for the three-dimensional visualisation of morphological structure called external marker-based automatic congruencing (EMAC) (Streicher *et al.*, 1997) and is therefore called ‘GeneEMAC’ (Streicher *et al.*, 2000). GeneEMAC combines the high resolution of histological sections with the methodological advantages of whole-mount staining. With the ex-

ception of a precision drill, the method does not require specialised laboratory equipment and covers a broad range of specimen sizes and resolutions. The hardware required for image acquisition and processing is a research-grade microscope equipped with a video camera and a standard computer system capable of capturing images from the video camera. Most of the image processing procedures are either supported or automatically performed by special routines for image analysis and processing (macros) written in a PASCAL-like programming language. These macros run under NIH-Image version 1.62, which is a public domain image analysis software developed at the National Institutes of Health (NIH, Bethesda, Maryland, USA). For final three-dimensional modelling and visualisation, any reconstruction software that is able to convert sets of binary tag image file format (TIFF) images into three-dimensional objects can be used (*e.g.*, Velocity² Pro™ from Image3, LLC, Salt Lake City, Utah, USA).

7.7.3

Specimen preparation

7.7.3.1 Whole-mount *in situ* hybridisation (WISH)

Because whole-mount *in situ* hybridisation (WISH) currently represents the standard technique for the overall visualisation of gene expression patterns in the embryo, the GeneEMAC protocol uses ordinary WISH preparations as the starting material. Overnight incubation in buffered formaldehyde is recommended in order to stabilise the colour deposit. Alternatively, it is possible to start from unstained material and to perform on-section *in situ* hybridisation. However, this approach is more time-consuming and more critical in terms of artificial staining intensity variation among individual sections.

7.7.3.2 Embedding

The WISH embryos are embedded in glycolmethacrylate resin (JB4; Polysciences Europe GmbH, Eppelheim, Germany) following the standard protocol. If on-section staining with molecular probes for *in situ* hybridisation or immunostaining is desired, a larger mesh size should be provided for better penetration of the probes using specialised embedding media.

7.7.3.3 Introduction of external markers

Since mounted histological sections vary in orientation, size, and distortion, a set of reliable landmarks (external markers) that serve as an objective reference for computer-driven realignment, scaling and rectification (congruencing) are required (Streicher *et al.*, 1997). Therefore, prior to sectioning, four parallel drill holes are introduced into the resin block. These are placed in a rectangular arrangement engaging the specimen. One of the drill holes must be placed asymmetrically in order to enable automatic detection of the section orientation by image analysis software (figure 7.13). Later, when sections are viewed with the microscope and digitised,

the drill holes must lie within the view field. Thus, the width and height of the rectangle must be chosen in accordance with the later microscopic magnification desired. In order to achieve maximum axial parallelism and correct rectangular positioning of the drill holes, the block holder is fitted onto a microscope cross stage that permits precise horizontal movements of the resin block. This microscope stage is mounted on a high precision drill stand holding a drilling machine. The diameter of the drills may range from 100 to 300 micrometres depending on the specimen size.

7.7.3.4 Capturing of a reference image

After drilling, a reference image of the block surface is taken using the same magnification as will be used to view the future sections. This serves as the template to bring the sections into congruence at a later stage.

7.7.3.5 Histological sectioning

The resin block is fixed onto the stage of a rotary microtome. Using a dissection microscope with a grid-plate mounted in its eyepiece, the block is positioned in such a way that the axes of the drill holes, which are visible as opaque cylinders within the transparent resin, are oriented perpendicularly to the sectioning plane. Subsequently, serial sections are cut with a steel knife. The sections are transferred to a warm water bath and allowed to stretch. Between twelve and twenty sections are collected and mounted on silane-coated slides. Finally, the sections are cover-slipped without any additional staining.

Use of resins permits sections of thin as one micrometre to be obtained. However, when deciding on the useful thickness, the two-dimensional resolution of the later digital images must be taken into account. The sectional thickness (resolution in the third dimension) should equal the x and y -pixel dimensions, so that the resolution is as homogenous as possible in all of the three dimensions that will make up the resulting voxels. For small vertebrate embryos (*e.g.*, a Theiler stage 17 mouse embryo), five to seven micrometres result in a good z -resolution.

7.7.4

Microscopy and digital image processing

7.7.4.1 Image capturing

The processed slides are viewed under a microscope equipped with a computer-linked video camera. The individual sections are roughly pre-oriented, placing the four holes near the corners of the microscopic view field. Two grey-scale images of each section, one phase-contrast and one bright-field image, are then captured (figure 7.13). Eight frames per image are automatically averaged, corrected for uneven illumination, and numbered using a capturing image analysis and processing routine. The images are stored in TIFF format. The phase-contrast images enable the automatic extraction of the external markers (drill holes) for the realignment procedure, and also allow for the later segmentation of the boundaries of morpho-

logical structures. Since no counter-stains are used, the bright field images are directly used for segmentation of the gene expression signals.

7.7.4.2 Image congruencing

The congruencing procedure consists of macro-driven image processing operations that realign and rectify all sectional images. First, the drill holes in each image of the phase-contrast stack are segmented. If the macro detects fewer than four drill holes in an image, the image is transferred to a separate folder, and the ill defined (e.g., debris filled) drill hole is segmented manually. In the next step, a congruencing routine computes the locations of the centres of gravity of the holes. These coordinates are compared with the original distances and locations in the reference image, and each image is automatically shifted, rotated, and scaled to its original size and position. The parameters of these operations are recorded and subsequently applied to the corresponding bright field image of the same section. Due to the asymmetric pattern of the drill holes (figure 7.13), the macro is also able to detect and correct inversely mounted sections. Both macros are programmed as loops that automatically process all images of a series, one after the other, without any need for supervision by an operator.

7.7.4.3 Image segmentation

In this step, the selected structures of interest are converted to isolated monochromatic objects, while all other structures are removed (figure 7.13). Segmentation is also performed by another computer routine in a largely automatic manner. Individual thresholds, which define structures of interest by their grey values, are set once in a single sample image and are then automatically adjusted to all other images using a dynamic thresholding algorithm (Streicher *et al.*, 1997). Using this procedure, gene expression signals and selected morphological features can be extracted from the bright field and phase-contrast images. Additional morphological features that cannot be defined by automated contour detection and filtering operations are segmented manually. This segmentation procedure results in multiple series of black and white TIFF images, one containing the gene expression patterns, and several others that each contain one of the relevant morphological structures.

7.7.4.4 Generation of a three-dimensional model

Three-dimensional models of the individually selected objects are reconstructed by generating a connecting surface that recombines the object's sectional outlines in the file set. By entering a value corresponding to the section thickness, the images of the stack are placed in the appropriate distances. The final three-dimensional models of gene expression patterns and of morphological structures can be visualised either separately, or as any combination of solid or transparent objects (figure 7.13). Each model or combination of models can be subject to further volumetric, stereometric, and geometric analyses.

7.7.5

Visualisations of models

The models resulting from the GeneEMAC method can be visualised in four different ways. (i) Conventional images of single models (e.g., the gene expression pattern) or combinations of models (gene expression and related morphological structures) (figures 7.13, 7.14 und 7.15). (ii) Stereoscopic pairs of colour images that can be fused to a single three-dimensional image by cross-eyed viewing. (iii) Animations in the QuickTime™ movie format derived from a sequence of preset views. (iv) Interactive QuickTime™ VR models that can be turned by the viewer, generated from series of views from different perspectives. Examples of these visualisation types can be viewed at <http://www.univie.ac.at/GeneEMAC/>.

7.7.6

Examples**7.7.6.1 Whole-organism screens**

In screens for relevant expression patterns of new genes and for interesting phenotypes in genetically modified specimens, usually the entire organisms are investigated. Figure 7.14 provides an example of a three-dimensional whole-body visualisation of the expression pattern of the neuronal marker gene *Scgn10* (Anderson *et al.*, 1985; Stein *et al.*, 1988) in a Theiler stage 17 (Theiler, 1972) mouse embryo.

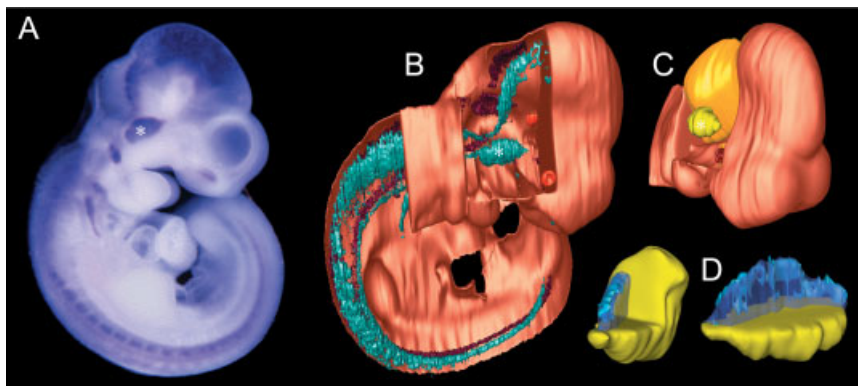


Figure 7.14 Reconstruction of a whole embryo. **A)** Theiler stage 17 WISH embryo labelled for *Scgn10*. **B)** Representation of the whole embryo with the body surface partially removed to expose the signal (right side *Scgn10* expression in blue; left side *Scgn10* expression in purple; lens pits in red). **C)** Cranio-lateral view of the head, windowed above the mesencephalic brain vesicle, showing the anlage of the trigeminal ganglion (*) protruding from the vesicle (yellow); lens pit in red. **D)** Distribution of the *Scgn10* signal (blue) inside the anlage of the trigeminal ganglion (yellow). The anterior dorsal quadrant (left) and the dorsal half of the morphological (right) ganglion are removed. *Scgn10* expression is shown in transparent blue, whereas maximum signal intensity areas are displayed as solid blue objects. With permission from *Nature Genetics* (Streicher *et al.*, 2000).

The specimen was reconstructed from 478 serial sections. Despite the lower microscopic resolution necessary for capturing complete sections through a whole embryo, it is possible to accurately visualise the topology of signal-expressing regions. Regions of particular interest can be enlarged and different signal intensities can be visualised. This immediate digital enlargement capability represents a major advancement over conventional screening techniques. In the present case, the enlargement of the anlage of the trigeminal ganglion reveals distinct sub-regions of maximum signal intensity that are indiscernible in a conventional study. In the dorsal portion of the ganglion three foci of maximum intensity can be distinguished, which coincide with the three parts of the future trigeminal nerve, indicating a very early molecular partitioning of the ganglion anlage.

7.7.6.2 High-resolution analysis

The study of topological details of gene expression at the cellular level require high microscopic resolution. If the specimen size and the drill holes are kept as small as possible, GeneEMAC allows to resolve regional differences in gene expression even within small cell populations. In the example shown in figure 7.15, the three-dimensional visualisation of the *Meox1* protein (formerly *Mox-1*; Candia *et al.*, 1992) in a developing somite reveals new aspects of its expression pattern. In particular, somite derivatives that appear to express no or low levels of *Meox1* in the whole mount (figure 7.15.A) can be clearly distinguished: the myotome, the central portion of the rostral sclerotome, and the outer layer of the dermomyotome.

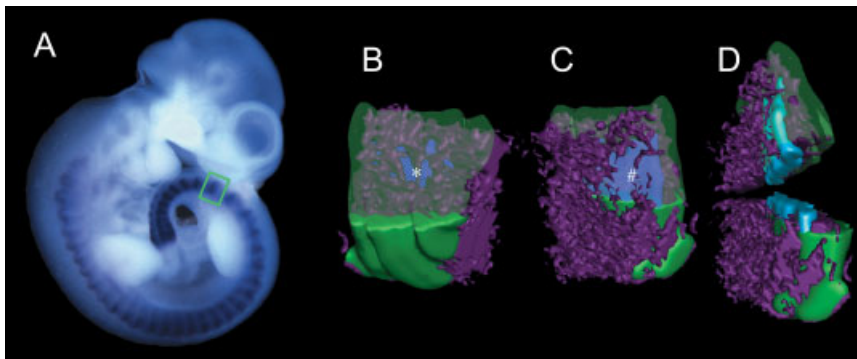


Figure 7.15 High-resolution reconstruction of a single somite. **A)** Theiler stage 18 WISH embryo hybridised for *Meox1*. The rectangle indicates the location of the reconstructed somite (sacral somite 1, according to Sporle & Schughart, 1997). Lateral view **B)**, medial view **C)**, and cranial view **D)** of the reconstructed somite separated into dorsal and ventral halves. Morphologically differentiated somites (dermomyotome and myotome) are coloured in green. The dorsal half is rendered transparent for a

view of the signal inside the somite. *Meox1* signals are coloured in purple. Signal-free compartments corresponding to the myotome are blue. Note that the signal transgresses the boundaries of the somite both medially and caudally, but is absent in the centre of the dermomyotome (*) and has a distinct signal-free area on the medial side of the anterior portion (#). With permission from *Nature Genetics* (Streicher *et al.*, 2000).

7.7.7

Discussion

In terms of specimen size and structural levels, the GeneEMAC method can be used for the entire range of light microscopy resolutions. Gene expression patterns in entire organisms, individual organs, or selected cell populations can be reliably visualised as three-dimensional models and can be combined with morphological features. The digital data are appropriate for systematic compilation and archiving in universally accessible internet databases (see the section on a database of three-dimensional visualisation of gene expression in this chapter).

The methodological concept of external marker-based automatic congruencing is also fully compatible with fluorescent microscopy. In this way, the expression patterns of multiple genes, visualised by different fluorochromes in separate channels, can be analysed simultaneously, assessing the study of co-expression and genetic networks. In addition, the three-dimensional models can be used for the quantification of gene expression parameters, and provide a basis for inter-specific comparisons and for theoretical mathematical models.

A clear limitation of the GeneEMAC concept is the inability to perform *in vivo* observations due to the necessity to physically section the specimens. If *in vivo* observation is indispensable, virtual sectioning technologies such as confocal laser scanning microscopy and magnetic resonance imaging (MRI) microscopy are the methods of choice. However, although MRI has almost reached microscopic resolution (Roush *et al.*, 1997; Smith *et al.*, 1999; Smith, 1999), and despite recent advances involving the detection of reporter gene expression in living animals (Louie *et al.*, 2000), the potential to visualise molecular targets is still very limited. CLSM techniques, on the other hand, permit precise gene expression studies even at single-cell and subcellular levels but are limited to small specimens or early developmental stages. CLSM reconstructions also lack the ability to visualise the accompanying histomorphology at a quality equivalent to light microscopy.

7.8

RNA-based gene expression databases and analyses tools

7.8.1

Introduction

This section describes the principles of constructing a database for mRNA expression with particular reference to expressed sequence tags (ESTs). The most important step in this process is to consider carefully in advance what analyses should be performed and what data should be gathered.

7.8.1.1 Choice of a database management systems

Several types of database management systems exist, including hierarchical, network and relational designs. Of these, relational database management systems have become the most commonly used. Relational databases organise data using a table structure called the 'relation'. Object databases can be regarded as one type of expanded relational database. Relational database management systems include expensive ones such as Oracle, although PostgreSQL (Momjian, 2000; <http://www.postgresql.org/>) and MySQL (<http://www.mysql.com/>), which can be obtained for free *via* the internet and which are sufficient in many cases for laboratory use. If UNIX is not available commercial databases such as Microsoft® Access may be used if no more than a few tens of thousands of sequences are to be registered (<http://www.microsoft.com>). However, even in this case, additional programming using Visual Basic will be required for full-scale analyses. Considering that many types of sequence analysis software such as Phred/Phrap have been developed using an operating system related to UNIX, one good option is to use Linux as the operating system together with PostgreSQL. AceDB database (a *Caenorhabditis elegans* database), which was developed by the Nematode Genome Project (<http://www.acedb.org/>), is such an example.

Because almost all relational database management systems are based upon SQL standard (Date & Darwen, 1996), there are a few differences regarding convenience of use as long as one is constructing a biological database dealing with up to several tens of thousands of sequences.

7.8.1.2 cDNA database based on relational database management systems

When constructing a relational database management system, the database formats for cDNAs are determined by the users themselves (figure 7.16). An identification number, a data source, a sequence length and a sequence of each clone, for example, may be registered in a single table. When cluster identification numbers are used to designate clustered clones sharing the same 3'-tag sequences, a cluster identification number table is often made, making it possible to refer to each cluster rather than to its clones. When each clone has two sequences, such as a 5'-tag and a 3'-tag, another table can be made. In this way, the table layers can be deepened as required (figure 7.17).

7.8.1.3 Spatial and temporal gene expression data

Although it depends on the biological system being studied, large-scale cDNA sequencing projects often simultaneously collect spatial and temporal gene expression data. When a database is constructed containing non-text data such as photographs of *in situ* hybridisations, it is necessary to store descriptions of the results observed by researchers. To do this, it is convenient to prepare checkboxes, using which Boolean screening, including 'AND/OR', can be done in later statistical analyses. However, since this Boolean type of description causes the loss of much observational data, free-formatted memoranda may be needed to describe the observations in full detail.

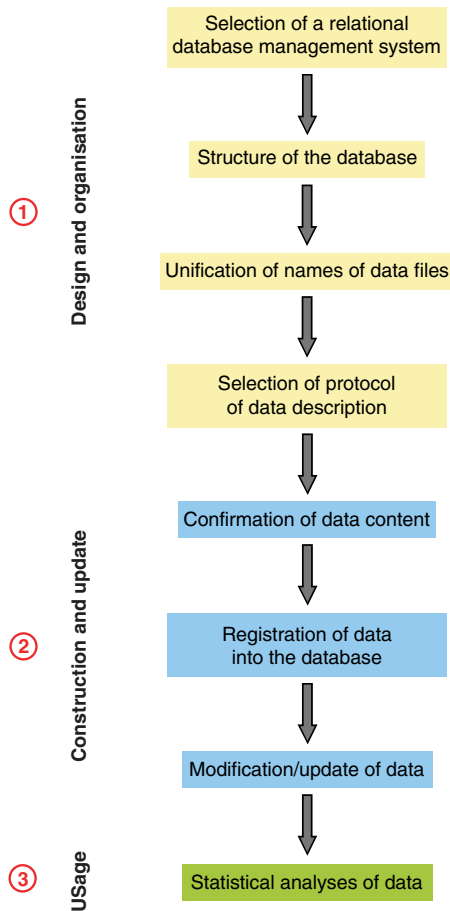


Figure 7.16 Flowchart for the construction of a relational database management system. See text for details.

7.8.1.4 Data format

The data formats of collaborators must be unified in order to store them in the database. For example, there should be a common protocol in order to process the text-formatted raw sequence data from auto-sequencers into a standard format. In general, the FASTA format (Pearson, 1990, 1994 and 2000) is most convenient and very easy to convert into another frequently used format called BLAST (basic local alignment search tool) (Altschul *et al.*, 1990, 1994 and 1996).

Unification of data formats is relatively easy if you choose software such as Phred/Phrep (see the Phred/Phrap/Consed System homepage at <http://bozeman.mbt.washington.edu/> for information). One must, however, pay attention to whether the sequences contain only the usual five characters (A for adenine, T for thymine, G for guanine, C for cytosine and N for any of the four nucleic acids adenine, cytosine, guanine or thymine), U (uracil) in addition to T (thymine), or universal codes (B for cytosine, guanine or thymine; D for adenine, guanine or thymine;

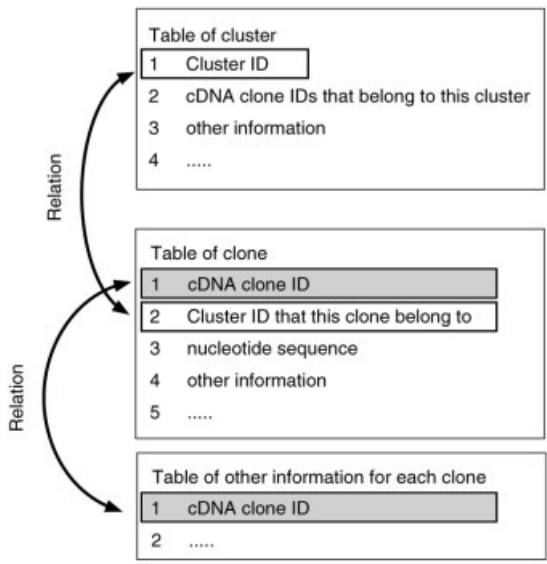


Figure 7.17 Relationship of tables in a relational database management system. Tables are made to be in relation to each other using the common items shared by the tables in the relational database management system. Note that some items may remain open when dealing with experimental data.

H for adenine, cytosine or thymine; I for deoxyinosine; K for guanine or thymine; M for adenine or cytosine; R for the purine bases adenine or guanine; S for cytosine or guanine; V for adenine, cytosine or guanine; W for adenine or thymine; Y for the pyrimidine bases cytosine or thymine), and whether vector sequences have been removed or masked by the letter X, and so on. Whether reverse sequences are converted to the complementary sequences is an elementary but important matter. This issue must be clarified from the start; it cannot be addressed later.

7.8.1.5 Types of data

After initial data are registered, there is often a need to register related data. Examples of indispensable items seen in typical EST databases are shown in table 7.5. Considering this point in advance makes it easier to design a database. For example, data that might be registered later include the expression data obtained from whole-mount *in situ* hybridisation studies and the predicted length of the mRNA. Furthermore, value-added data such as the results of sequence homology searches against sequences of other organisms or results of motif searches are often registered together in the database.

There are some cases where further ideas are required. Items that consume a lot of memory such as pictures are better registered separately as individual files linked to the database. Depending on the relational database management system that is used, one may have to define the data type of each data item before registration: a character format for data files and sequences, a number (or flow decimal number) format for sequence length, and a date format for data registration date, for instance. When accompanied by picture data, in particular, analysed data such as ‘expressed in muscle’ or ‘signal disappeared at stage 15’ are often registered to-

Table 7.5 Examples of indispensable items seen in typical expressed sequence tag databases.

<i>Field Name</i>	<i>Examples</i>	<i>Data type</i>	<i>Importance</i>
Clone ID	Clone ID	Char	Essential
Length	Predicted insert length	Float8	Optional
Sequence	<i>e.g.</i> , AGTCAGTC....	Varchar	Essential
Cluster ID	ID for redundant cDNA clones	Char	Optional
Date	Date of last data modification	Date	Essential
Similar genes	Result of similarity searches	Varchar	Optional
Others	...	...	...

gether. For the purpose of subsequent statistical analysis, one should unify the protocol of description. One of the best ways to do this, is to make the registration view in the form of a selection list format, which also helps to avoid typing errors. Note that remarks or memoranda with a free format should be minimised in this case, as they are of almost no use for statistical analyses.

It is impossible to completely prevent invalid cDNA sequences from contaminating a database when collecting a large amount of sequence information; for example, contamination by insert recombination or bacterial sequences may be impossible to eliminate completely, but it may be a good idea to make several checkboxes which allow removal of some kinds of data after registration.

7.8.1.6 Use of other databases

Many databases focusing on gene expression are available on the internet. Using these before will provide many helpful tips towards developing ones own database. The annual January issues of *Nucleic Acids Research*, which is a special annual issue focusing on databases, is a good source of references (*Nucleic Acids Research Database* issues since 1996). However, the best way to learn about all databases described in this book is to visit them. The brief descriptions are too short for the authors to offer any advice to those who may wish to publish their own small, specialised expression databases on the web. Some of the issues involved have been discussed at length elsewhere, but the authors are happy to offer assistance to anyone planning similar databases of their own.

7.8.2

ASDB – The Alternative Splicing Database

The organisation of eukaryotic genes into modular exon units is thought to have had major importance in evolution since it allowed shuffling of these modular units in allowing generation of novel protein structures and functions. Gelfand and co-workers demonstrated that one in every three human genes might give rise to alternatively spliced transcripts (Mironov *et al.*, 1999). Similar estimates of 30 to 38 percent have been reported by others (Brett *et al.*, 2000; Hanke *et al.*, 1999; Kan *et al.*, 2001).

In 1998, a public Alternative Splicing Database (ASDB) was created as a central repository of data on alternatively spliced genes (Dralyuk *et al.*, 2000; Gelfand *et al.*, 1999). This is an ambitious undertaking because a uniform reporting method for such data does not yet exist. ASDB consists of two divisions: ASDB(proteins), which contains information on the protein products of alternatively spliced genes, and ASDB(nucleotides), which contains complete sequences of alternatively spliced genes. Although in many respects the SWISS-PROT database is well curated, the information content is incomplete with reference to alternative splicing and does not allow for automatic retrieval of such entries. ASDB(proteins) was built to improve access to this information.

First, protein sequences were selected from SWISS-PROT by means of a full text search for the words 'alternative splicing' and 'varsplc'. To find proteins which represent alternatively spliced variants of a single gene, an automated clustering procedure was developed using similarity analysis. Protein sequences retrieved from SWISS-PROT were compared all-against-all in order to find common sequence fragments using a string comparison tool. Each cluster is represented in the database by the multiple global alignment of its members. Different colours in these sequence alignments show common fragments and variable areas, allowing for easy identification of regions produced by alternative splicing.

In the current version of ASDB(proteins), entries are marked with different symbols to allow for easy differentiation among the following three types. (i) Proteins which are part of the ASDB clusters and the corresponding multi-alignments, (ii) proteins whose SWISS-PROT entries contain information on different variants, and (iii) proteins for which no information on variants is available in any database at the present time. The ASDB(nucleotides) division was generated by collecting all GenBank entries containing the words 'alternative splicing' and further selection of those entries that contain complete cDNA sequences. This is a new division that will be expanded.

ASDB can be accessed *via* the internet at <http://cbcg.lbl.gov/asdb>. The database can be searched using Medline, SWISS-PROT, and GenBank identifiers and accession numbers. ASDB contains internal links between entries and/or clusters, as well as external links to Medline, GenBank and SWISS-PROT.

7.8.3

AsMamDB – The Alternative Splice Database of Mammals

AsMamDB is an alternative splicing database of mammals (Ji *et al.*, 2001) that can be accessed at <http://166.111.30.65/ASMAMDB.html>. The current version of AsMamDB (actual state: February, 2002) contains 1,563 alternatively spliced genes of human, mouse and rat. Of these, 899 genes are from human, 431 are from mouse and 233 from rat. Each gene is associated with an entry consisting of five files – a flat file, a FASTA file, an alignment file, a gene structure file and a functional segment file.

Flat files provide descriptions of genes, their locations in chromosomes, the products of these genes, and the tissues in which they are expressed. In addition, links to related entries of Medline, UniGene and GenBank are provided, and various transcripts of genes are annotated. In FASTA files, DNA and mRNA sequences are provided in FASTA format for users' convenience. In alignment files, alternative splicing patterns are represented by multiple alignments of the various transcripts of the gene. In the gene structure files, gene structures are illustrated by distributions of exons, introns and various regulatory elements. Splicing structures may not be completely precise, since they are generated automatically by means of a heuristic alignment tool. The tool cannot guarantee all the splicing structures to be correct, though it performs well in comparison with other such tools. In functional segment files, nucleotide sequences of exons, introns, promoters, enhancers and other regulatory elements are provided in files of FASTA format.

In order to query the database, one can use both keywords and nucleotide sequences. If keywords, such as accession number and gene name are submitted, entries containing these keywords will be returned. If one submits a nucleotide sequence, the submitted sequence will be compared with the representative sequences in the database using BLAST (Altschul *et al.*, 1997). Entries that contain the most similar sequences will be returned. Human genes can also be accessed through chromosome figures. By clicking a specific region of a chromosome, alternatively spliced genes in that region will be listed.

As part of the database, a web interface of Asalign, a multiple alignment tool for revealing alternative splicing patterns, is provided. When users submit a cluster of sequences in FASTA format, the tool will give an overview of their alignment.

7.8.4

BodyMap – An anatomical gene expression database of human and mouse

The BodyMap database is based on locally produced unique gene expression data of human and mouse and is accessible at <http://bodymap.ims.u-tokyo.ac.jp> (Kawamoto *et al.*, 1999 and 2000; Okubo *et al.*, 1992; Sese *et al.*, 2001). The data include more than 300,000 site-directed EST sequences collected from non-biased cDNA libraries. As of December 2001, 60 human and 39 mouse tissues were included. Libraries were all constructed by 3'-directed *Mbo* I cloning (Okubo *et al.*, 1992).

in which transcripts from one gene are represented by exactly the same fragments referred to as gene signature. As a result of gene signature counting by random sequencing and referencing to the database RefSeq and UniGene, a spreadsheet showing the occurrence of each transcript in each library (geneID $\times$ tissue name), the expression matrix, is generated periodically. BodyMap provides three categories of queries. Selection of tissue name returns a matrix permuted in the descending order of selected tissue column. Rows of matrix can be selected by gene identification numbers and by keywords. Selection by expression patterns is facilitated with gene ranking system (Sese *et al.*; 2000) that order the genes according to the statistically calculated similarity to the query pattern. In addition, BodyMap prepares detailed information for each gene signature as a card. To provide more accurate anatomical distribution pattern for each gene signature, BodyMap is now incorporating data produced by the PCR-based expression profiling method introduced amplified fragment length polymorphism (iAFLP; see chapter 4).

7.8.5

The CYTOMER[®] Gene Expression Database on human organs and cell types

An important aspect of each gene is its expression pattern. Up to now, most tools developed for genome annotation have emphasised the identification of open reading frames in order to deduce their potential products and elucidate their biochemical function using sophisticated sequence or structure homology searches. However, assessing the biological and physiological function of a gene is impossible without knowing when, where and under what conditions the gene is expressed.

To overcome this lack of knowledge, advanced technology is increasingly being applied to the collection of large amounts of data on gene expression, for example, by means of DNA microarray technology. Many researchers in the pharmaceutical industry hope to identify new target genes for drug development using such approaches. However, use of these data is hampered by the lack of a standard for their representation. This makes it difficult to compare the data generated at one site, let alone that generated by different laboratories.

These problems have been tackled by international consortia such as that coordinated by the European Bioinformatics Institute (EBI) in Hinxton, United Kingdom, called ArrayExpress (<http://www.ebi.ac.uk/microarray/ArrayExpress/>). These initiatives have already been successful in setting up standards for the expression patterns but have highlighted the urgent need for ontologies of, for example, expression sources (Brazma *et al.*, 2000).

Several international working groups have been established, one of which is dedicated to the development of ontologies for sample description, including information on the tissues and cell lines involved. Pioneering work has already been done for the mouse system at the Jackson Laboratory (Bar Harbor, Maine, USA) in cooperation with Edinburgh University (Edinburgh, United Kingdom), the Edinburgh Mouse Atlas (<http://genex.hgu.mrc.ac.uk/Databases/Anatomy/>) (Davidson *et al.*, 2001a).

In the following, an ontology of the anatomical, morphological structures of the human organism, and the cell types constituting these structures is presented. The underlying trees have been generated for all human embryonic stages and for the adult organism.

The exact definition of the term 'ontology' is still going on. Nevertheless, we decided to call CYTOMER[®] an ontology because it is a concise description of principal, relevant entities and their potential relations to each other (Schulze-Kremer, 2002). In empirical items, it is a prototype-based ontology since it distinguishes subtypes by a comparison with a typical member or prototype for each subtype (Sowa, 2000).

CYTOMER[®] has been constructed and is maintained as a relational database system which aims to provide a comprehensive overview on all gene expression sources in the human (Chen *et al.*, 1999). Gene expression sources are organs, tissues and cell types at various developmental stages. Therefore, CYTOMER[®] is a database of physiological systems (table 'system'), developmental stages (tables 'stage' and 'period'), anatomical structures and substructures (table 'organ') and the constituting cell types (table 'cell') in different organisms or species (table 'species'). The central table of CYTOMER[®] is 'HUB' which is a list that links entries of the five other tables. The HUB table represents anatomical and histological knowledge on the identity of the cells that occur, their function, the organs in which they occur, the developmental stages at which they are present and the species in which they arise.

The most extensive tables of CYTOMER[®] are organ and cell. The organ table represents an ontology of anatomical structures and morphological substructures. It is hierarchically organised as a tree-like directed graph, starting with the entry 'human body of developmental stage *n*' as the 'root' (level 0). The tree for the adult organism, which is the most complex one, proceeds through 80 nodes of level 1, the 'primary organs', exhibits 2,079 inner nodes and terminates with 6,321 end nodes ('leaves'). The hierarchy depth varies greatly between the branches, ranging between one to eleven levels.

The leaves of the organ tree, the finest definable morphological structures, mostly consist of a single cell type, therefore exhibiting one criterion of a 'tissue'. Cell types as they are listed in the cell table are connected exclusively to these end nodes of the organ tree. In cases where a cell type is known to occur within an organ, but the structure of its occurrence is undefined, an additional dummy leaf is attached to the node connected to that cell type.

The cell table is currently under construction. When complete, it will include the international, English and German cell name as well as synonyms. There will also be short descriptions of cell location and function. Moreover, since a hierarchical order exists between the cells, reference to the respective cell parent of each entry will also be given.

In the relational system, organ end nodes and cells are linked through the HUB table. Any of these combinations may be assigned to a physiological function and, thus, to one of the 13 items of the system table. This was done since many organs can be assigned to more than one physiological system, for instance the brain can

be assigned both to the nervous system and the endocrine system. However, when the functions are traced to the detailed morphological structures, our experience has been that cells of a defined type within one of the end structures can always be assigned to a defined physiological system.

The CYTOMER[®] database is applied to map expression patterns of genes and gene products. Up to now, its main use is to represent expression patterns of transcription factors as they are given in the TRANSFAC database (Wingender *et al.*, 2001). However, mechanisms are in preparation to allow any kind of expression pattern to be linked with CYTOMER[®]. Therefore, CYTOMER[®] will be an essential part of BIOBASE's ExProfile system (BIOBASE GmbH, Wolfenbüttel, Germany) which is designed to store, manage and interpret gene expression patterns and profiles.

At the time of writing, the organ-table hierarchy of the human adult is complete. The organ ontology is freely available on the internet under <http://www.biobase.de:8080/index.html>.

7.8.6

Database of three-dimensional visualisation of gene expression

The best way to organise a systematic, generally accessible compilation of three-dimensional gene expression data is as a searchable database on an internet platform. In conjunction with the GeneEMAC method, the structure for a three-dimensional visual gene expression database shown in figure 7.18 was proposed.

Three-dimensional visualisations of individual models can be stored in a database arranged in the following hierarchy: species > gene > developmental stage > structural level (whole embryo, body part, developmental subunit). Every model is further classified in terms of the organs, tissues, and cells represented. Within the directory of each model, various kinds of visualisations are organised by media type: images (in joint photographic expert group, JPEG, graphic format), stereoscopic images (in JPEG graphic format), movies (as QuickTime[™] movies), and interactive models (as QuickTime Virtual Reality[™] movies). Source data, such as views of the original whole mount specimens, all images of the serial sections used, and high-resolution images of selected sample sections or of additional stains can be archived and linked to the three-dimensional model.

The user interface proposed for querying the database ('chooser') offers four categories of search terms, which can be combined by Boolean operators. They facilitate both, structure and gene-oriented queries. The response takes the form of a choice of available structural levels and a subsidiary overview of available representations ('3D-selector'). Together with each model, direct links are provided to the source images as well as to corresponding entries in external text-oriented gene databases.

Several examples of three-dimensional visualisations of gene expression patterns organised according to this preliminary structure can be accessed at <http://www.univie.ac.at/GeneEMAC/>.

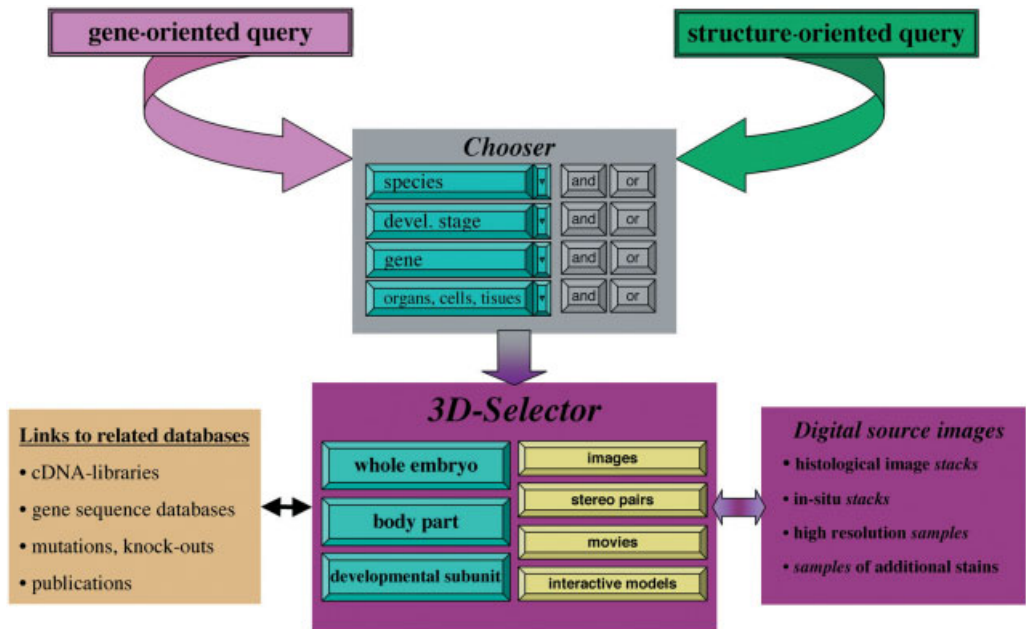


Figure 7.18 Schematic organisation chart for a visual three-dimensional gene expression database. The figure is reproduced from Streicher *et al.* (2000) with permission from *Nature Genetics*.

7.8.7

dbEST – The Database of Expressed Sequence Tags

dbEST is a division of GenBank that contains a vast collection of expressed sequence tags (ESTs). Most expressed sequence tags are short (300 to 500 base pairs) reads from cDNA molecules derived from mRNA expressed in a variety of cell types and organisms. Historically, dbEST is enriched in 3'-end reads, many of which do not contain much coding sequence, since they are derived from 3'-untranslated regions. The latest release of dbEST (actual state: November 2001) represents 364 species of animals, plants, and bacteria, and contains over 9.5 million entries (see table 7.3). The two organisms with the highest expressed sequence tag counts are human and mouse, with 3,900,480 and 2,363,201 dbEST entries respectively.

In addition to the expressed sequence tag sequence, dbEST contains the following information: publication details (title, authors, year, and status of the accompanying paper); expressed sequence tag library description (library name and organism, as well as a number of optional fields such as organ, tissue, cell type, cell line, or additional free-text description); submitter's contact information; annotation of library tags found in a given expressed sequence tag sequence, indication of the presence of poly(A) tail, start and end of high-quality sequence, *etc.* Genomic mapping information is also available for a subset of expressed sequence tags.

As with all divisions of GenBank, access to dbEST is a free to all, and it is one of the default databases available for a subset of specialised basic local alignment search tool (BLAST) searches at the National Center for Biotechnology Information (NCBI, Bethesda, Maryland, USA). It is especially powerful when used in a translated BLAST search (tBLASTn or tBLASTx), where a protein sequence or a translated nucleotide sequence is used as a query against a translated nucleotide database. The corresponding BLAST pages (see the listing at <http://www.ncbi.nlm.nih.gov/BLAST/>) include all ESTs, human only ESTs, mouse only ESTs, and other ESTs among the database choices.

The dbEST homepage allows Entrez-style searches to be performed. It is, also possible to limit a search to dbEST while narrowing down regular NCBI nucleotide Entrez queries.

The 'UniLib' button provides a link to the website of the Unified Library Database (<http://www.ncbi.nlm.nih.gov/UniLib/>), which provides a library-wide view of expressed sequences (expressed sequence tags and serial analysis of gene expression (SAGE) tags). The Unified Library Database can be searched by organ or tissue type, library name, organism, library type (expressed sequence tags *versus* SAGE tags), and the minimal number of sequences in a library. The results provide detailed information about library construction, sequence count, and links to related libraries.

At the bottom of the dbEST webpage, there are links to the Integrated Molecular Analysis of Genomes and their Expression consortium (IMAGE; <http://image.llnl.gov/>) and related websites. The IMAGE consortium, initiated in 1993, is a popular resource of high-quality cDNA clones for microarray studies. The ultimate goal of this consortium is to construct expression arrays where every cDNA from a genome of interest is present, for a number of commonly studied species.

The dbEST website also provides links to Human Genome Resources, UniGene, LocusLink of the National Center for Biotechnology Information and to the Cancer Genome Anatomy Project (CGAP) web pages of the National Cancer Institute (NCI, Bethesda, Maryland, USA).

7.8.8

DDD – Digital Differential Display

Digital Differential Display (DDD) is a tool from the National Center for Biotechnology Information (NCBI, Bethesda, Maryland, USA) which enables the comparison of gene expression levels among the various expressed sequence tag (EST) libraries of dbEST (Boguski *et al.*, 1993). These comparisons allow the identification of genes which differ among libraries of different tissues. This makes it possible to determine which genes may be contributing to a cell's unique characteristics, for example, those which make a muscle cell different from a skin or liver cell. Along similar lines, DDD can be used to try to identify genes for which the expression levels differ between normal, premalignant, and cancerous tissues, or between different stages of development.

The ESTs in dbEST are assigned to genes by UniGene (Wheeler *et al.*, 2001). UniGene groups the sequences of dbEST into organism specific clusters, each cluster

representing a unique gene. Currently, UniGene is available for six vertebrates (human, rat, mouse, cow, zebrafish and frog), two insects (fly and mosquito) and five plants (thale cress, wheat, barley, rice and corn). UniGene is updated monthly, incorporating the new dbEST entries into its clusters. The algorithm for generating the UniGene clusters has several stages, and not all of the ESTs in dbEST are always included in the final set of clusters. For each library which has clustered sequences, UniGene provides a library summary page, which gives the biological source, library construction details, and a list of the various UniGene clusters to which the ESTs have been assigned.

As in UniGene, the DDD resource is organism specific and is available from the UniGene website for that organism (<http://www.ncbi.nlm.nih.gov/UniGene/>). For those libraries which have sequences in UniGene, DDD lists the title and tissue source. A link to the UniGene summary page provides additional information about the library. From the libraries listed, the user can select two for comparison. DDD then displays those genes for which the frequency of the transcript is significantly different between the two libraries. The output includes, for each gene, the frequency of its transcript in each library, the title of the gene's corresponding UniGene cluster, and a link to a page describing the cluster. Results are sorted by significance, with the genes having the largest differences in frequencies displayed at the top. Libraries can be added sequentially to the analysis, and DDD will perform an analysis on each possible library/gene pair combination. Similarly, groups of libraries can be pooled together and compared with other pools or single libraries. DDD uses the Fisher Exact test to determine significance of the differences of transcript frequencies.

DDD restricts analysis to libraries with over 1,000 sequences in UniGene. For human data, this is less than six percent of the total number of libraries represented in UniGene, but nevertheless includes 350 libraries of 79 tissue types. Mouse is also well represented by 249 libraries representing 88 tissue types. For wheat, the most poorly represented organism in UniGene, DDD includes only nine libraries from seven tissues.

This requirement for deeply-sequenced libraries and use of the Fisher test places limitations on what genes can be found to differ among the currently available libraries. Without large number of sequences in each pool, differences in rarely expressed genes usually do not achieve statistical significance. Furthermore, the wide variety of tissue types, cell types, histology and methods of generating the libraries can make it difficult to attribute significant differences to any one aspect of the libraries. These issues underscore the need for more libraries to be made public and the need for the comparisons to be made using proper controls. Libraries generated by the Cancer Genome Anatomy Project (CGAP) (<http://cgap.nci.nih.gov/>) will be especially valuable in this regard (Krizman *et al.*, 1999), since this project has generated many human libraries from a wide variety of tissue types.

The DDD tool is publicly available as a resource on the UniGene website, but not as a stand-alone application. The UniGene clusters are updated monthly, to make use of the new ESTs which are deposited into dbEST and to exploit updated resources such as LocusLink (Pruitt & Maglott, 2001), that are used in the UniGene

clustering algorithm. Updating UniGene can also entail the deletion of various sequences. Since dbEST is a public database, the submitting authors can choose to withdraw their entries, which sometimes results in entire libraries being removed.

7.8.9

The Genexpress IMAGE Knowledge Base of the Human Genome and Transcriptomes

The Integrated Molecular Analysis of Genomes and their Expression (IMAGE) consortium (Lennon *et al.*, 1996) has established a public resource of cDNA clones representing the majority of the transcripts of the human genome. The sequences derived from these clones have been used to map the corresponding human genes in the frame of the radiation hybrid consortium (Deloukas *et al.*, 1998). The sequence and mapping data of the human genes and transcripts have been grouped in 63,000 clusters in the Genexpress Index (Houlgatte *et al.*, 1995) and integrated with expression data obtained by differential hybridisation to macroarrays and microarrays of IMAGE cDNA clones. This made it possible to establish a dynamic resource of genes, transcripts and proteins which can be studied in the context of the physiology and pathologies of the biological systems, with the goal of developing new therapeutic strategies. The Genexpress IMAGE Knowledge Base modules of the human brain (Piétu *et al.*, 1999a), and muscle (Piétu *et al.*, 1999b) transcriptomes are available at <http://idefix.upr420.vjf.cnrs.fr/IMAGE>. These contain information on sequence, mapping and expression.

Sequence data. The sequence data within the resource consists of clusters of cDNA and mRNA sequences with a schematic representation of their relative positions and polyadenylation sites, cross-indexing with UniGene and TIGR indices, sequence alignments, deduced consensus and putative coding regions, annotation of putative alternative splicing and overlapping genes, electronic expression profiles and active links to several databases, such as dbEST and GenBank (<http://www.ncbi.nlm.nih.gov/>), the European Molecular Biology Laboratory (EMBL; <http://www.embl.org/>), SWISS-PROT (<http://www.expasy.org/sprot/>), the Genome Database (GDB; <http://www.gdb.org/>), Radiation Hybrid Database (<http://www.ebi.ac.uk/RHdb/>), and the Whitehead Institute (<http://www.wi.mit.edu/>).

Mapping data. The mapping data for each cDNA cluster consists of the data which is available from cytogenetic (GenAtlas), genetic (Généthon, Marshfield), physical (Whitehead Institute YAC resource), and genic (GeneMap'99, radiation hybrid maps from Stanford and Whitehead) map sources; a unique validated map location for each gene based on mapping data integration and co-localised orphan diseases retrieved on the basis of cytogenetic and/or linkage data (OMIM, GenAtlas).

Expression data. The expression data in the resource consists of expression profiles of 5,058 human gene transcripts in seven human tissues represented by an array of 7,451 clones from the first IMAGE consortium cDNA library from infant brain, and of more than 800 human gene transcripts represented by an array of 1,091 human skeletal muscle cDNA clone inserts. Seven gene transcripts were found to be transcribed preferentially or specifically in brain. For these transcripts

Northern blot analyses of mRNA from eight adult and four fetal tissues, and from 15 distinct regions of brain were performed. In four instances, further documentation of the sites of expression was generated by *in situ* hybridisation of ratbrain tissue sections.

The Genexpress IMAGE Knowledge Base provides users with a working and navigation space containing curated, annotated, integrated data for their genes of interest. Additional modules will be added as data becomes available from studies of neuromuscular physiopathology and development, from animal models of neurodegenerative disorders, from investigation into resistance to cancer chemotherapy, and from exploration of the role of genome instability in disease. An assistant for browsing through genomic sources on the internet is in development. In addition, data are collected which are automatically saved in structured XML documents with a view to facilitating data exchange and database interoperability.

7.8.10

The Gencarta™ Database

The Gencarta™ Database (Compugen, Jamesburg, New Jersey, USA) contains annotated sequence information representing the transcriptome and proteome of several major organisms, currently human, mouse, rat and zebrafish (actual state: September 2002). By presenting a comprehensive view of these organisms' mRNA and protein aggregates, including alternative splice forms of genes, Gencarta facilitates the analysis of expression experiments based on technologies such as DNA microarrays (Hu *et al.*, 2001), mass-spectrometry and the various tag-based methods such as SAGE™.

Gencarta is based on Compugen's LEADS™ technology, which uses multiple sequence analysis methods and annotation tools to process public, raw sequence data derived from whole genome and EST (expressed sequence tag) sequencing efforts. A distinct feature of the LEADS™ process is its ability to model various biological phenomena, such as alternative splicing (Matloubian *et al.*, 2000; Sorek & Amitay, 2001), polymorphisms, homologies and repeats, and remove numerous sequencing artefacts, including chimeric sequences, intron contaminations and sequencing errors.

Gencarta holds information on over 40,000 human genes and over 150,000 mRNA transcripts. A graphic user interface allows viewing and browsing the data, and a query mechanism enables database searches and identification of genes by accession numbers, sequence homology, domains, tissue association, SAGE tags and other criteria.

Gencarta is available on a subscription basis to academic and commercial institutes. Further information on Gencarta and access to an interface presenting a pre-selected group of genes is available through <http://www.gencarta.com/>.

7.8.11

GEO – Gene Expression Omnibus Database

Gene Expression Omnibus (GEO) is a public repository of high-throughput mRNA expression data that have been generated by a variety of methods. The database is designed to hold data sets obtained from all high-throughput expression profiling platforms now in use, including spotted microarrays, hybridisation filters, high-density oligonucleotide arrays, and serial analysis of gene expression (SAGE) experiments. The database accepts experimental data from any source and organism. Organisms currently represented include *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Zea mays*, *Escherichia coli*, and *Helicobacter pylori*.

The four basic entities of the GEO database are: submitter, platform, series, and sample (<http://www.ncbi.nlm.nih.gov/geo/info/>). The entity 'platform' contains detailed information on all physical reagents used in a given expression profiling experiment. This may include the selection of clones or set of clones used, the printing procedures for microarrays, lists of unique clone or oligonucleotide identifiers, and the corresponding GenBank accession numbers of the sequences being surveyed. The section 'sample' captures the actual gene expression data, and the details of mRNA collection and experimental conditions. The section 'series' summarizes common information for a related set of samples (such as sample order in a time course or a dose response experiment), as well as any data analyses that have been performed by the submitter (e.g., clustering of genes or samples based on the similarity of their expression profiles).

An accession number is assigned to each platform, series, and experiment submitted to GEO, to facilitate further use of the data by others, as well as to streamline the update procedures. For example, if new or corrected clone identity information becomes available later, the submitter can update the corresponding platform record.

At the time of this writing (actual state: December 2001), GEO contains data from 707 samples (GEO accessions GSM1 through GSM767) that are derived from 21 platforms (GPL4 through GPL24). Twelve experimental series (GSE1 through GSE12) have been annotated as well. Entering any of these accession numbers into the GEO search engine ('display tool') displays the corresponding platform, series, or sample record, which contains links to the other two types of components where available. For instance, a sample (GSM) record will provide links to the platform record for a given experiment, and to any annotated series the sample is a part of. Choosing 'view = full' or 'view = data' in the sample browser window will display the hybridisation values, which can also be downloaded as a tab-delimited ASCII (American standard code for information interchange) table.

Another powerful way of searching GEO entries is *via* the Entrez probe set interface (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo>). There, one can search for desired terms present in free text description, organism, etc.

7.8.12

GXD – The Mouse Gene Expression Database

The Mouse Gene Expression Database (GXD) is a community resource of information on gene expression in the laboratory mouse. Its objective is to integrate different types of data on endogenous gene expression and to make these freely available in an appropriate format. (Ringwald *et al.*, 2001). Data are acquired from the literature by curation staff and *via* electronic submission from laboratories. New expression data are made available on a daily basis.

Currently, GXD captures data from RNA *in situ* hybridisation and immunohistochemistry, Northern and Western blots, RT-PCR, RNase protection assays and cDNAs. The database will be extended to include microarray data and other types of expression data. By combining different types of expression data, GXD will provide more complete information on the transcripts and proteins produced by individual genes, and on the location, timing, and quantity of gene expression. In addition, information on variation in gene expression in different mouse strains and mutants are provided. Integration is achieved by storing primary expression data. Such primary expression data include the timing and tissue of expression, the genetic origin of the sample, the number and sizes of detected bands and sequence information. The molecular probe used, the expression assay type and the experimental conditions are also given. New data and new assay types can be added readily in this format. Novel insights from new data can be represented dynamically.

The spatial and temporal patterns of gene expression are described in a standardised way using an extensive dictionary of anatomical terms for the mouse (Bard *et al.*, 1998). The anatomical dictionary names the tissues and structures for each developmental stage, and organises the terms hierarchically from body region or system to tissue to tissue substructure. This model facilitates an integrated description of expression patterns for various assays with differing spatial resolution, and allows computational analysis of expression patterns at different levels of detail. Expression records are linked to digitised images of the original expression data.

GXD is integrated with the Mouse Genome Database (MGD) to enable a combined analysis of genotype, expression, and phenotype data. It also has comprehensive links to other databases such as PubMed, OMIM, sequence databases and databases from other species. As part of the Gene Ontology project (The Gene Ontology Consortium, 2001), shared controlled vocabularies for biological processes, molecular functions and cellular components have been assembled and have been assigned to mouse genes and their products. These classification schemes provide important search parameters for expression data. New query and display tools are constantly developed. Links to other resources that will extend GXD's utility for analysis of gene expression information will be added as they arise. GXD is implemented in the Sybase™ relational database management system. The database is accessible through the Mouse Genome Informatics website at <http://www.informatics.jax.org/> or directly at http://www.informatics.jax.org/menus/expression_menu.shtml.

7.8.13

ISG Database – Interferon-Stimulated Gene Database

The interferons are a broad family of diverse cytokines consisting in humans of the type I species with twelve interferon- α subtypes, interferon- ω and β and the type II species, containing a single member, interferon- γ . Interferons are primarily produced during viral infection and play an essential role in innate immunity by inducing the synthesis of proteins that inhibit the replication and spread of viral, bacterial and parasitic pathogens. They also modulate immune responses and have antiproliferative activities on different cells. Because of these properties, interferons are used to treat certain viral infections, specific cancers and multiple sclerosis (reviewed in Stark *et al.*, 1998).

The gene products regulated by interferons, known as interferon-stimulated genes (ISGs) are the primary effectors of the interferon response. While the functions of most of the interferon-stimulated genes remain to be elucidated, nearly all of the well studied interferon-stimulated genes play pivotal roles in host defence (Stark *et al.*, 1998; de Veer *et al.*, 2001). In order to obtain a more comprehensive understanding of the gene products induced by interferons, data from extensive microarray analyses of RNA samples collected from experiments on human and murine cell lines treated with interferon- α , interferons- β or interferons- γ were compiled (Der *et al.*, 1998; and unpublished data). The combined experimental data sets have identified over 330 potential interferon-stimulated genes. In order to obtain information on interferon-induced cellular phenotypes, the interferon-stimulated genes were assigned into functional categories based on the current literature, homology to other functionally characterised proteins and new defined experiments (de Veer *et al.*, 2001).

The ISG Database classifies interferon-stimulated genes using a series of progressively more specific functional groups. The data is freely available at <http://www.lerner.ccf.org/labs/williams/>. Individual entries can be retrieved by searching using keywords or by selecting a distinct functional category from the drop down menu. Selecting a functional category retrieves all entries assigned to it. Each individual entry is linked to GenBank and UniGene (NCBI, Bethesda, Maryland, USA) databases, from where the complete gene and corresponding protein sequence can be obtained. Each entry is also linked to the GeneCards database (Weizmann Institute of Science, Rehovot, Israel) that contains information on other gene names, links to literature, homologous sequences and a short functional description of the gene if available. Each gene entry also lists and links to the other functional categories it is assigned to in a hierarchical order. Simply clicking on a listed category retrieves the entries contained within it.

The ISG Database was constructed by assembling the individual entries into an extensive Microsoft Excel™ file and then converting the file to a tab-delimited text format. The text file was transferred to a computer running the Linux operating system. A second tab-delimited text file that mapped gene symbols to the functional categories defined in the study was similarly prepared and processed. The database can be interrogated *via* a web browser which provides output in the form of hypertext markup language (HTML) tables.

The ultimate goal of this effort is to compile a comprehensive archive listing all interferon-stimulated genes with links to expression data obtained by analysing further microarray experiments. Bryan William's laboratory (Lerner Research Institute, Cleveland, Ohio, USA) has used the information contained in the ISG Database to construct a custom microarray containing cDNAs from almost all the genes represented in the ISG Database. This array is being used to generate expression profiles following interferon treatment of a number of cell types.

7.8.14

The Kidney Development Database

The mammalian kidney has become a well-established system for investigating the molecular mechanisms of organogenesis in higher animals. The most important reasons for this are: (i) that kidneys develop well in organ culture, and are therefore accessible to experimental manipulation, (ii) that renal development involves most key developmental processes, such as induction, proliferation, apoptosis, differentiation, mesenchymal and epithelial morphogenesis, vascular and neural development, and (iii) that abnormalities of kidney development account for many human congenital diseases.

Research into kidney development at a molecular level began in earnest in the early 1980s. In an attempt to keep track of this information, gene expression data were first summarised in printed tables (Bard *et al.*, 1994). However, it soon became clear that the best way of disseminating the information would be as a database accessible *via* the internet. This database, one of the first gene expression databases to be available in this fashion, was launched in 1995, and has been updated continuously ever since.

The scope of the Kidney Development Database is deliberately narrow. It covers gene expression, mutant phenotypes and effects of culture manipulations during early kidney development. The data are stored and presented in a simple manner, partly because when the database was first designed, most web browsers supported only text.

Compared with gene sequence databases, which hold only one-dimensional information, databases of gene expression have to handle four-dimensional information (three spatial dimensions, and also time). One approach to solving this problem would be to hold three-dimensional 'pictures' of gene expression at all developmental times, but this is impractical for real data; generally, publications describing gene expression show only 'snapshots' at a few locations and times. The development of the kidney, however, is unusual, in that it is centrifugal – as the centre of the kidney matures, the outer edges begin to develop and to undergo the sequence of developmental changes that the centre has already undergone. This process is continuous, so that throughout foetal development, early stages of development can be found in the cortex of the kidney and later stages are found towards the medulla (Davies & Bard, 1998). Actual developmental time is therefore of less importance than it is on other systems (such as brain or limb) and can to all intents and purposes be ignored within the limits addressed by the database.

To reduce complex expression data to simple tables, the continuous processes of development of excretory nephrons and the branching of collecting ducts are divided into a few conventional 'stages', which are explained in a mini-essay attached to the database. These stages are then used as the basis of the gene expression tables, each of which describes the expression of classes of genes (*e.g.*, *hox* genes, matrix receptors, *etc.*) as 'present', 'absent' or 'data not available' in each of the stages of tissue development (Davies *et al.*, 1997).

The database can be searched by gene name or by gene type, and a second version of the database (available from the same home-page) can also be searched by expression pattern, so that a user can make queries such as "tell me all of the genes expressed in collecting ducts but not in the surrounding mesenchyme". Mutant phenotypes and the effects of blocking antibodies can be searched by gene or protein name, and are also cross-referenced from the gene expression tables. All entries are linked to full references.

The chief limitation of the kidney development database, as it currently stands, is that it cannot interact with other databases, in either direction. The main reason for this is lack of standardisation for gene expression data and databases (Verbeek *et al.*, 1999).

7.8.15

MAGEST – The Maboya Gene Expression Patterns and Sequence Tags Database

The Maboya Gene Expression Patterns and Sequence Tags Database (MAGEST) is a database for maternal gene expression information for the ascidian *Halocynthia roretzi* (Kawashima *et al.*, 2000). This ascidian, a lower chordate, has become an animal model in developmental biological research. It shows simple developmental processes, many of which are known to rely on the localised maternal factors (Satoh, 1994).

To study the nature of the maternal genetic information, several kinds of data have been deposited in MAGEST. Currently, MAGEST contains the expressed 3' and 5'-tag sequences from a fertilised egg cDNA library, the results of similarity searches and the expression data from whole-mount *in situ* hybridisation (WISH) experiments. In addition, a database, called translated MAGEST, for the amino acid fragment sequences predicted from the EST data sets has been constructed (Kawashima *et al.*, 2002). MAGEST, implemented in the PostgreSQL relational database system, is accessible through the internet at <http://www.genome.ad.jp/magest/>.

Each cDNA in MAGEST is given a unique clone identification number consisting of six alphanumerical characters. After removal of vector DNA sequences and ambiguous regions that contain stretches of unknown bases 'N' from raw sequences, the processed data is registered in MAGEST. These sequences are used as query sequences for BLAST homology searches against GenBank (NCBI, Bethesda, Maryland, USA) at the nucleotide level and also against the a non-redundant protein sequence database of GenBank constructed from several databases at the amino acid level. All 3'-expressed sequence tag (EST) sequences are compared

with each other to examine the numbers of redundant genes, which should reflect the population of maternal mRNAs. Based on these search results, each clone is annotated.

WISH is carried out for the eight cell stage, the 110 cell stage and the bud-tailed stage of embryonic development to obtain information on the localisation and expression sites of the clones during embryogenesis. Expression data is classified according to each blastomere or tissue. This classification is based on the cell lineage study of the ascidian (Nishida, 1987). A clone showing no specific expression is classified as 'overall' when the staining is seen over the entire embryo, or 'ISH-done' when no signal is detected in the *in situ* (ISH) hybridisation. The original images of WISH are provided with the classification data. Using this database, maternal mRNA expression has been used to generate a catalogue of ascidian development (Makabe *et al.*, 2001).

7.8.16

MethDB – The DNA Methylation Database

DNA methylation is a common and fundamental characteristic of most genomes. Specific DNA methyltransferases are responsible for the post-replicative methylation in the carbon-5 position of cytosine. Because DNA methylation is coding-neutral (*i.e.* cytosine and 5-methylcytosine are identically transcribed into mRNA and subsequently translated into proteins), nature has used DNA methylation to encode additional information that is not stored in the genetic code. This sort of heritable, but potentially reversible modification of the genetic material is now accepted to code for 'epigenetic' information. Methylation in the regulatory regions of genes can cause for instance repression of the transcription of this gene. For reviews on this topic, see, for example, Costello & Plass (2001).

DNA methylation data can be very heterogenous in nature and range from the global estimation of the total amount of 5-methylcytosine in a DNA sample to the precise sequence of methylated cytosine in a single DNA molecule. The DNA Methylation Database (MethDB) was created to store the available data in a single common and freely available resource (<http://www.methdb.net>) (Grunau *et al.*, 2000 and 2001). It contains 720 DNA methylation content values, 4,430 sequence specific methylation patterns and profiles for a total of 77 genes or loci in 42 species (actual state: September 4th, 2001). In MethDB, methylation content refers to the average relative 5-methylcytosine content of a DNA sample.

A methylation pattern is the sequence of the five nucleotide bases (including 5-methylcytosine) in a single DNA molecule. Methylation profiles are representations of the average methylation at cytosine sites along sequences. Both, global changes in the methylation state of DNA and local alterations in the 5-methylcytosine patterns can be of epigenetic importance.

To search MethDB, the user enters www.methdb.net and chooses 'search'. To meet the requirements of both novice and expert users two basic access routes are available: (i) a simple search field or a browse mode and (ii) four detailed query masks to search in methylation content data, in methylation profiles and pat-

terns, as well as in environmental conditions that can influence the DNA methylation state. In addition, it is possible to match the experimental results of the user with methylation content values in the database.

In the simple search field, single search terms like ‘tumor’, ‘diet’, “human” or ‘kidney’ can be entered. Logical operators cannot be used to combine them. Pressing the “Quick Search” button delivers a list of up to 50 experiments that match the search terms. Clicking on the ‘details’ link provides the user with background information about the experiment. In general, hypertext links lead to local information about the term that is linked-out or give direct access to other data sources on the internet (PubMed, OMIM, GenBank, on-line encyclopaedias, *etc.*).

Users who have already precise ideas of what they are looking for, and who are for instance interested in methylation patterns, will follow the link on the main search pages that brings them to a specific search mask. Here, species name, sex, tissue and the gene of interest can be entered. The user may choose a tumour or healthy tissue phenotype, or both. Search terms are connected with the logical operator ‘AND’. By choosing ‘IS NOT’ from the menu in front of the search field, entries can be excluded. Blank search fields are ignored and all entries are shown. Again, the query delivers a list of experiments that match the search terms. A click on the ‘details’ link gives access to information about the experiment including a graphical representation of the methylation profiles or patterns, a link to the investigated sequence with exon/intron structure and location of the regulatory regions (if available), and expression data of the gene that was studied. The method is described that was used to generate the methylation data. If PCR was used, the reaction conditions are available. The utilisation of the query masks for methylation content data and environmental conditions follow the same scheme. Data sources for MethDB are the related literature and direct submissions.

7.8.17

EMAGE – The Edinburgh Mouse Atlas Gene Expression Database

The study of normal and aberrant gene function requires tools for comparative analysis of very large amounts of information that is essentially spatio-temporal in nature (*e.g.*, gene expression, mutant phenotype, cell behaviour and cell lineage, *etc.*). One approach is to map the data into a common framework. The Edinburgh Mouse Atlas Project (EMAP) aims to provide such a framework for bioinformatics at the organism level. In EMAP, attention was focused on the mouse embryo, but the tools that have been built are generic and can be extended to other systems. The Edinburgh Mouse Atlas Gene Expression (EMAGE) database uses the EMAP framework to bring diverse *in situ* gene expression data into a common context for query and analysis. EMAP can also be applied to other spatially organised data in the embryo, for example, the textual part of the framework forms the basis of the GXD gene expression database. There are also plans to use EMAP as a framework for a database of mutant phenotypes. See Davidson *et al.* (2001a) for a review.

The EMAP atlas is based on the paper atlases of Kaufman (Kaufman, 1992) and Theiler (Theiler, 1989). There are three components: a stage-by-stage, hierarchical anatomical nomenclature (Bard *et al.*, 1998), a series of three-dimensional volumetric models of embryos at consecutive stages and, linking these, a set of 'painted' anatomical domains in the models corresponding to the major named parts (Brune *et al.*, 1999). Each model shows the histological structure of the embryo in photographic detail and can be digitally re-sectioned in any plane or viewed as a whole embryo. This facilitates comparison with the user's specimens of corresponding stage. Three-dimensional models are being made available in a progressive series of draft releases with increasing resolution and functionality. EMAP is available in the internet at <http://genex.hgu.mrc.ac.uk/> and on CD-ROM.

Many important aspects of *in situ* gene expression data cannot be adequately captured in textual form. This is especially striking when tissues and organs are in the process of forming. The EMAGE database addresses this problem by capturing gene expression data using both the textual and space/time components of the EMAP atlas framework. The database will provide a resource for the scientific community, including data from large-scale expression screens as well as data from individual, hypothesis-driven experiments. Using published, standard procedures that include quality control checks, each expression pattern will be mapped to the appropriate model embryo in the atlas. Thus, large numbers of patterns will be accessible to spatial queries and can be analysed together in the appropriate embryonic context. The raw gene expression images will be stored alongside the mapped data so that the original pattern can still be examined – and, if desired, re-mapped to facilitate more detailed three-dimensional analysis. The application of this concept ranges from low-resolution mapping to facilitate simple database searching, to advanced three-dimensional contextual analysis and hypothesis generation (Davidson & Baldock, 2001).

Any gene expression data of adequate quality that can be attributed to spatial regions in the embryo will be accepted, *e.g.*, *in situ* hybridisation patterns, protein immunohistolocalisation, reporter expression patterns, microarray assays of expression in spatially defined regions, for example, in tissue samples captured by laser dissection microscopy. Raw data may be in a variety of forms, *e.g.*, images of whole-mount preparations or of histological sections, or textual descriptions. The expression pattern must be accompanied by minimal information on specimen and probe/antibody that will allow the results to be interpreted. Selected published, retrospective data will also be included in the database.

The data in this internet-accessible, public database will be free or at nominal cost to academic sites. Data can be submitted and interrogated *via* interactive, Java-based interfaces. Expression patterns can be mapped at the bench by the researcher or an editorial office staff with feedback from the researcher. Data may be submitted in electronic form as images or as actual specimens. In addition to the public database, a personal database is available that has the same data input and supports simple data management queries. Entry of data into a personal database can be used as an initial step in submission to the public database.

The EMAGE database with selected gene expression patterns will be accessible from <http://genex.hgu.mrc.ac.uk/> from September 1st, 2002 or earlier. It is planned to link the EMAGE database with the GXD database to form an integrated resource – the Mouse Gene Expression Information Resource (MGEIR) (Ringwald *et al.* 1994).

7.8.18

RAD – The RNA Abundance Database

High-throughput technologies for gene expression analysis require novel data management systems. Before these data can be shared for purposes of publication and/or further analysis the experimental annotations, the methodology used, the raw data, and the processing of the data need to be described in sufficient detail. The RNA Abundance Database (RAD) was built to respond to these needs (Stoeckert *et al.*, 2001) and in recognition of the heterogeneity of systems studied by current and potential collaborators. A main intention was to develop a common resource that would hold gene expression data from filter arrays, Affymetrix® arrays, two-colour microarrays, and SAGE, and that would also be extensible to novel technologies. Such a resource would allow investigators to extract information on gene expression regardless of the method used. Thus, RAD was developed as a generic system to capture gene expression data from various types of experiments performed in a variety of biological systems. RAD also contains technology for mining gene expression data.

RAD is implemented using a commercial relational database management system (Sybase™ and Oracle™). The RAD schema is available online at <http://www.cbil.upenn.edu/cgi-bin/RAD2/schemaBrowserRAD.pl>. Files for generating RAD are available at <http://www.cbil.upenn.edu/downloads/RAD/>. The RAD schema is conceptually organised into groups of tables containing particular types of information (see <http://www.cbil.upenn.edu/RAD2/schema.html>). During the development of RAD and subsequent to it, an international grass-roots effort arose to facilitate the adoption of standards in data representation for microarray experiments. As a result of the participation in the Microarray Gene Expression Database (MGED) group the guidelines for minimal information about a microarray experiment (MIAME) have been incorporated (Brazma *et al.*, 2001). A microarray gene expression (MAGE) object model and XML-based language have been developed for data exchange.

A website demonstrating the features of RAD using microarray, Affymetrix® array and SAGE studies is available at <http://www.cbil.upenn.edu/RAD2/>. Moreover, experiments stored in RAD are also presented in websites focused on the malarial parasite, *Plasmodium falciparum*, (The *Plasmodium* Genome Resource Database; PlasmoDB) and on the endocrine pancreas (The Endocrine Pancreas Consortium Database; EPConDB). At PlasmoDB (<http://plasmodb.org>), SAGE and microarray experiments are used to provide additional lines of evidence that genes predicted in the emerging genomic sequence are real and to indicate where they may be expressed (Bahl *et al.*, 2002). At EPConDB (<http://www.cbil.upenn.edu/EPConDB/>). Data and experimental information on microarray experiments

using pancreatic tissues from human and mouse can be queried. Queries can combine expression annotations (*e.g.*, tissue) with annotations associated with genes of interest (*e.g.*, gene ontology functions and chromosomal locations). This powerful functionality is made possible by a sister relational database, Genomic Unified Schema (GUS), designed to integrate genomic, transcribed, and protein sequences along with their associated features (Davidson *et al.*, 2001b). PlasmoDB and EP-ConDB demonstrate the capacity of RAD to contain disparate types of gene expression data and to extend the value of those data by combining them with sequence and sequence annotations.

7.8.19

The Rochester Muscle Database

The Rochester Muscle Database is a catalogue of SAGE (serial analysis of gene expression) tags obtained from normal human *vastus lateralis* muscle. The database is a free public resource and can be viewed or downloaded at <http://www.urmc-rochester.edu/smd/CRC/SWindex.html>. The website has instructions for using and downloading the database, and links to other webpages that facilitate identification of the genes corresponding to the SAGE tags. Over 21,000 unique SAGE tags were detected from a total of approximately 108,000 tags. An equal number (approximately 54,000) were obtained from a pool of RNA extracted from muscle of eight healthy young men (at the age of 21 to 24 years) and a pool of RNA extracted from eight healthy older men (at the age of 66 to 77 years). Every tag detected in the study is listed along with the number of times it was detected in the sample from young muscle and the number of times it was detected in the sample from old muscle. Details of the study have been published (Welle *et al.*, 1999 and 2000).

Muscle was obtained from subjects in good health and with normal neuromuscular function as determined by a panel of laboratory tests and examination by a neuromuscular disease specialist. None was involved in weight lifting or extreme levels of aerobic exercise. All subjects were asked to refrain from activities more strenuous than walking for three days before the biopsy. They received a standard breakfast about 90 minutes before the biopsy. Needle muscle biopsies were obtained from the *vastus lateralis* muscles. The tissue was removed within a few minutes of administration of the local anaesthetic. It was quickly frozen in liquid nitrogen, then stored at -70 degrees centigrade until RNA extraction. The total RNA was extracted from each piece of tissue, then was pooled for extraction of polyadenylated RNA. The polyadenylated RNA was examined according to the protocol provided by the originators of the SAGE method (Velculescu *et al.*, 1995).

The database demonstrates a very high abundance of mitochondrial mRNA in human muscle, accounting for 17 percent of the tags catalogued. Other abundant transcripts include those encoding contractile proteins, enzymes of glucose and glycogen metabolism, and ribosomal proteins. The results correlate well with transcript concentrations inferred from databases of longer ESTs derived from muscle

RNA (Bortoluzzi *et al.*, 2000). They also correlate well with relative mRNA levels determined with the Affymetrix® HG-U95A high-density oligonucleotide array (unpublished data).

7.8.20

SAGEmap – The serial analysis of gene expression tag to gene mapping database

The serial analysis of gene expression (SAGE) tag to gene mapping database (SAGEmap) is a public data repository (Lal *et al.*, 1999; Lash *et al.*, 2000) designed to ease the analysis and comparison of SAGE data. In SAGE, gene expression profiles in various biological samples are determined indirectly, by obtaining counts of short, nine to eleven base pair tags (Zhang *et al.*, 1997) that are assigned to genes. At the time of this writing (actual state: December 2001), 100 SAGE libraries are represented in SAGEmap, with the total of 4,280,231 tags.

Given both the limited information content of such oligonucleotide tags, and the frequency of errors in many GenBank (NCBI, Bethesda, Maryland, USA) sequences derived from expressed sequence tag (EST) sequences, it is not always possible to produce a unique tag-to-gene or gene-to-tag mapping. Basing the tag assignment on UniGene (NCBI, Bethesda, Maryland, USA) sequences instead of expressed sequence tags is one possible way to increase the reliability of mapping. Errors in sequencing of the tags themselves, leading to errors in tag counts, must also be considered, especially where rare transcripts are concerned.

Information on two-way mapping between SAGE tags and UniGene cluster identifications can be searched or downloaded. The SAGE data itself can be searched by either the tag sequence, or the UniGene identification (follow the ‘Tag to Gene Mapper’ or the ‘Gene to Tag Mapper’ links on the left-hand side of the main SAGEmap webpage). If the tag or the UniGene identification are present in the database, the search results will be displayed in table format showing SAGE library names where this tag is found, followed by the tag frequency (count per million), the graphic spot showing the tag’s relative expression, the tag count, and the total number of tags in each of the reported libraries. At the bottom of the results page, the summary of UniGene clusters found for this tag is displayed.

The SAGEmap site also links to such National Center for Biotechnology Information resources as Virtual Northern, xProfler, and Library Browser (follow the ‘About tools’ link on the left for more information about them), as well as to the submission tool, and to the detailed descriptions of all the procedures involved in establishing the mapping between tags and genes (explore the links under the information heading). Other links provide access to the public Cancer Genome Anatomy Project (CGAP) of the National Cancer Institute (NCI, Bethesda, Maryland, USA) and to two pre-compiled “Differential expression analysis queries” (colon cancer *versus* normal colon, and brain tumour *versus* normal brain). In each of these analysis queries, two normal tissue samples are compared with two malignant samples of the same tissue type, and the lists of genes that best distinguish between these groups are available for browsing and downloading.

7.8.21

SGD – The *Saccharomyces* Genome Database and its Expression Connection

As microarray technology becomes more widely used, it is becoming ever more difficult for researchers to access and interpret data that are usually presented in heterogeneous formats by widely scattered research groups. As an aid to yeast researchers, the *Saccharomyces* Genome Database (SGD; accessible at <http://genome-www.stanford.edu/Saccharomyces/>) (Dwight *et al.*, 2002) has created a web-based tool to allow a researcher to use gene names to query datasets derived from yeast experiments using microarrays. SGD's Expression Connection is currently querying eleven datasets produced by laboratories at six locations (Rosetta Inpharmatics, Kirkland, Washington; Stanford University, Stanford, California; Cold Spring Harbor Laboratories, Cold Spring Harbor, New York; The Whitehead Institute, Cambridge, Massachusetts; University of California, San Francisco, California; University of Missouri, Columbia, Missouri, USA) (Ball *et al.*, 2001). The datasets were derived from the following publications: Cho *et al.* (1998), Chu *et al.* (1998), DeRisi *et al.* (1997), Ferea *et al.* (1999), Gasch *et al.* (2000 and 2001); Lyons *et al.* (2000), Ogawa *et al.* (2000), Roberts *et al.* (2000), Spellman *et al.* (1998) and Wyrick *et al.* (1999). As more datasets become available, SGD curators will retrieve and format them for use with Expression Connection. Queries of Expression Connection can be made at the following URL: <http://genome-www4.Stanford.EDU/cgi-bin/SGD/expression/expressionConnection.pl>.

Using software developed at SGD, downloaded data from several microarray studies are simultaneously queried to display the expression pattern of a given gene and its most similarly expressed neighbours. When retrieved, the results display the gene names, a red-to-green display indicating the relative expression level of the gene as detected on the array, as well as the gene product's biological process, molecular function and cellular component as curated at SGD using the Gene Ontology (Ashburner *et al.*, 2000). In addition to the red/green display, time series data are represented as a line graph. Since a single gene product may have more than one Gene Ontology annotation, Expression Connection permits users to follow a link to a webpage that details all the Gene Ontology annotations for the group of genes being viewed. Each microarray data report in Expression Connection is complemented by hyperlinks to SGD, to a browser that allows users to explore the microarray data as well as to any supplementary websites provided by the authors of each study.

7.8.22

SMD – The Stanford Microarray Database

The Stanford Microarray Database (SMD) was initially established to support research within the joint laboratory of David Botstein and Patrick O. Brown and has now been extended to support the entire Stanford community (Sherlock *et al.*, 2001). Over time, the role of this database has expanded considerably. Now, in addition to storing both raw and normalised data acquired from micro-

array experiments performed by Stanford researchers and their collaborators, SMD also provides analysis tools, public access to published microarray results, and the source code underlying a working, scalable, open-source microarray database. The following section describes these various aspects of SMD.

SMD is the largest academic microarray database in the world, containing 18,793 experiments (as of November 2001). Local researchers acquire raw data from two-channel, spotted array experiments using either GenePix (http://www.axon.com/GN_GenePixSoftware.html) or Scanalyze (<http://rana.lbl.gov/EisenSoftware.htm>) software. Data from single or multiple experiments can be loaded into SMD. The experimenter determines the method of data normalisation upon entry into the database. The normalisation method, individual spot flagging, and the description and classification of each experiment can be changed by the experimenter at any time. Because of its importance to interpretation, biological annotation of spotted samples is frequently updated. Data can be selected or filtered on a spot-by-spot basis using any spot parameter, and the resulting data can be analysed by hierarchical clustering or self-organising maps. Additional analysis tools provide measurements of reproducibility, spot-metric plotting, histogram distributions of channel intensities and ratios, expression 'histories' of spotted samples, and distributions of ratios physically across an array.

Although not a substitute for the fledgling international repositories for published microarray data (Gene Expression Omnibus at the National Center for Biotechnology Information; <http://www.ncbi.nlm.nih.gov/geo/> and ArrayExpress at the European Bioinformatics Institute; <http://www.ebi.ac.uk/arrayexpress/>) (Gardiner-Garden & Littlejohn, 2001), SMD endeavours to make local data world-viewable either upon publication or at the experimenter's discretion. Utilising the same integrated analysis tools that are available to Stanford researchers, world-users can analyse organised and annotated experiment sets associated with publication records. Alternatively, all public data may be downloaded by FTP for offline analysis. Currently, the results of 1,567 hybridisations within SMD are public, associated with 34 distinct publications (actual state: November 2001).

Both the database structure and the overlying software interface have been released to the community in the hope that they will benefit other groups and stimulate additions and improvements. Compartmentalisation of institution-specific information into configuration modules and a package-installer make it easy for other institutions to use SMD. SMD uses Oracle™ as the database management system, Perl for software interfaces, and Apache for the web server.

The web-based interface of SMD allows researchers to access data remotely using common computer platforms. In addition, software improvements on the server are reflected on all client machines instantly. Access to experimental data is regulated at both the individual user and research group level, allowing both collaboration and confidentiality. Since SMD can serve an entire institution, individual labs need not to purchase or develop and maintain separate database systems. Free access to the SMD source code allows groups to customise the software to accommodate new analysis tools and tailor the software to their own needs. As microarray analysis techniques develop, the open nature of SMD will allow new methods to

be incorporated as they become available. The SMD package continues to be improved both at Stanford and elsewhere. Updated software is periodically released, and the research community is encouraged to submit new tools and enhancements to the project.

The SMD staff are working to make the database compliant with the standards recommended by the Microarray Gene Expression Database Group (MGED; <http://www.ebi.ac.uk/microarray/MGED/>). In order to provide the suggested 'minimum information about a microarray experiment' (MIAME) (Brazma *et al.*, 2001), improvements to sample annotation, experimental design, and array element annotation will be made. Upcoming enhancements to the Stanford Online Universal Resource for Clones and ESTs (SOURCE; <http://genome-www.stanford.edu/source>) will continue to incorporate the latest biological annotations from publicly accessible databases for human, mouse, and rat clones. Used in conjunction with SMD, SOURCE is an invaluable tool for interpreting microarray experimental results. SMD can be accessed at <http://genome-www.stanford.edu/microarray>.

7.8.23

TRIPLES – The Database of Transposon-Insertion Phenotypes, Localisation, and Expression in *Saccharomyces cerevisiae*

TRIPLES is the web-accessible Database of Transposon-Insertion Phenotypes, Localisation, and Expression in *Saccharomyces cerevisiae*. This is a relational database housing approximately half a million data points generated from an ongoing study using large-scale transposon mutagenesis to characterise gene function in yeast (Kumar *et al.*, 2000a). At present, TRIPLES encompasses three data sets (phenotypic data, protein localisation data and expression data) derived from analysis of nearly 3,800 yeast genes (60 percent of the yeast genome); each data set is described in detail. In particular, TRIPLES provides data regarding the subcellular localisation of over 1,000 yeast proteins drawn from immunofluorescence analysis of approximately 12,000 strains carrying epitope-tagged proteins. This localisation data is supplemented with over 500 fluorescence micrographs of staining patterns observed upon immunolocalisation of epitope-tagged proteins, establishing TRIPLES as one of the largest visual libraries of yeast protein localisation and a central database of yeast gene function. TRIPLES may be accessed free-of-charge at <http://ygac.med.yale.edu>.

The methodology of the transposon-based study underlying TRIPLES is straightforward (Kumar *et al.*, 2000b). Transposon mutagenesis is performed on a genome-wide scale using a specially designed multifunctional transposon, such that each insertion may be used to derive a variety of informative alleles including reporter gene fusions, gene disruptions, and epitope-tagged alleles (Ross-MacDonald *et al.*, 1999). Gene fusions to transposon-encoded *lacZ* provide a means of generating expression profiles identifying sequences translated under given growth conditions. As this *lacZ* reporter is terminated by a series of stop codons, transposon insertion also results in truncation of its host gene, thereby potentially generating disruption alleles for subsequent phenotypic analysis. Fi-

nally, by means of Cre-lox recombination, an integrated transposon insertion (carrying two lox sites) may be modified such that the bulk of the transposon is excised, leaving behind a short stretch of epitope-coding sequence. These epitope-tagged alleles can be used to generate corresponding tagged proteins for immunolocalisation. By this approach, a single insertion is sufficient to yield expression, phenotypic, and localisation data. Three data sets are maintained and updated in TRIPLES.

All data within TRIPLES may be conveniently accessed through a variety of search options. Each data set may be queried for records regarding any gene of interest. Alternatively, searches may be initiated by selecting from a list of controlled vocabulary terms describing each data set. For example, TRIPLES may be queried for all tagged proteins localising to the nucleus by initiating a search of localization data with 'nucleus' chosen as the localisation. To help integrate these varied data sets, the results from a given search may be used to initiate further searches of similar or different data types. Furthermore, composite searches can be used to retrieve records from multiple data sets for any given gene or insertion. Users may also browse data within TRIPLES by means of a convenient graphic map accessible through the 'Display Insertion Map' link incorporated into the basic search forms. The resulting map displays all transposon insertions within a chosen gene; composite records corresponding to each insertion may be viewed by simply clicking on an insertion sign of interest.

Data within TRIPLES is principally presented in tabular format and may be downloaded as tab-delimited text for import into any standard spreadsheet or data analysis software. The TRIPLES site provides a variety of output-formatting options. Reports may be custom-formatted to display only those fields of greatest interest. In addition, users may sort output reports by data field in order to group results in a logical manner. Each complete data set is also available for downloading as a flat file. Since the phenotypic data set often is very large, the opportunity is provided to download results from each phenotypic assay individually, thereby generating files more amenable to standard spreadsheet analysis.

In total, TRIPLES constitutes a unique resource fostering yeast molecular biology. It is a medium through which both data and reagents may be shared among the scientific community. Unlike other yeast databases, TRIPLES presents online order forms linked to each output report, so that academic users may immediately request any desired strain generated in this project free of charge by submitting a completed form. TRIPLES also provides access to corresponding data within a variety of relevant sites including the *Saccharomyces* Genome Database (SGD) (Ball *et al.*, 2001), Yeast Proteome Database (YPD) (Costanzo *et al.*, 2001), and GenBank (Wheeler *et al.*, 2001). Furthermore, the TRIPLES database is featured in the new SGD tool 'Function junction' (Ball *et al.*, 2001), a helpful resource offering access to data from a variety of functional genomic projects. Collectively, these resources exemplify a collaborative spirit integral to continued advancements in the field of molecular biology.

7.8.24

UTRdb and UTRsite – The Specialised Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs

UTRdb is a specialised database of 5' and 3'-untranslated region (UTR) sequences of eukaryotic mRNAs (Pesole *et al.*, 2000a). 5'-UTR sequences are defined as the mRNA regions from the cap site to the starting codon, whereas 3'-UTR sequences are the mRNA regions from the stop codon to the poly(A) starting site.

UTRdb is automatically generated from the EMBL nucleotide database (Stoesser *et al.*, 2001) through a syntactic parsing of the feature table of the relevant EMBL entries that allows definition of UTR regions even when these have not been reported in the primary EMBL entry. UTRdb entries contain additional information not present in the primary EMBL or GenBank database such as completeness or incompleteness of the UTR, number of spanning exons, presence of repetitive elements and presence of UTR-specific functional patterns for which experimental evidence of a role in post-transcriptional regulation of gene expression exists. All annotated functional patterns are collected in the UTRsite database against which unknown mRNA or genomic sequences can be compared for UTR-specific regulatory elements.

To eliminate redundancy entries with a similarity and overlapping degree with longer entries in UTRdb of greater than 95 and 90 percent respectively are removed (Grillo *et al.*, 1996). Possible vector DNA contamination is also annotated. The current version of UTRdb (release October 2001) contains 247,548 entries, 38 percent of which proved to be redundant, totalling 64,060,991 nucleotides. UTRdb is organised in eight sequence collections, one for each of the eukaryotic divisions in the EMBL and GenBank databases (human, rodent, other mammals, other vertebrates, invertebrates, plant and fungi) plus viruses.

All UTRdb and UTRsite internet resources implemented for retrieval and functional analysis of 5' and 3'-UTR sequences are publicly accessible at <http://bighost.area.ba.cnr.it/BIG/UTRHome/> (Pesole & Liuni, 1999). These include retrieval of UTRdb and UTRsite using the Sequence Retrieval System (SRS) (Etzold *et al.*, 1996), which is also available at the European Bioinformatics Institute internet server (<http://srs.ebi.ac.uk:80/>), UTRscan and UTRblast. The UTRscan utility allows the enquirer to search submitted sequences for any of the patterns collected in UTRsite. The UTRblast utility allows database searches against fully annotated UTRdb entries.

The availability of UTRdb and UTRsite databases may allow the user to carry out systematic investigations of the structural features of mRNA UTRs, to study their evolutionary dynamics and to reveal novel regulatory elements by pattern discovery algorithms (Davuluri *et al.*, 2000; Pesole *et al.*, 1997; Pesole *et al.*, 1999; Pesole *et al.*, 2000b; Pesole *et al.*, 2000c; Pesole *et al.*, 2001).

7.8.25

yMGV – The Yeast Microarray Global Viewer

Microarrays are used to provide a genome-wide response to a specific biological question. Often, experimenters use only a small part of the dataset, while much potentially interesting data is discarded. Mining of available transcriptome data thus provides a new way for biologists to find interesting hints on gene function. Normalisation of microarray procedure, information and file format are being improved (see <http://www.mged.org>) which will permit more rapid development of this approach.

The Yeast Microarray Global Viewer (yMGV) is the biggest freely accessible yeast transcriptome database in the world. It is available at <http://transcriptome.ens.fr/ymgv/> and it provides a simple intuitive interface to mine transcriptome data. To date (November 2001), this database contains 1,030 whole genome yeast transcriptome analyses from 50 publications, encompassing most of the yeast microarray datasets published since 1997. Data are those furnished by authors and have not been filtered or normalised. The description and statistics of each dataset are available online.

yMGV may be used to investigate three distinct questions: (i) Under what conditions is a gene transcriptionally regulated? (ii) What are the genes regulated by specified conditions? (iii) Which are the genes sharing similar transcription profiles?

The following will briefly describe these three database features. Full description of the database has been previously published in Le Crom *et al.* (2002) and Marc *et al.* (2001).

The basic feature requests the user to provide the name of a gene or of an open reading frame and to specify the cut-off for differences in gene expression. For example, if one chooses to view experiments where PDR5 (a yeast ABC transporter) transcription is altered by a factor of three or more, yMGV will display a one page report containing the gene information from the *Saccharomyces* Genome Database (Ball *et al.*, 2000), selected transcriptome data and links to PDR5 pages in external databases. Transcriptome data are displayed as a set of histograms for each publication in which the threefold limit is reached. Each set of histograms represents the $\log_2$ ratio observed in the corresponding experiments. A click on this representation displays more details (ratio, experimental conditions, mean and standard deviation of experiment data, distribution graphics, *etc.*). The second feature asks the user to enter a criterion for gene selection defined by experiment name and cut-off for significance. yMGV then searches for genes which fulfil all criteria and lists all genes sharing this regulation profile. The third feature is still undergoing testing at the time of writing. In this feature, the transcription profile of a given gene is compared to that of other genes using a similarity coefficient which ranges from plus one (identical profile) to minus one (opposite profile). The user enters a gene and a limit for the similarity coefficient, the yMGV output is an ordered list of genes fulfilling these criteria. A link allows comparison of visually selected gene transcriptional profiles.

7.9

Protein-based gene expression databases and analyses tools

7.9.1

Introduction to protein-based gene expression databases

Proteins are the main agents of physiological activity in a cell or an organism. The protein profile is the complete inventory of proteins of a cell or tissue. The composition of the protein profile depends on the *in vivo* situation which is affected by disease mechanisms, drug effects and other influences. Conversely, the protein profile affects cell health, cell response to drugs and so on. The characterisation of protein profiles and changes within them reveals important information for the understanding of disease and therapy as well as for the unravelling of fundamental cellular processes (Cavalcoli, 2001; Chakravarti *et al.*, 2000).

Complete genome sequences of several organisms have become available during the last few years (see the Entrez Genomes website of the National Center for Biotechnology Information at <http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html> for an actual overview). The idea of systematically mapping and identifying the total protein complement of a cell – the proteome – arose over 20 years ago when the separation of proteins from total cell extracts by two-dimensional gel electrophoresis was developed (Anderson *et al.*, 2001; James, 1997). This is noteworthy, because the effort to produce an profile of all human proteins – the human protein index (HPI) began even before the initiation of the human genome programme (Anderson *et al.*, 2001). However, there are two main reasons why the characterisation of mRNA expression profiles is more widespread than the characterisation of proteins: First, mRNA expression profiling is inherently simpler and more scalable than protein analytical technology. During the last decades, several different approaches have been developed for mRNA profiling (see, for example, chapters 3 and 4 of this book). Second, the annotation of the human genome indicates a surprisingly low number of somewhere between 30,000 and 40,000 genes (Lander *et al.*, 2001; Rappsilber & Mann, 2002; Venter *et al.*, 2001), whereas the estimated number of proteins encoded by these genes is at least two to three times higher (Rappsilber & Mann, 2002) because of widespread post-transcriptional and post-translational modifications of mRNAs and proteins in the cell (see chapter 1).

New and improved technologies for the separation, characterisation and identification of proteins (chapter 5 of this book) have now started to make possible what was begun more than 20 years ago. The resolving power of two-dimensional polyacrylamide gel electrophoresis and its full analytical and preparative potential have been improved with special emphasis on reproducibility and standardisation of protein spot patterns, enhanced protein detection sensitivity, computer analysis and database development. This has allowed a very large amount of data to be generated, primarily in the form of two-dimensional polyacrylamide gels combined with information on the proteins identified from the gel spots.

The idea of storing this data in a comprehensive, accessible database that integrates structural, functional and physiological information also goes back to the end of the 1980s (Aebersold & Leavitt, 1990; Celis *et al.*, 1989 and 1990a; Neidhardt, 1989; Santaren, 1990; Simpson *et al.*, 1992a). Primordial protein expression databases based on two-dimensional polyacrylamide gel electrophoresis were already constructed at this time (see table 7.6 for examples). These databases soon linked the pure protein identity data to information on structure, function and physiological role and to genome data (*e.g.*, Celis *et al.*, 1991c and 1993; Leffers *et al.*, 1996). Table 7.8, the NCI 2DWG Image Meta-Database at <http://www-lecb.ncifcrf.gov/>

Table 7.6 Early protein databases.

Name	References
Comparative two-dimensional gel protein database of the intact and regenerating newt limbs	Tsonis <i>et al.</i> (1993)
Comprehensive database of proteins from the urine of patients with bladder cancer	Rasmussen <i>et al.</i> (1996)
Comprehensive two-dimensional gel protein database of non-cultured unfractionated normal human epidermal keratinocytes	Celis <i>et al.</i> (1991b, 1992 and 1994)
Gene-protein database of <i>Escherichia coli</i> K-12	VanBogelen <i>et al.</i> (1990)
HL-60 protein database	Manabe <i>et al.</i> (1995)
Human lymphoid protein database	Hanash <i>et al.</i> (1998)
Rat liver epithelial (RLE) cell nuclear protein database	Wirth <i>et al.</i> (1991 and 1993)
REF52 protein database	Garrels & Franza (1989); Garrels <i>et al.</i> (1990)
MRC-5 human embryonic lung fibroblast two-dimensional gel cellular protein database	Celis <i>et al.</i> (1990a)
Mouse liver protein database	Giometti <i>et al.</i> , 1992
<i>Saccharomyces cerevisiae</i> protein database	Garrels <i>et al.</i> (1994); Latter <i>et al.</i> (1995)
Transformed amnion cells (AMA) master database	Celis <i>et al.</i> (1990b and 1991a)
Two-dimensional protein database of human heart	Kovalyov <i>et al.</i> (1995)
Two-dimensional database of mouse liver proteins	Fountoulakis <i>et al.</i> (2000 and 2001)
Two-dimensional gel protein database of <i>Saccharomyces cerevisiae</i>	Boucherie <i>et al.</i> (1996)
Two-dimensional rat brain protein database	Fountoulakis <i>et al.</i> (1991)
Two-dimensional polyacrylamide gel electrophoresis protein database of adult <i>Drosophila melanogaster</i>	Ericsson <i>et al.</i> (1997)

2dwgDB/ and the Partial List of Web 2D Electrophoretic Gel Databases at <http://www.lecb.ncifcrf.gov/EP/table2Ddatabases.html> are helpful for overviews on protein databases that were available in the internet (Please note: these databases are not maintained any longer).

In the last 20 years, much effort has been devoted to the standardisation of protein expression analyses (Appel *et al.*, 1996a), to the establishment of protein expression databases in the internet (Hoogland *et al.*, 1997; guidelines for building a standardised two-dimensional database at <http://www.expasy.org/ch2d/fed-rules.html>), and to detection, identification and image analysis. Recent developments include the mining of protein data (Oh *et al.*, 1999), the functional annotation of proteins (Protein Information Resource, PIR, at <http://pir.georgetown.edu/>) (Wu *et al.*, 2002), development of databases of proteins and peptides evolved *in vitro* (Artificially Selected Proteins/Peptides Database, ASPD, at <http://www.sgi.sccc.ru/mgs/gnw/aspd/>) (Valuev *et al.*, 2002), development of databases for the studying of cellular networks of protein interactions (Database of Interacting Proteins, DIP™, at <http://dip.doe-mbi.ucla.edu>) (Xenarios *et al.*, 2002) and the further enhancement of established resources such as SWISS-PROT and SWISS-2DPAGE (Appel *et al.*, 1993, 1994, 1996; Hoogland *et al.*, 1998, 1999 and 2000; Jung *et al.*, 2000; Pasquali *et al.*, 1996; Reymond *et al.*, 1997; Sanchez *et al.*, 1995-2000; Sarto *et al.*, 1997; Tonella *et al.*, 1998 and 2000; Yan *et al.*, 1997; for further references, see the following section on SWISS-2DPAGE) at the Swiss Institute of Bioinformatics at <http://www.expasy.org/sprot/> and <http://www.expasy.org/ch2d/>.

In the following sections some of the protein expression databases that are currently available in the internet are described.

7.9.2

The Proteome Analysis Database

The Proteome Analysis Database provides comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms (Apweiler *et al.*, 2001a). Currently, the database contains information for the proteins from nearly 60 complete genomes including those of the eukaryotic species *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens*.

The analyses are performed on non-redundant sets of SWISS-PROT and TrEMBL entries (Bairoch & Apweiler, 2000) representing each complete proteome, based on genome sequence submissions and additional knowledge researched by SWISS-PROT curators. Additional sets are available for mouse and human incorporating supplementary data from the EnSEMBL database (<http://www.ensembl.org/>). The sets provide an unbiased resource for analysis and similarity searches.

The statistical and comparative analyses are compiled using the databases InterPro (Apweiler *et al.*, 2001b; <http://www.ebi.ac.uk/interpro/>), CluSTr (Kriventseva *et al.*, 2001a; <http://www.ebi.ac.uk/clustr>) and Gene Ontology (GO) (Ashburner *et al.*, 2000; <http://www.geneontology.org>) and contains structural information derived from the Homology-Derived Structures of Proteins (HSSP) resource (Holm &

Sander, 1999; <http://www.embl-ebi.ac.uk/dali/>) and the Protein Data Bank (PDB) database (Berman *et al.*, 2000; <http://www.rcsb.org/pdb/>).

The Proteome Analysis Database is accompanied by a programme that has been designed to carry out InterPro-based proteome comparisons (<http://www.ebi.ac.uk/proteome/comparisons.html>) for any combination of proteomes in the database.

The database provides a variety of ways to query and compare proteome composition. The combination of both structural information and functional classification provides depth. For example, systematically conserved proteins that are likely to have orthologues across species and be involved in a common core biology can be identified. Conserved families that are missing in a given genome or proteins unique to a particular species may help to define the species.

7.9.3

SWISS-2DPAGE – A two-dimensional polyacrylamide gel electrophoresis database

SWISS-2DPAGE is a curated, added-value database (Sanchez *et al.*, 1995; Hoogland *et al.*, 2000), containing data on proteins which have mainly been identified using two-dimensional protein gel electrophoresis. Recently, the database has been extended to include data from one-dimensional techniques. The SWISS-2DPAGE database was launched in 1993 and is maintained collaboratively by the Swiss Institute of Bioinformatics and the Central Clinical Chemistry Laboratory of the Geneva University Hospital. The input data consists of annotated images from scanned representations of one- and two-dimensional protein gels. In addition to experimental data (isoelectric point, molecular weight, amino acid composition, peptide masses, and the protein identification method: gel matching, co-migration, immunoblotting, amino acids composition, microsequencing, peptide mass fingerprinting, tandem mass spectrometry), various annotations are also supplied on the level of physiological and pathological protein expression, polymorphisms, post-translational modifications, and bibliographical references. In addition, cross-references are provided to Medline, to various federated two-dimensional databases (at this time: COMPLUYEAST-2DPAGE, ECO2DBASE, HSC-2DPAGE, PHCI-2DPAGE, PMMA-2DPAGE, Siena-2DPAGE, and YEPD) and to SWISS-PROT, the protein knowledgebase (Bairoch & Apweiler, 2000), which itself provides links to many other biological databases (EMBL, GenBank, PROSITE, OMIM, *etc.*).

Information storage in SWISS-2DPAGE is based on structured flat files and images. Obviously, different kinds of lines are defined, each beginning with a two-character code, which indicates the type of data stored in the current line. For standardisation purpose, most of the SWISS-2DPAGE line types and formats have been retained from SWISS-PROT. These contain general information such as the protein name, protein description, and the species, bibliographical references, and cross-references to other databases. Only a few line types are specific to the SWISS-2DPAGE database. The master line indicates available reference maps for the protein which has been identified. The images line lists the one- and two-dimensional gel images that are associated with the protein entry. One- and

two-dimensional lines contain experimental data specific to one- and two-dimensional gels, respectively.

The current release (actual state: November 2001) contains 3,398 identified spots, corresponding to 861 different protein entries, in 33 reference maps. Seventeen are from human cells or tissues (colorectal epithelia cells, kidney, liver, lymphoma, platelet, red blood cells, soluble nuclear proteins and matrix from liver tissue), body fluids (cerebrospinal fluid, plasma), cell lines (colorectal adenocarcinoma cells, erythroleukemia cells, hepatoblastoma carcinoma-derived cells, hepatoblastoma carcinoma-derived cells and proteins secreted by these cells, macrophage-like cells, nucleolar proteins from HeLa cells, promyelocytic leukemia-derived cells). Six maps are from mouse cells or tissues (brown adipose tissue, gastrocnemius muscle, liver, pancreatic islet cells, soluble nuclear proteins and matrix from liver tissue, white adipose tissue). The other maps derive from *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli* (for various isoelectric point ranges) and *Saccharomyces cerevisiae*.

The SWISS-2DPAGE database is freely available for academics and non-profit organisations. It is submitted to an annual subscription fee for commercial entities. SWISS-2DPAGE can be downloaded by anonymous FTP (<ftp://ftp.expasy.org/databases/swiss-2dpag/>). However the most efficient and user-friendly way to browse interactively in SWISS-2DPAGE is through the ExPASy proteomics web server (<http://www.expasy.org/ch2d/>). From there, SWISS-2DPAGE can be accessed by different ways, which include various keyword searches (protein name or description, gene name, authors, and full text) and graphical queries (by selecting a spot on a chosen reference map).

7.10

Further gene-expression databases in the internet

Large-scale expression analyses create enormous quantities of data which must be managed, analysed and presented. The traditional way of presenting such data in scientific journals is not well suited for large amounts of data. For this reason, several researchers provided access to their data on the internet. The popularity, the ease of use of the internet and access to it make the internet the ideal resource for the publication of database. It is therefore not surprising that an increasing number of databases of different contents have been published during the last years.

A detailed description of all public and commercial databases available in the internet goes far beyond the scope of this book. Therefore, we listed further RNA and protein-based gene expression databases in tables 7.7 and 7.8. For every database, the internet link, references – if available – and a short description of its content is given. Because of the dynamic nature of the internet the list of databases described in this book cannot be regarded as definitive.

Table 7.7 Further RNA-based databases available in the internet.

Database name	Abbreviations	Internet link, references	Additional information
A <i>Caenorhabditis elegans</i> Database	ACeDB	http://www.acedb.org/ Hope <i>et al.</i> (1997), Kelly (2000), Martinelli <i>et al.</i> (1997), Stein & Thierry-Mieg (1998)	ACeDB was originally developed for the <i>Caenorhabditis elegans</i> genome project since 1989. It is used both for managing data within genome projects, and for making data available to the wider scientific community.
Array Database at the National Human Genome Research Institute	ArrayDB	http://genome.nhgri.nih.gov/arraydb/ Ermolaeva <i>et al.</i> (1998)	ArrayDB provides an interface for the mining and analysis of microarray gene expression data. Microarray slide images and information about the clones are available. Each clone is linked to external databases, such as UniGene, Entrez, dbEST, KEGG, and GeneCards. It is also possible to query a range of related experiments in order to see changes in the expression pattern of a gene over different experiments.
A <i>Xenopus laevis</i> database	Axeldb	http://www.dkfz-heidelberg.de/abt0135/axeldb.htm Gawanitka <i>et al.</i> (1998), Pollet <i>et al.</i> (2000)	Axeldb is a database focussing on gene expression in the frog <i>Xenopus laevis</i> . It contains data of a large-scale <i>in situ</i> hybridisation screening in <i>Xenopus laevis</i> embryos. The goals of this project are to identify genes by the characterisation of their expression pattern, to partially sequence the corresponding cDNAs and to maintain a database collecting the results.
cDNA xProfiler	–	http://cgap.nci.nih.gov/Tissues/xProfiler	The cDNA xProfiler allows comparison of the expression of genes between two pools of cDNA libraries, where each library can be either a single library or several libraries. For a gene to be 'present' in a library, there must be at least one EST sequence in the UniGene cluster. The tool lists all genes found in both pools and categorises them as 'unique' and 'non-unique' and as 'known' and 'unknown'. Finally, the distribution of these genes between both samples is shown.

Table 7.7 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Digital Gene Expression Displayer	DGED	http://cgap.nci.nih.gov/Tissues/GXS	DGED analyses the differences in gene expression between two pools of libraries. Unlike the cDNA xProfiler, which lists every gene in both groups, DGED finds only the statistically significant differences, based on the sequence odds ratio and a χ -squared test.
Expressed Gene Anatomy Database	EGAD	http://www.tigr.org/tdb/egad/egad.shtml	EGAD was constructed by extraction and curation of sequences from GenBank to create a non-redundant set of human transcript sequences and non-human transcript sequences. Each human transcript is linked to a human gene. Cellular roles have been assigned and tissue distribution information is also available. Extensive expression information including positions of overlapping expressed sequence tags is available.
Endocrine Pancreas Consortium Database	EPConDB	http://www.cbil.upenn.edu/EPConDB/	For information about EPConDB, see "RAD – The RNA Abundance Database"
Gene Expression Database for the Molecular Pharmacology of Cancer	–	http://discover.nci.nih.gov/nature2000/ Additional data at: http://dtp.nci.nih.gov/ Myers <i>et al.</i> (1997), Ross <i>et al.</i> (2000), Scherf <i>et al.</i> (2000)	The database arose out from the NCI60 microarray expression project. cDNA microarrays were used to assess gene expression profiles in 60 human cancer cell lines derived from the National Cancer Institute. Using the microarray data, bioinformatics were linked with chemoinformatics by correlating gene expression and drug activity patterns in these cells.
Genzyme Molecular Oncology SAGE Database	–	http://www.sagenet.org/	SAGE data from over 600,000 transcripts are available for downloading as SAGE projects and Microsoft Access™ databases. These files include SAGE data from human transcripts from colorectal tissue, pancreatic cancer cells and tissue, mouse fibroblast transcripts and several yeast transcripts.

Table 7.7 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
GenBank	–	http://www.ncbi.nlm.nih.gov/ Benson <i>et al.</i> (1999, 2000, 2002), Wheeler <i>et al.</i> (2000, 2001, 2002)	GenBank is a database of the National Institutes of Health (NIH). The annotated collection of all public DNA sequences contains more than 14,397,000,000 bases in 13,602,000 records (October 2001). It is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at the National Center for Biotechnology Information (NCBI). GLS is a gene expression tool for a single cDNA library or a group of libraries. First, it finds all the expressed genes that meet a chosen criteria of organism and library characteristics. Then it summarises or organises these genes based on a selected criteria. Finally, it classifies them into unique and non-unique, and known and unknown genes.
Gene Library Summarizer	GLS	http://cgap.nci.nih.gov/Tissues/LibrarySummarizer	
Mammalian Gene Collection	MGC	http://mgc.nci.nih.gov/index.html Strausberg <i>et al.</i> (1999)	The MGC seeks to identify full open reading frames for each human and mouse gene. Over 80 cDNA libraries derived from human tissue and cell lines, and mouse tissue were produced. More than 500,000 expressed sequence tags have been read, and more than 20,000 non-redundant human and mouse putative full open reading frames have been assembled.
Nematode Expression Pattern Database	NextDB	http://nematode.lab.nig.ac.jp/	NEXTDB integrates data from an expression pattern project of the <i>Caenorhabditis elegans</i> genome that has been constructed through EST analysis and whole-mount <i>in situ</i> hybridisation. This database contains visual expressions of relationships among cosmids, predicted genes and the cDNA clones, <i>in situ</i> hybridisation images and tag sequences.

Table 7.7 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Prostate Expression Database	PEDB	http://www.pedb.org/ Hawkins <i>et al.</i> (1999), Nelson <i>et al.</i> (2000, 2002), Pritchard <i>et al.</i> (2001)	The public PEDB currently comprises over 85,000 expressed sequence tags from 38 prostate tissue and cell type-specific cDNA libraries. The PEDB started as an effort to construct a molecular fingerprint of the normal, preneoplastic, and neoplastic human prostate and allows virtual expression analysis. Data is linked to other tools, like dbEST, GenBank, Entrez, CGAP, UniGene, <i>etc.</i>
<i>Plasmodium</i> Genome Resource Database	PlasmoDB	http://plasmodb.org Bahl <i>et al.</i> (2002), Ben Mamoun <i>et al.</i> (2001), Munasinghe <i>et al.</i> (2001), Patankar <i>et al.</i> (2001), Watanabe <i>et al.</i> (2001)	For information about PlasmoDB, see "RAD – The RNA Abundance Database"
Database of the Translational Recoding Events	RECODE	http://recode.s.utah.edu/ Baranov <i>et al.</i> (2001)	RECODE is a database of the utilisation of ribosomal frameshifting, translational bypassing and mRNA-specific codon redefinition for gene expression.
SAGEmap Virtual Northern	–	http://www.ncbi.nlm.nih.gov/SAGE/sagevm.cgi Lal <i>et al.</i> (1999), Lash <i>et al.</i> (2000), Riggins & Strausberg (2001), Wheeler <i>et al.</i> (2000, 2001, 2002)	This tool extracts up to four possible SAGE tags and orientation signals from a given sequence. Orientation signals allow the most likely tag to be chosen. Following the tag links in the output table the expression level of the tag in different SAGE libraries is shown and information how it is represented in the rest of the sequences in GenBank and UniGene is given.
SAGEmap xProfiler		http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi For references, see SAGEmap Virtual Northern	This tool for pooling and comparison of SAGE library data has been designed for differential-type analyses. Similar libraries can be placed into one of two groups based on their characteristics (<i>e.g.</i> , normal colon, and colon cancer). Comparisons are then made between the two groups using a statistical test developed specifically for SAGE data.

Table 7.7 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Tooth Development Database	–	http://bite-it.helsinki.fi/ Nieminen <i>et al.</i> (1998)	The database provides information on gene expression in the developing tooth. The aim of the database is to provide a tool for learning and comparing the expression patterns of different genes and for detecting coexpression. The data in this database was generated by <i>in situ</i> hybridisation, whole-mount <i>in situ</i> hybridisation, immunohistochemistry and other methods.
Yeast genome-wide expression database	–	http://web.wi.mit.edu/young/expression/ Causton <i>et al.</i> (2001), Holstege <i>et al.</i> (1998), Wyrick & Young (2002)	The database contains detailed information on yeast mRNA populations (the transcriptome, apparent half-life and transcriptional frequency of mRNAs) and genome-wide expression data (dependence of gene expression on components of the transcriptional machinery and dependence of expression of a gene on multiple components of the machinery).

Table 7.8 Further protein-based databases available in the internet. For a reference to standardised two-dimensional PAGE databases; see Appel *et al.* (1996a). For an overview of protein databases in the internet, see the WORLD-2DPAGE (<http://www.expsys.org/ch2d/2d-index.html>) from which most of the data in this table is derived. Please note that some of the databases are no longer maintained, but may nevertheless contain useful data.

Database name	Abbreviations	Internet link, references	Additional information
<i>Arabidopsis</i> mitochondrial proteome project database	AMPP	http://www.gartenbau.uni-hannover.de/genetik/A_MPP Kruft <i>et al.</i> (2001), Werhahn & Braun (2002)	The <i>Arabidopsis</i> mitochondrial proteome project at the University of Hanover (Germany) provides data about proteomes of mitochondria isolated from cells in suspension characterised by high resolution two-dimensional gel electrophoresis and mass spectrometry. Currently, the proteomes of selected <i>Arabidopsis</i> knock-out mutants are under investigation with particular reference to nuclear genes of mitochondrial proteins.
Argonne protein mapping group server	–	http://www.anl.gov/BIO/PMG/projects/ Giometti <i>et al.</i> (1992, 1995, 1997), Williams <i>et al.</i> (1998)	The Protein Mapping Group at the Argonne National Laboratory (Argonne, Illinois, USA) studies normal protein expression and changes in protein expression induced by pathological events or exposure to environmental stimuli. The database currently includes maps of proteins from mouse liver, human breast cells, and the hyperthermophilic organism, <i>Pyrococcus furiosus</i> .
Australian National University Two-Dimensional Polyacrylamide Gel Electrophoresis Database	ANU-2DPAGE	http://semele.anu.edu.au/2d/2d.html Chen <i>et al.</i> (2000), Mathesius <i>et al.</i> (2001), Weiller <i>et al.</i> (2001)	ANU-2DPAGE (Australian National University, Canberra, Australia) provides information on the proteomes of <i>Oryza sativa</i> (anthem) and <i>Medicago truncatula</i> (root). Proteomes of <i>Sinorhizobium meliloti</i> , <i>Medicago truncatula</i> leaf and <i>Medicago truncatula</i> seed are under construction.
Clusters of SWISS-PROT and TrEMBL protein database		http://www.ebi.ac.uk/clustr/ Apweiler <i>et al.</i> (2001), Kriventseva <i>et al.</i> (2001)	CluSTr offers automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. Clustering is based on analysis of all pairwise comparisons between protein sequences. The database provides links to databases, such as HSSP, PDB and InterPro. Clusters for five complete eukaryotic proteomes (<i>Arabidopsis thaliana</i> , <i>Caenorhabditis elegans</i> , <i>Drosophila melanogaster</i> , <i>Saccharomyces cerevisiae</i> and <i>Homo sapiens</i>) and 43 prokaryotic proteomes are available.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Embryonic stem cell database	–	http://www.dur.ac.uk/~db10nh1/2DPAGE/	This database at the University of Edinburgh (United Kingdom) provides information on protein abundance in embryonic stem cells.
<i>Haemophilus</i> 2D Protein Database	–	http://www.abdn.ac.uk/~mmb023/2dhome.htm Cash <i>et al.</i> (1995, 1997)	The <i>Haemophilus</i> 2D protein database at the University of Aberdeen (Scotland, United Kingdom) is being developed to define the proteome of <i>Haemophilus influenzae</i> and other members of the genus.
Heart High-Performance 2-DE Database	HP-2DPAGE	http://www.mdc-berlin.de/~emu/heart/ Jungblut <i>et al.</i> (1992, 1994), Müller <i>et al.</i> (1996), Otto <i>et al.</i> (1994, 1996)	The Heart High-Performance 2-DE Database at the Max Delbrück Center (Berlin, Germany) contains data on 3,300 proteins of human heart ventricle. The identification of proteins was improved by the combination of amino-terminal and internal sequencing, amino acid analysis, immunostaining, MALDI-MS peptide mass fingerprinting, post source decay MALDI-MS, and ladder sequencing by carboxypeptidase P.
Heart Science Centre 2D-PAGE database	HSC-2DPAGE	http://www.harefield.nthames.nhs.uk/nhli/protein/ Corbett <i>et al.</i> (1994), Dunn <i>et al.</i> (1997), Evans <i>et al.</i> (1997) and Pleissner <i>et al.</i> (1996, 1997)	The HSC-2D PAGE protein database was established and is maintained by the Protein Biochemistry Group within the Heart Science Centre at Harefield in the United Kingdom. The database contains information on myocytes and endothelial cells from the ventricle of the human heart, as well as data derived from rat and dog heart (ventricle) tissue.
Human Colon Carcinoma Protein Database	–	http://www.ludwig.edu.au/jpsl/jpslhome.html Chataway <i>et al.</i> (1998), Ji <i>et al.</i> (1994, 1997), Rasmussen <i>et al.</i> (1998), Simpson (1991), Simpson <i>et al.</i> (1992b, 2000), Ward <i>et al.</i> (1990), Ward & Simpson (1991)	The Human Colon Carcinoma Protein Database at the Ludwig Institute for Cancer Research (Melbourne, Australia) includes proteome data of human breast carcinoma cell line MDA-MB231, human colon carcinoma cell line LIM1215, human placenta and mouse colon crypts.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Human and mouse 2-D PAGE databases	–	http://proteomics.cancer.dk/ Celis <i>et al.</i> (1996, 1998)	The human and mouse 2-D PAGE databases of the Danish Centre for Human Genome Research at the University of Aarhus in Denmark provides data derived from several human (<i>e.g.</i> , keratinocytes, transitional cell carcinomas, bladder squamous cell carcinomas, urine, fibroblasts) and mouse (<i>e.g.</i> , kidney) samples.
Human Myocardial Two-Dimensional Electrophoresis Protein Database	HEART-2DPAGE	http://userpage.chemie.fu-berlin.de/~pleiss/dhzb.html Baker <i>et al.</i> (1992), Jungblut <i>et al.</i> (1994), Müller <i>et al.</i> (1996), Pleissner <i>et al.</i> (1996-1997)	The HEART-2DPAGE database at the German Heart Institute in the Virchow Clinic of the Humboldt University Berlin (Germany) is a two-dimensional polyacrylamide gel electrophoresis database for human heart proteins. It contains data on proteins identified on 2D-PAGE maps of ventricle and atrium of human heart.
INRA Maize Proteome Database	–	http://aestivum.moulon.inra.fr/imgd/	The INRA Maize Proteome Database at the Institut National de la Recherche Agronomique (INRA; Gif-sur-Yvette, France) provides data of the maize proteome and genome.
Mitochondrial proteome database	MitoPick	http://www-dsv.cea.fr/thema/MitoPick/Mito2D.html	MitoPick (Life Sciences Division, Commissariat à l'Énergie Atomique, CEA; Paris, France) offers information on mitochondrial proteins.
Parasite Host Cell Interaction 2D-PAGE database	PHCI-2DPAGE	http://www.gram.au.dk/ Shaw <i>et al.</i> (1999)	The PHCI-2DPAGE database at the University of Aarhus (Denmark) is a two-dimensional polyacrylamide gel electrophoresis database with information on the <i>Chlamydia pneumoniae</i> VR1310 proteome and interferon-γ-induced HeLa liver cell proteins.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Parasite proteome maps	–	http://www.aber.ac.uk/~mpgwww/Proteome/ProtMaps.html Barrett <i>et al.</i> (2000), Jefferies <i>et al.</i> (2000)	The parasite proteome maps at the Institute of Biological Sciences (University of Wales, Aberystwyth, United Kingdom) provide proteome data derived from <i>Caenorhabditis elegans</i> , <i>Fasciola hepatica</i> , <i>Hymenolepis diminuta</i> , <i>Moniezia expansa</i> and <i>Necator americanus</i> .
PhosphoProtein Database	PPDB	http://www-lecb.ncifcrf.gov/phosphoDB/	The purpose of the preliminary PhosphoProtein Database at the Laboratory of Experimental and Computational Biology of the National Cancer Institute (Bethesda, Maryland, USA) is to promote a systematic approach to the study of cellular protein phosphorylation.
Plant Plasma Membrane Database	PPMdb	http://sphinx.rug.ac.be:8080/ppmdb/index.html Sahnoun <i>et al.</i> (2000)	The PPMdb database at Gent University (Belgium) offers data on plasma membrane proteins expressed in the plant <i>Arabidopsis thaliana</i> . This database is not being updated and will be replaced by the <i>Arabidopsis</i> Functional Protein Database (AFPdb).
Protein database of haematopoietic cell lines	BPP-2DPAGE	http://www.smbh.univ-paris13.fr/lbtp/Biochemistry/biochimie/bque.htm Müller <i>et al.</i> (1999), Poirier <i>et al.</i> (2001)	The BPP-2DPAGE database at the Laboratoire de Biochimie et Technologie des Protéines (Bobigny, France) contains data on proteins identified on various 2-D PAGE reference maps of haematopoietic cell lines (DG75, HL60, K562, PRI and TF1).
Protein database of the Statens Serum Institute	SSI-2DPAGE	http://www.ssi.dk/en/forskning/tbimmun/Protein_database/protein_database.htm Rosenkrands <i>et al.</i> (2000), Weldingh <i>et al.</i> (2000)	The Protein database of the Statens Serum Institute in Copenhagen (Denmark) offers data on major proteins from the cell wall, short term culture filtrate and cytosol of the <i>Mycobacterium tuberculosis</i> strain H37RV.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Protein Disease Database	PDD	http://www-lecb.ncifcrf.gov/PDD/	The PDD at the Laboratory of Experimental and Computational Biology of the National Cancer Institute (Bethesda, Maryland, USA) is a relational database designed to answer questions regarding protein patterns found in common body fluids with respect to several disease conditions. Data is derived from human plasma, cerebrospinal fluid and urine. The PDD is no longer being maintained or updated.
Protein expression database for the molecular pharmacology of cancer	–	http://discover.nci.nih.gov/host/prot.html Myers <i>et al.</i> (1997), Ross <i>et al.</i> (2000), Scherf <i>et al.</i> (2000)	The Developmental Therapeutics Program of the National Cancer Institute (Bethesda, Maryland, USA) has screened over 60,000 chemical compounds and a larger number of natural product extracts for their ability to inhibit growth of 60 cancer cell lines of leukemias and lymphomas, as well as cancers of breast, prostate, ovary, lung, colon, kidney, and central nervous system origin. As part of that effort, a two-dimensional gel electrophoresis-based protein expression database covering all cell types was developed.
Proteome 2D-PAGE database	–	http://www.mpiib-berlin.mpg.de/2D-PAGE/ Bumann <i>et al.</i> (2001), Jungblut <i>et al.</i> (1999, 2000), Mollenkopf <i>et al.</i> (1999), Thiede <i>et al.</i> (2000)	The Proteome 2D-PAGE database at the Max Planck Institute for Infection Biology (Berlin, Germany) contains data derived, amongst others, from the European Bacteria Project (EBP). Samples have been isolated from <i>Chlamydia pneumoniae</i> , <i>Helicobacter pylori</i> , <i>Mycobacterium tuberculosis</i> , <i>Borrelia garinii</i> , <i>Francisella tularensis</i> LVS, <i>Mycoplasma pneumoniae</i> M129, human Jurkat T-cells and mouse mammary gland.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Proteome®'s BioKnowledge® Library		http://www.proteome.com/databases/index.html	The BioKnowledge® Library at Incyte Genomics, Inc. (Palo Alto, California, USA) consists of multi-species volumes that allow searching between volumes using common protein and gene characteristics. Model organism databases are freely available over the internet for academic users, while products of constantly increasing value are available for corporate subscribers. The library, available to commercial subscribers, consists of individual proteome databases for each model organism, and includes analysis software. In addition to data derived from <i>Homo sapiens</i> , <i>Saccharomyces cerevisiae</i> , <i>Schizosaccharomyces pombe</i> <i>Caenorhabditis elegans</i> , MycoPathDB™ offers data derived from <i>Aspergillus flavus</i> , <i>Aspergillus fumigatus</i> , <i>Aspergillus niger</i> , <i>Blastomyces dermatitidis</i> , <i>Candida albicans</i> , <i>Candida dubliniensis</i> , <i>Candida glabrata</i> , <i>Candida guilliermondii</i> , <i>Candida krusei</i> , <i>Candida lusitanae</i> , <i>Candida parasilosis</i> , <i>Candida pseudotropicalis</i> , <i>Candida tropicalis</i> , <i>Coccidioides immitis</i> , <i>Cryptococcus neoformans</i> , <i>Histoplasma capsulatum</i> and <i>Pneumocystis carinii</i> .
Human proteome survey	HumanPSD™	Costanzo <i>et al.</i> (2001), Hodges <i>et al.</i> (1999, 2002)	
Yeast proteome	YPD™		
<i>Schizosaccharomyces pombe</i> proteome	PombeDB™		
<i>Caenorhabditis elegans</i> proteome	WormPD™		
Fungal pathogens' proteomes	MycoPathDB™		
Proteome Database for Aging Research	TMIG-2DPAGE	http://proteome.tmig.or.jp/2D/ Toda <i>et al.</i> (1998, 2000), Toda (2001)	The TMIG-2DPAGE database at the Tokyo Metropolitan Institute of Gerontology (Tokyo, Japan) presents data of cell line proteomes (normal human diploid fibroblasts TIG-3, fibroblasts derived from patients with Werner's syndrome, C57BL/6 mouse brain tissues, Epstein-Barr virus-transformed human B-lymphoblastoids and human neuroblastoma SH-SY5Y).

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Purkyne Military Medical Academy 2D-PAGE Database	PMMA-2DPAGE	http://www.pmma.pmfhk.cz/ Greibenova et al. (2000), Hernychova et al. (2000), Jungblut et al. (1999), Kovarova et al. (1994), Stulik et al. (1997-2001)	The PMMA-2DPAGE database at the Purkyne Military Medical Academy in Hradec Králové (Czech Republic) is a two-dimensional polyacrylamide gel electrophoresis database with data derived from human colorectal carcinoma and <i>Francisella tularensis</i> .
Rat Serum Protein Study Group database	–	http://linux.farma.unimi.it/ Eberini et al. (1999, 2000), Haynes et al. (1998), Miller et al. (1998, 1999), Wait et al. (2001)	This database at the Department of Pharmacological Science of the University of Milan (Italy) contains data derived from rat serum (control Sprague-Dawley female and male rats, turpentine-treated rats, diabetic rats, stroke-prone rats), urine (control female and male rats, stroke-prone rats), and cerebrospinal fluid (control and stroke-prone rats).
Rat Serum Protein Database	–	http://depts.washington.edu/~ruedilab/info/ratserum/ratserum.html Eberini et al. (1999), Haynes et al. (1998), Miller et al. (1998, 1999), Sironi et al. (2001), Wait et al. (2001)	The Rat Serum Protein Database at the Abersold Protein Laboratory of the University of Washington (Seattle, Washington, USA) offers data from two-dimensional gel electrophoresis separations of rat serum (normal and acute phase).
SIENA-2DPAGE	–	http://www.bio-mol.unisi.it/2d/2d.html Bini et al. (1996, 1997), Birkelund et al. (1997) Liberatori et al. (1997)	The SIENA-2DPAGE database is a two-dimensional polyacrylamide gel electrophoresis database at the Department of Molecular Biology of the University of Siena (Italy). The data was derived from human breast ductal carcinoma, histological normal tissue and amniotic fluid, as well as from <i>Chlamydia trachomatis</i> and <i>Caenorhabditis elegans</i> .
<i>Streptomyces coelicolor</i> Reference Gel	–	http://gepasi.dbs.aber.ac.uk/S_Coeli/ReferenceGel/refgel.html	This internet site at the University of Wales (Aberystwyth, United Kingdom) offers the <i>Streptomyces coelicolor</i> reference gel with corresponding protein data.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
SWISS-2DPAGE	–	http://www.expasy.org/ch2d/ Appel <i>et al.</i> (1993-1996), Hoogland <i>et al.</i> (1998-2000), Jung <i>et al.</i> (2000), Pasquali <i>et al.</i> (1996), Reymond <i>et al.</i> (1997), Sanchez <i>et al.</i> (1995-2000), Sarto <i>et al.</i> (1997), Tonella <i>et al.</i> (1998, 2000), Yan <i>et al.</i> (1997)	The SWISS-2DPAGE database at the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland) is an annotated two-dimensional polyacrylamide gel electrophoresis (2DPAGE) and SDS-PAGE database. The actual release 15.0 contains 861 protein entries identified in 33 reference maps from human, mouse, <i>Arabidopsis thaliana</i> , <i>Dichytelium discoideum</i> , <i>Escherichia coli</i> and <i>Saccharomyces cerevisiae</i> samples.
Swiss-Czech Proteomics Server	SWICZ	http://proteom.biomed.cas.cz/ Grünenfelder <i>et al.</i> (2001), Vohradsky <i>et al.</i> (1997)	The SWICZ database is collaborative project between the Institute of Microbiology (Academy of Sciences in Prague, Czech Republic) and the Division of Molecular Microbiology (Biozentrum, University of Basel, Switzerland). It offers developmental two-dimensional PAGE databases for <i>Caulobacter crescentus</i> and <i>Streptomyces coelicolor</i> .
SWISS-PROT	–	http://www.ebi.ac.uk/swissprot/ http://www.expasy.org/sprot/ Appel <i>et al.</i> (1993-1996), Bairoch & Apweiler (1996-2000), Gasteiger <i>et al.</i> (2001), Hoogland <i>et al.</i> (1998-2000), Jung <i>et al.</i> (2000), O'Donovan <i>et al.</i> (2002), Pasquali <i>et al.</i> (1996), Reymond <i>et al.</i> (1997), Sanchez <i>et al.</i> (1995-2000), Sarto <i>et al.</i> (1997), Tonella <i>et al.</i> (1998, 2000), Yan <i>et al.</i> (1997)	The SWISS-PROT database maintained by the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland) and the European Bioinformatics Institute (EBI; Hinxton, United Kingdom) consists of protein sequence entries. It contains high-quality annotation, is non-redundant and cross-referenced to many other databases. The actual release 40.0 contains 101,602 sequence entries, comprising 37,315,215 amino acids abstracted from 91,880 references.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
ToothPrint	–	http://bioc111.otago.ac.nz:591/tooth/home.htm Hubbard <i>et al.</i> (2001)	ToothPrint is a two-dimensional gel database of dental tissue proteins at the University of Otago (Dunedin, New Zealand). It provides protein expression data in enamel cells and matrix derived from rat.
<i>Toxoplasma gondii</i> proteome map	–	http://www-public.rz.uni-duesseldorf.de/~hfischer/	This internet site at the Institute for Medical Microbiology and Virology of the Heinrich-Heine-University Düsseldorf (Germany) offers a proteome map of the tachyzoite <i>Toxoplasma gondii</i> .
TrEMBL	–	http://www.ebi.ac.uk/trEMBL/ Bairoch & Apweiler (1996-2000), Junker <i>et al.</i> (2000), Lang (1997), Möller <i>et al.</i> (1999), O'Donovan <i>et al.</i> (1999)	The TrEMBL database contains translations of all coding sequences present in the Nucleotide Sequence Database of the European Molecular Biology Laboratory (EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom), which are not yet integrated into SWISS-PROT. The goal of the current TrEMBL release is to achieve synchronisation with the SWISS-PROT protein knowledgebase.
Two-dimensional polyacryl-amide gel electrophoresis database for broncho-alveolar lavage fluid proteins	BALF 2D	http://www.umh.ac.be/~biochim/BALF2D.html Wattiez <i>et al.</i> (1999, 2000)	The BALF 2D database at the Department of Biological Chemistry of the University of Mons-Hainaut (Belgium) is a database with information of proteins present in broncho-alveolar lavage fluids derived from humans and mice.
Washington University Inner Ear Protein Database	–	http://oto.wustl.edu/thc/innerear2d.htm Thalmann <i>et al.</i> (1992, 1994, 1997)	The Washington University Inner Ear Protein Database is a result of long-term ongoing studies at the Department of Otolaryngology at Washington University (St. Louis, Missouri, USA) in collaboration with numerous investigators from around the world. The database contains information on protein abundance in human perilymph and endolymph, organ of corti, otoconia, tectorial membrane derived from Guinea pig.

Table 7.8 (continuing)

Database name	Abbreviations	Internet link, references	Additional information
Yeast 2D-PAGE Protein Database	YEAST 2D-PAGE	http://yeast-2dpage.gmm.gu.se/	The YEAST 2D-PAGE database at the Lundberg Laboratory of the University of Gothenburg (Sweden) contains information about two-dimensional PAGE-resolved proteins from a number of species of yeast (<i>Saccharomyces cerevisiae</i> , <i>Debaryomyces hansenii</i> , <i>Schizosaccharomyces pombe</i> , <i>Candida albicans</i> , and <i>Rhodotorula glutinis</i>); however, the main focus will be on <i>Saccharomyces cerevisiae</i> .
Yeast protein database COM-PLUYEAST-2DPAGE	COM-PLUYEAST-2DPAGE database	http://babbage.csc.ucm.es/2d/ Pardo <i>et al.</i> (1999, 2000), Pitarch <i>et al.</i> (1999, 2001)	The COMPLUYEAST-2DPAGE database at the Universidad Complutense de Madrid (Spain) is a two-dimensional polyacrylamide gel electrophoresis database and provides data of the yeast proteome.
Yeast Protein Map	YPM	http://www.ibgc.u-bordeaux2.fr/YPM/ Boucherie <i>et al.</i> (1996), Perrot <i>et al.</i> (1999)	The YPM database at the Institute of Biochemistry and Cellular Genetics (Bordeaux, France) contains data on soluble yeast proteins synthesised by strain S288C exponentially growing on glucose and synthesised by the same strain in response to heat shock.
<i>Xylella</i> 2D Database	–	http://proteome.ibi.unicamp.br/databases/index.htm	The <i>Xylella</i> 2D Database at the Laboratory of Protein Chemistry (Laquip Unicamp, São Paulo, Brazil) provides information on the proteome of the bacteria <i>Xylella fastidiosa</i> .

7.11

Summary

As this chapter documents, bioinformatics is a wide term and describes a field that is evolving very rapidly. Thus, any description must be seen as a snapshot in time and not as a definitive statement. It is however clear that the approaches outlined in this chapter will have a fundamental and wide-ranging effect on research, diagnostics and therapeutics. Expression analyses and multi-parametric testing will become part of routine laboratory diagnostics, and the study of the interaction between gene expression, genetic polymorphisms and xenobiotic compounds – the field which has been termed pharmacogenomics – holds great promise for the design of individualised therapies that are more effective and cause fewer side effects.

In research it is often important to simplify in order to understand. This is nowhere more true than in the field of bioinformatics where the problem is not to see the wood for the trees but rather to understand the global pattern of forestation. We hope that this chapter makes some small contribution to this task.

7.12

References**Internet resources**

AceDB (*Caenorhabditis elegans* Database) at the Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.acedb.org/>.

Arabidopsis mitochondrial proteome project database (AMMP; University of Hanover, Hanover, Germany); available at <http://www.gartenbau.uni-hannover.de/genetik/AMPP>.

Argonne protein mapping group server (Argonne National Laboratory; Argonne, Illinois, USA); accessible at <http://www.anl.gov/BIO/PMG/projects/>.

ArrayDB (Array Database) at the National Human Genome Research Institute (NHGRI; National Institutes Of Health NIH; Bethesda, Maryland); available at <http://genome.nih.gov/arraydb/>.

ArrayExpress at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus; Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/arrayexpress/>.

Artificially Selected Proteins/Peptides Database (ASPD; Institute of Cytology and Genetics, Novosibirsk, Russia); available at <http://www.sgi.sgcc.ru/mgs/gnw/aspd/>.

ASDB (Alternative Splicing Database) hosted by the Lawrence Berkeley National Laboratory (operated by the University of California for the United States Department of Energy, USA); available at <http://cbcg.lbl.gov/asdb>.

AsMamDB (Alternative Splice Database of Mammals) at the Tsinghua University (Peking, China); available at <http://166.111.30.65/ASMAMDB.html>.

Australian National University Two-Dimensional Polyacrylamide Gel Electrophoresis Database (ANU-2DPAGE; Australian National University, Canberra, Australia); accessible at <http://semele.anu.edu.au/2d/2d.html>.

Axeldb (*Xenopus laevis*) Database at the Deutsches Krebsforschungszentrum, (DKFZ; Heidelberg, Germany); available at <http://www.dkfz-heidelberg.de/abt0135/axeldb.htm>.

Bioinformatics.Org (USA); available at <http://bioinformatics.org>.

Bioinformatics Template Library (BTL); available at <http://www.cryst.bbk.ac.uk/classlib/bioinf/BTL.html>.

BLAST resources of the National Center of Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/BLAST/>.

- BodyMap (The Human and Mouse Gene Expression Database) at the Institute for Molecular and Cellular Biology (Osaka University, Osaka, Japan); available at <http://body-map.ims.u-tokyo.ac.jp/>.
- Catch (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.biobase.de>.
- cDNA xProfiler at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/Tissues/xProfiler>.
- CGAP (Cancer Genome Anatomy Project) at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/>.
- CGAP tools of the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/Tools>.
- Cister (Cis-element Cluster Finder; Boston University; Boston, Massachusetts, USA); available at <http://sullivan.bu.edu/~mfrith/cister.shtml>.
- Cluster of Orthologous Groups (COG) database of the National Center for Biotechnology Information (NCBI; National Library of Medicine, Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/COG/>.
- Cluster software from Michael Eisen; available at <http://rana.lbl.gov/EisenSoftware.htm>.
- CluStr (Clusters of SWISS-PROT and TrEMBL protein database) at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/clustr/>.
- COMPEL (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.gene-regulation.com/pub/databases.html#transcompel>.
- ConsInspector (GSF – Forschungszentrum für Umwelt und Gesundheit GmbH, Neuherberg, Germany); available at <http://www.gsf.de/biodv/consinspector.html>.
- Consed on the Phred/Phrap/Consed System homepage; available at <http://www.phrap.org/>.
- CorePromoter Finder (Computational Biology and Bioinformatics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA); available at <http://argon.cshl.org/genefinder/CPROMOTER/>.
- Cross-match on the Phred/Phrap/Consed System homepage; available at <http://www.phrap.org/>.
- CYTOMER Gene Expression Database (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.biobase.de:8080/index.html>.
- Danish Centre for Human Genome Research's 2-D PAGE Databases (University of Aarhus, Aarhus, Denmark); available at <http://proteomics.cancer.dk/>.
- Database of Interacting Proteins (DIP™; University of California, Berkeley, California, USA); available at <http://dip.doe-mbi.ucla.edu>.
- dbEST (Database of Expressed Sequence Tags) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/dbEST/>.
- DDD (Digital Differential Display) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?ORG=Hs>.
- DGED (Digital Gene Expression Displayer) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/Tissues/GXS>.
- Dragon Promoter Finder (DPF; Laboratories for Information Technology, Kent Ridge Digital Labs, Singapore); available at <http://sdmc.krdl.org.sg:8080/promoter/>.
- EBI (European Bioinformatics Institute; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom) internet server; available at <http://www.ebi.ac.uk> and <http://srs.ebi.ac.uk:80/>.
- Edinburgh Mouse Atlas (Jackson Laboratory, Bar Harbor, Maine, USA, in cooperation with the Edinburgh University, Edinburgh, United Kingdom); available at <http://genex.hgu.mrc.ac.uk/Databases/Anatomy/>.
- EGAD (Expressed Gene Anatomy Database) at The Institute of Genomic Research (TIGR; Rockville, Maryland, USA); available at <http://www.tigr.org/tdb/egad/egad.shtml>.
- EMAGE (Edinburgh Mouse Atlas Gene Expression Database) at the University of Edinburgh (Edinburgh, United Kingdom); available at <http://genex.hgu.mrc.ac.uk/>.
- EMBL (European Molecular Biology Laboratory, European Bioinformatics Institute, EBI; Heidelberg, Germany); available at <http://www.embl.org/>.

- Embryonic stem cell database (University of Edinburgh, Edinburgh, United Kingdom); available at <http://www.dur.ac.uk/~dbi0nh1/2DPAGE/>.
- EnsEMBL is a joint project between the EMBL (European Molecular Biology Laboratory, Heidelberg, Germany) and the Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom) to develop a software system which produces and maintains automatic annotation on eukaryotic genomes; available at <http://www.ensembl.org/>.
- Entrez Genomes website of the National Center for Biotechnology Information (Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html>.
- EPConDB (Endocrine Pancreas Consortium Database) at the University of Pennsylvania (Philadelphia, Pennsylvania, USA); available at <http://www.cbil.upenn.edu/EPConDB/>.
- EPD (Eukaryotic Promoter Database) of the Swiss Institute of Bioinformatics and Swiss Institute for Experimental Cancer Research (Epalinges s/Lausanne, Switzerland); available at <http://www.epd.isb-sib.ch>.
- ESTscan at the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland); available at <http://www.ch.embnet.org/software/ESTScan.html>.
- ExPASy proteomics web server of the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland); available at <http://www.expasy.org/ch2d/>.
- ExPASy server links (Swiss Institute of Bioinformatics, SIB; Geneva, Switzerland) at Amos' WWW links page at <http://www.expasy.org/alinks.html>.
- FASTA resources of the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/FASTA/>.
- FastM (Genomatix Software GmbH, Munich, Germany); available at <http://www.genomatix.de/cgi-bin/fastm2/fastm.pl>.
- FirstEF (Computational Biology and Bioinformatics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA); available at <http://www.cshl.org/mzhanglab/>.
- Framefinder tool; available at <http://fe.hgmp.mrc.ac.uk/gslater/estate.tar.gz>.
- FunSiteP (Institute Cytology and Genetics; Union of Soviet Socialist Republics Academy of Science, Novosibirsk, Russia); available at <http://compel.bionet.nsc.ru/FunSite/fsp.html>.
- Gapped BLAST at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <ftp://ftp.ncbi.nih.gov/blast/>.
- GDB (Genome Database, USA); available at <http://www.gdb.org/>.
- GenBank at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/>.
- Gencarta Database (Compugen, Tel Aviv, Israel); available at <http://www.gencarta.com/>.
- GeneData AG (Basel, Switzerland); available at <http://www.genedata.com>.
- GeneEMAC at the University of Vienna (Vienna, Austria); available at <http://www.univie.ac.at/GeneEMAC/>.
- Gene Expression Database for the Molecular Pharmacology of Cancer at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at [http://discover.nci.nih.gov/nature2000/Additional data at: at http://dtp.nci.nih.gov/](http://discover.nci.nih.gov/nature2000/Additional%20data%20at%20http://dtp.nci.nih.gov/).
- Gene Ontology™ (GO) consortium; available at <http://www.geneontology.org>.
- GenePix software from Axon Instruments, Inc. (Union City, California, USA); available at http://www.axon.com/GN_GenePixSoftware.html.
- GeneSpring™ from Silicon Genetics (Redwood City, California, USA); available at <http://www.sigenetics.com>.
- Genexpress IMAGE Knowledge Base of the Human Genome and Transcriptomes (Centre National de la Recherche Scientifique, CNRS; Villejuif, France); available at <http://idefix.upr420.vjf.cnrs.fr/IMAGE>.
- GenomeNet (Bioinformatics Centre at the Institute for Chemical Research of the Kyoto University, Kyoto, Japan); available at <http://www.genome.ad.jp>.
- Genzyme Molecular Oncology SAGE Database (USA); available at <http://www.sagenet.org/>.
- GEO (Gene Expression Omnibus Database) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/geo/>.
- GLS (Gene Library Summarizer) at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://cgap.nci.nih.gov/Tissues/LibrarySummarizer>.

- GMEP (genome-mean expression profile) software from Michael Eisen; available at <http://rana.lbl.gov/EisenSoftware.htm>.
- GXD (The Mouse Gene Expression Database) at the Jackson Laboratory (Bar Harbor, Maine, USA); available at http://www.informatics.jax.org/menus/expression_menu.shtml.
- Haemophilus* 2D Protein Database (University of Aberdeen, Aberdeen, Scotland, United Kingdom); accessible at <http://www.abdn.ac.uk/~mmb023/2dhome.htm>.
- Heart High-Performance 2-DE Database (Max Delbrück Center, Berlin, Germany); available at <http://www.mdc-berlin.de/~emu/heart/>.
- Heart Science Centre 2D-PAGE database (Heart Science Centre, Harefield, United Kingdom); accessible at <http://www.harefield.nthames.nhs.uk/nhli/protein/>.
- HGMP (Human Genome Mapping Project; UK MRC HGMP Resource Centre, Hinxton, Cambridge, United Kingdom); available at <http://www.hgmp.mrc.ac.uk/>.
- Hidden Markov model profiler (HMMER) at the Washington University in St. Louis (St. Louis, Montana, USA); available at <http://hmmerr.wustl.edu/>.
- HomoloGene database of the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/HomoloGene/>.
- Human Colon Carcinoma Protein Database (Ludwig Institute for Cancer Research, Melbourne, Australia); accessible at <http://www.ludwig.edu.au/jpsl/jpslhome.html>.
- Human Mitochondrial Genome Database (MITOMAP; Center for Molecular Medicine, Emory University, Atlanta, Georgia, USA); available at <http://www.gen.emory.edu/mitomap.html>.
- Human and mouse 2-D PAGE databases (Danish Centre for Human Genome Research, University of Aarhus, Aarhus, Denmark); available at <http://proteomics.cancer.dk/>.
- Human Myocardial Two-Dimensional Electrophoresis Protein Database (German Heart Institute, Virchow Clinic of the Humboldt University Berlin, Berlin, Germany); accessible at <http://userpage.chemie.fu-berlin.de/~pleiss/dhzb.html>.
- IMAGE (Integrated Molecular Analysis of Genomes and their Expression) consortium; available at <http://image.llnl.gov/>.
- Information for software developers from Developer Shed, Inc. (St. George, Utah, USA) at <http://www.devshed.com>.
- Information on FASTA by William R. Pearson (University of Virginia, Charlottesville, Virginia, USA); available at <http://www.people.virginia.EDU/~wrp/pearson.html/>.
- Information on Microsoft® Access at the Microsoft® homepage at <http://www.microsoft.com/>.
- INRA Maize Proteome Database (Institut National de la Recherche Agronomique, INRA; Gif-sur-Yvette, France); available at <http://aestivum.moulon.inra.fr/imgd/>.
- InterPro (European Bioinformatics Institute, EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/interpro>.
- ISG Database (Interferon-Stimulated Gene Database) at the Lerner Research Institute (Cleveland, Ohio, USA); available at <http://www.lerner.ccf.org/labs/williams/>.
- The Source for Java™ Technology at <http://java.sun.com>.
- Kidney Development Database at the University of Edinburgh (Edinburgh, United Kingdom); available at <http://golgi.ana.ed.ac.uk/kidhome.html>.
- Links to database services of the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/Databases/index.html>.
- MGC (Mammalian Gene Collection) at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at http://mgc.nci.nih.gov/index_html.
- MAGE-ML (Microarray Gene Expression Data Markup Language) group at the MGED; available at <http://sourceforge.net/projects/mged/>.
- MAGE-OM (Microarray Gene Expression Data Object Model) group at the MGED (Microarray Gene Expression Database) group at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://sourceforge.net/projects/mged/>.
- MAGEST (Maboya Gene Expression Patterns and Sequence Tags Database) at the Kyoto University (Kyoto, Japan); available at <http://www.genome.ad.jp/magest/>.

- MaskerAid at the Washington University in St. Louis (St. Louis, Montana, USA); available at <http://sapiens.wustl.edu/maskeraid>.
- Match (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.gene-regulation.com/pub/programs.html#match>.
- MatInspector (Genomatix Software GmbH, Munich, Germany); available at <http://www.genomatix.de/cgi-bin/matinspector/matinspector.pl>.
- MATRIX SEARCH (Center for Information Technology, National Institute of Health, NIH; Bethesda, Maryland, USA); available at <http://www.cbs.umn.edu/software/sigmatrrix.html>.
- McPromoter V3.0 (Version 3.0 of Markov Chain Promoter Finder; University of Erlangen, Germany); available at <http://promoter.informatik.uni-erlangen.de/>.
- MethDB (DNA Methylation Database) at the Centre National de la Recherche Scientifique (Montpellier, France); available at <http://www.methdb.net>.
- MGED (Microarray Gene Expression Database) group at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.mged.org>.
- MGI (Mouse Genome Informatics) website of the Jackson Laboratory (Bar Harbor, Maine, USA); available at <http://www.informatics.jax.org/>.
- MIAME (Minimum Information about a Microarray Experiment) group at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.mged.org/Annotations-wg/index.html>.
- Mitochondrial proteome database (MitoPick; Life Sciences Division of the Commissariat à l'Énergie Atomique, CEA; Paris, France); accessible at <http://www-dsv cea.fr/thema/MitoPick/Mito2D.html>.
- MySQL™ homepage at <http://www.mysql.com/>.
- National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/>.
- NCI 2DWG Image Meta-Database (Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA); available at <http://www-lecb.ncifcrf.gov/2dwgDB/>.
- Neural Network Promoter Prediction (NNPP) programme (Berkeley Drosophila Genome Project, BDGP; Lawrence Berkeley National Lab, Berkeley, California, USA); available at http://www.fruitfly.org/seq_tools/promoter.html.
- NextDB (Nematode Expression Pattern Database) at the National Institute of Genetics (Mishima, Japan); available at <http://nematode.lab.nig.ac.jp/>.
- Normalisation working group of the MGED (Microarray Gene Expression Database) group at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.dnachip.org/mged/normalisation.html>.
- Object-oriented Transcription Factors Database (ooTFD; Institute for Transcriptional Informatics, IFTI; Pittsburgh, Pennsylvania, USA); available at <http://www.ifti.org/>.
- Online tools of the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland); available at <http://www.expasy.org/tools/>.
- Ontology working group of the MGED (Microarray Gene Expression Database) group at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at http://www.cbil.upenn.edu/Ontology/MGED_ontology.html.
- Oxford English Dictionary online (OED online); available at <http://dictionary.oed.com/entrance.dtl>.
- Parasite Host Cell Interaction 2D-PAGE database (PHCI-2DPAGE; University of Aarhus, Aarhus, Denmark); accessible at <http://www.gram.au.dk/>.
- Parasite proteome maps (Institute of Biological Sciences, University of Wales, Aberystwyth, Wales, United Kingdom); accessible at <http://www.aber.ac.uk/~mpgwww/Proteome/ProtMaps.html>.
- Partial List of Web 2D Electrophoretic Gel Databases (Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA); available at <http://www-lecb.ncifcrf.gov/EP/table2Ddatabases.html>.
- PatSearch (Gesellschaft für Biotechnologische Forschung mbH, GBF mbH, Braunschweig, Germany); available at <http://transfac.gbf.de/cgi-bin/patSearch/patsearch.pl>.
- PEDB (Prostate Expression Database) at the Fred Hutchinson Cancer Research Center

- (Seattle, Washington, USA); available at <http://www.pedb.org/>.
- Comprehensive Perl Archive Network (CPAN) at <http://www.cpan.org>.
- PhosphoProtein Database (; Laboratory of Experimental and Computational Biology, National Cancer Institute, Bethesda, Maryland, USA); accessible at <http://www-lecb.ncifcrf.gov/phosphoDB/>.
- PHP group; available at <http://www.php.net>.
- Phrap (phragment assembly programme) on the Phred/Phrap/Consed System homepage; available at <http://www.phrap.org/>.
- Phred on the Phred/Phrap/Consed System homepage; available at <http://www.phrap.org/>.
- Phred/Phrap/Consed System homepage at <http://bozeman.mbt.washington.edu/>.
- Plant Plasma Membrane Database (PPMdb; Gent University, Gent, Belgium); accessible at <http://sphinx.rug.ac.be:8080/ppmdb/index.html>.
- PlasmoDB (*Plasmodium* Genome Resource Database) of the *Plasmodium falciparum* Genome Project; available at <http://plasmodb.org>.
- PMMA-2DPAGE at Purkyne Military Medical Academy (Czech); accessible at <http://www.pmma.pmfhk.cz/>.
- PostgreSQL homepage at <http://www.postgresql.org/>.
- PromFind (IUBio Archive, Indiana University, Bloomington, Indiana, USA); available at <http://iubio.bio.indiana.edu/soft/molbio/mswin/mswin-or-dos/profin11.exe>.
- Promoter 2.0 (Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark); available at <http://www.cbs.dtu.dk/services/promoter>.
- PromoterInspector (Genomatix Software GmbH, Munich, Germany); available at <http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl>.
- PromoterScan (Center for Information Technology, National Institute of Health, NIH; Bethesda, Maryland, USA); available at <http://bimas.dcrf.nih.gov/molbio/proscan/>.
- Protein database of haematopoietic cell lines (BPP-2DPAGE ; Laboratoire de Biochimie et Technologie des Proteines, Bobigny, France); accessible at <http://www-smbh.univ-paris13.fr/lbtp/Biochemistry/biochimie/bque.htm>.
- Protein database of the Statens Serum Institute (SSI-2DPAGE; Statens Serum Institute, Copenhagen, Denmark); available at http://www.ssi.dk/en/forskning/tbimmun/Protein_database/protein_database.htm.
- Protein Disease Database (PDD; Laboratory of Experimental and Computational Biology, National Cancer Institute, Bethesda, Maryland, USA); accessible at <http://www-lecb.ncifcrf.gov/PDD/>.
- Protein expression database for the molecular pharmacology of cancer at the National Cancer Institute (NCI; Bethesda, Maryland, USA); available at <http://discover.nci.nih.gov/host/prot.html>.
- Protein Information Resource (PIR; National Biomedical Research Foundation, Georgetown University Medical Center, Northwest Washington, District of Columbia, USA); available at <http://pir.georgetown.edu/>.
- Proteome Analysis Database at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/proteome/>.
- Proteome 2D-PAGE database at the Max Planck Institute (MPI) for Infection Biology (Berlin, Germany); accessible at <http://www.mpiib-berlin.mpg.de/2D-PAGE/>.
- Proteome[®]s BioKnowledge[®] Library (Yeast Proteome Database, YPD[™]; Human Proteome Survey Database, HumanPSD[™]; *Caenorhabditis elegans* Proteome Database, WormPD[™]; MycoPath Proteome Database, MycoPathDB[™]; *Schizosaccharomyces pombe* Proteome Database, PombeDB[™]) at Proteome, Inc. (Beverly, Massachusetts, USA); available at <http://www.proteome.com/databases/index.html>.
- Proteome Database for Aging Research (TMIG-2DPAGE; Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan); accessible at <http://proteome.tmig.or.jp/2D/>.
- Purkyne Military Medical Academy 2D-PAGE Database (PMMA-2DPAGE; Purkyne Military Medical Academy, Czech); accessible at <http://www.pmma.pmfhk.cz/>.
- RAD (RNA Abundance Database) at the University of Pennsylvania (Philadelphia, Pennsylvania, USA); available at <http://www.cbil.upenn.edu/RAD2/>.
- Radiation Hybrid Database at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/RHdb/>.

- Rat Serum Protein Study Group database (Department of Pharmacological Science, University of Milan, Milan, Italy); accessible at <http://linux.farma.unimi.it/>.
- Rat Serum Protein Database (Aebersold Protein Laboratory, University of Washington, Seattle, Washington, USA); accessible at <http://depts.washington.edu/~ruedilab/info/ratserum/ratserum.html>.
- RECODE (Database of the Translational Recoding Events) at the University of Utah (Salt Lake City, Utah, USA); available at <http://recode.genetics.utah.edu/>.
- RepeatMasker at the University of Washington (Seattle, Washington, USA); available at <http://ftp.genome.washington.edu/>.
- Research Genetics (ResGen™; Huntsville, Alabama, USA); available at <http://www.res-gen.com>.
- Rochester Muscle Database at the University of Rochester (Rochester, New York, USA); available at <http://www.urmc.rochester.edu/smd/CRC/SWindex.html>.
- R resources of the Comprehensive R Archive Network (CRAN) at <http://cran.r-project.org/>.
- RZPD (German Resource Center Primary Database) at the Deutsches Ressourcenzentrum für Genomforschung GmbH, Berlin, Germany); available at <http://www.rzpd.de/>.
- SAGEmap (serial analysis of gene expression tag to gene mapping database) at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/SAGE/>.
- SAGEmap Virtual Northern at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/SAGE/sagevn.cgi>.
- SAGEmap xProfiler (National Cancer Institute, NCI; Bethesda, Maryland, USA) available at <http://www.ncbi.nlm.nih.gov/SAGE/sagexp-setup.cgi>.
- SAM (Sequence Alignment and Modeling System) at the University of California, (Santa Cruz, California, USA); available at <http://www.cse.ucsc.edu/research/compbio/sam.html>.
- Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.sanger.ac.uk>.
- Scanalyze software from Michael Eisen; available at <http://rana.lbl.gov/EisenSoftware.htm>.
- SGD (Saccharomyces Genome Database) at the Stanford University (Palo Alto, California, USA); available at <http://genome-www.stanford.edu/Saccharomyces/> and <http://genome-www4.stanford.edu/cgi-bin/SGD/expression/expressionConnection.pl>.
- SIENA-2DPAGE at the Department of Molecular Biology at the University of Siena (Siena, Italy); accessible at <http://www.bio-mol.unisi.it/2d/2d.html>.
- SIGNAL SCAN (Center for Information Technology, National Institute of Health, NIH; Bethesda, Maryland, USA); available at <http://bimas.dcrn.nih.gov/molbio/signal/index.html>.
- SITE (Softberry, Inc., Mount Kisco, New York, USA); available at <http://www.softberry.com/>.
- SMD (Stanford Microarray Database) at the Stanford University (Palo Alto, California, USA); available at <http://genome-www.stanford.edu/microarray/>.
- SOURCE (Stanford Online Universal Resource for Clones and ESTs) at the Stanford University (Palo Alto, California, USA); available at <http://genome-www.stanford.edu/source>.
- SQL tutorial of Refsnes Data (Norway) at <http://www.w3schools.com/sql>.
- STACK (Sequence Tag Alignment and Consensus Knowledgebase; South African National Bioinformatics Institute, University of Western Cape, Bellville South Africa); available at <http://www.sanbi.ac.za/Dbases.html>.
- Stanford Yeast Database at the Stanford University (Palo Alto, California, USA); available at <http://cellcycle-www.stanford.edu/>.
- Terry Speed's Microarray Data Analysis Group (Department of Statistics, University of California, Berkeley, California, USA) at <http://www.stat.berkeley.edu/users/terry/zarray/Html/links.html>.
- Streptomyces coelicolor* Reference Gel (University of Wales, Aberystwyth, Wales, United Kingdom); accessible at http://gepasi.dbs.aber.ac.uk/S_Coeli/ReferenceGel/refgel.html.
- Swiss-Czech Proteomics Server (SWICZ; Institute of Microbiology, Academy of Sciences in Prague, Czech Republic, and Division of Molecular Microbiology, Biozentrum, University of Basel, Basel Switzerland); accessible at <http://proteom.biomed.cas.cz/>.
- SWISS-2DPAGE (two-dimensional polyacrylamide gel electrophoresis database) at the Swiss Institute of Bioinformatics (SIB; Gen-

- eva, Switzerland); available at <http://www.expasy.org/ch2d/>.
- SWISS-PROT at the Swiss Institute of Bioinformatics (SIB, Geneva, Switzerland); available at <http://www.expasy.org/sprot/>.
- The R Project for Statistical Computing at <http://www.r-project.org/>.
- TIGEM (Telethon Institute of Genetics and Medicine, Naples, Italy); available at <http://www.tigem.it/>.
- The TIGR (The Institute of Genomic Research; Rockville, Maryland, USA) gene indices; available at <http://www.tigr.org/tdb/tgi.shtml>.
- Tooth Development Database at the University of Helsinki (Helsinki, Finland); available at <http://bite-it.helsinki.fi/>.
- ToothPrint (University of Otago, Dunedin, New Zealand); accessible at <http://bioc111.otago.ac.nz:591/tooth/home.htm>.
- Toxoplasma gondii* proteome map (Institute for Medical Microbiology and Virology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany); accessible at <http://www-public.rz.uni-duesseldorf.de/~hfischer/>.
- TRANSCompel (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.gene-regulation.com/pub/databases.html#transcompel>.
- Transcription Factor Database (TFD; Institute for Transcriptional Informatics, IFTI; Pittsburgh, Pennsylvania, USA); available at <http://www.ifti.org/>.
- TRANSFAC database (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.biobase.de> or <http://www.gene-regulation.com/pub/databases.html#transfac>.
- TRANSPLOER (BIOBASE Biological Databases GmbH, Wolfenbüttel, Germany); available at <http://www.biobase.de/>.
- TrEMBL at the European Bioinformatics Institute (EBI; Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.ebi.ac.uk/trembl/>.
- TRIPLES (Transposon-Insertion Phenotypes, Localisation, and Expression in *Saccharomyces cerevisiae* Database) at the Yale University (New Haven, Connecticut, USA); available at <http://ygac.med.yale.edu>.
- TRRD (Transcription Regulatory Region Database; Institute Cytology and Genetics, Union of Soviet Socialist Republics Academy of Science, Novosibirsk, Russia); available at <http://www.bionet.nsc.ru/trrd/celcyc/>.
- TSSG (Softberry, Inc., Mount Kisco, New York, USA); available at <http://www.softberry.com>.
- TSSW (Softberry, Inc., Mount Kisco, New York, USA); available at <http://www.softberry.com>.
- Two-dimensional polyacrylamide gel electrophoresis database for broncho-alveolar lavage fluid proteins (BALF 2D; Department of Biological Chemistry, University of Mons-Hainaut, Mons-Hainaut, Belgium); accessible at <http://www.umh.ac.be/~biochim/BALF2D.html>.
- UniGene website at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/UniGene/>.
- Unified Library Database at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <http://www.ncbi.nlm.nih.gov/UniLib/>.
- UTRdb and UTRsite (Specialised Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs) at the University of Milan, Milan, Italy); available at <http://bighost.area.ba.cnr.it/BIG/UTRHome/>.
- Visual Basic Bookmark at <http://www.vb-bookmark.com/>.
- Washington University Inner Ear Protein Database (Department of Otolaryngology, Washington University, St. Louis, Missouri, USA); accessible at <http://oto.wustl.edu/thc/innerear2d.htm>.
- Whitehead Institute (Cambridge; Massachusetts, USA); available at <http://www.wi.mit.edu/>.
- WORLD-2DPAGE of the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB; Geneva, Switzerland); accessible at <http://www.expasy.org/ch2d/2d-index.html>.
- Wise2 at the Sanger Institute (Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom); available at <http://www.sanger.ac.uk/Software/Wise2>.
- World Wide Web Consortium (W3C) at <http://www.w3.org> and in particular on XML at <http://www.w3.org/XML>.
- Wu-BLAST at the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, USA); available at <ftp://ftp.ncbi.nih.gov/blast/>.

- WWW Interface to Sequence Analysis Software Tools (W²H) at the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ; Heidelberg, Germany) at <http://industry.ebi.ac.uk/w2h/>.
- Yeast 2D-PAGE Protein Database (YEAST 2D-PAGE; Lundberg Laboratory, University of Gothenburg, Gothenburg, Sweden); accessible at <http://yeast-2dpage.gmm.gu.se/>.
- Yeast genome-wide expression database at the Whitehead Institute for Biomedical Research (Cambridge, Massachusetts, USA); available at <http://web.wi.mit.edu/young/expression/>.
- Yeast protein database (COMPLUYEAST-2DPAGE database; Universidad Complutense de Madrid, Madrid, Spain); accessible at <http://babbage.csc.ucm.es/2d/>.
- Yeast Protein Map (YPM; Institute of Biochemistry and Cellular Genetics, Bordeaux Cedex, France); accessible at <http://www.ibgc.u-bordeaux2.fr/YPM/>.
- yMGV (Yeast Microarray Global Viewer) at the Ecole Normale Supérieure (Paris, France); available at <http://www.transcriptome.ens.fr/ymgv/>.
- Xylella 2D Database (Laboratory of Protein Chemistry, Laquip Unicamp, Sao Paulo, Brazil); accessible at <http://proteome.ibi.unicamp.br/databases/index.htm>.
- ### Cited publications
- Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **1991**; 252(5013):1651-1656.
- Aebersold, R., and Leavitt, J. Sequence analysis of proteins separated by polyacrylamide gel electrophoresis: towards an integrated protein database. *Electrophoresis* **1990**; 11(7): 517-527.
- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, Hudson, T. J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**; 403(6769):503-511.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(12):6745-6750.
- Altmann, C. R., Bell, E., Sczyrba, A., Pun, J., Bekiranov, S., Gaasterland, T., and Brivanlou, A. H. Microarray analysis pipeline: Microarray-based analysis of early development in *Xenopus laevis*. *Developmental Biology* **2001**; 236(1):64-75.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **1990**; 215(3):403-410.
- Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. Issues in searching molecular sequence databases. *Nature Genetics* **1994**; 6(2):119-129.
- Altschul, S. F., and Gish, W. Local alignment statistics. *Methods in Enzymology* **1996**; 266:460-480.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **1997**; 25(17):3389-3402.
- Anderson, D. J., and Axel, R. Molecular probes for the development and plasticity of neural crest derivatives. *Cell* **1985**; 42(2):649-662.
- Anderson, N. G., Matheson, A., and Anderson, N. L. Back to the future: the human protein index (HPI) and the agenda for post-proteomic biology. *Proteomics* **2001**; 1(1):3-12.
- Appel, R. D., Sanchez, J. C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., and Hochstrasser, D. F. SWISS-2DPAGE: a database of two-dimensional gel electrophoresis images. *Electrophoresis* **1993**; 14(11):1232-1238.
- Appel, R. D., Sanchez, J. C., Bairoch, A., Golaz, O., Ravier, F., Pasquali, C., Hughes, G. J., and Hochstrasser, D. F. The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Research* **1994**; 22(17):3581-3582.
- Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R., Golaz, O., Pasquali, C., and Hochstrasser, D. F. Federated two-dimensional electrophoresis database: a simple means of

- publishing two-dimensional electrophoresis data. *Electrophoresis* **1996a**; 17(3):540-546.
- Appel, R. D., Sanchez, J. C., Bairoch, A., Golaz, O., Ravier, F., Pasquali, C., Hughes, G. J., and Hochstrasser, D. F. The SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis, its status in 1995. *Nucleic Acids Research* **1996b**; 24(1):180-181.
- Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J., and Zdobnov, E. M.; The InterPro Consortium. InterPro — an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **2000**; 16(12):1145-1150.
- Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E. V., Mittard, V., Mulder, N., Phan, I., and Zdobnov, E. Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Research* **2001a**; 29(1):44-48.
- Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., and Servant, F. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* **2001b**; 29(1):37-40.
- Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(21):11319-11324.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **2000**; 25(1):25-29.
- Bahl, A., Brunk, B., Coppel, R. L., Crabtree, J., Diskin, S. J., Fraunholz, M. J., Grant, G. R., Gupta, D., Huestis, R. L., Kissinger, J. C., Labo, P., Li, L., McWeeney, S. K., Milgram, A. J., Roos, D. S., Schug, J., and Stoeckert, C. J., Jr. PlasmoDB: The *Plasmodium* Genome Resource. An integrated database providing tools for accessing and analyzing mapping, expression and sequence data (both finished and unfinished). *Nucleic Acids Research* **2002**; 30(1):87-90.
- Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Research* **1996**; 24(1):21-25.
- Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Research* **1997**; 25(1):31-36.
- Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Research* **1998**; 26(1):38-42.
- Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research* **1999**; 27(1):49-54.
- Bairoch, A., and Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **2000**; 28(1):45-48.
- Bajic, V. B., Seah, S. H., Chong, A., Zhang, G., Koh, J. L. Y., and Brusic, V. Dragon Promoter Finder: recognition of vertebrate RNA Polymerase II promoters. *Bioinformatics* **2002**; 18(1):198-199.
- Baker, C. S., Corbett, J. M., May, A. J., Yacoub, M. H., and Dunn, M. J. A human myocardial two-dimensional electrophoresis database: protein characterisation by microsequencing and immunoblotting. *Electrophoresis* **1992**; 13(9-10):723-726.
- Baldi, P., and Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **2001**; 17(6):509-519.
- Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Kasarskis, A., Scafe, C. R., Sherlock, G., Binkley, G., Jin, H., Kasper, M., Orr, S. D., Schroeder, M., Weng, S., Zhu, Y., Botstein, D., and Cherry, J. M., Inte-

- grating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Research* **2000**; *28*(1):77-80.
- Ball, C., Jin, H., Sherlock, G., Weng, S., Matlese, J. C., Andrada, R., Binkley, G., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Schroeder, M., Botstein, D., and Cherry, J. M. *Saccharomyces* Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic Acids Research* **2001**; *29*(1):80-81.
- Bard, J. B. L., McConnell, J. E., and Davies, J. A. Towards a genetic basis for kidney development. *Mechanisms of Development* **1994**; *48*(1):3-11.
- Bard, J. B. L., Kaufman, M. H., Dubreuil, C., Brune, R. M., Burger, A., Baldock, R. A., and Davidson, D. R. An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mechanisms of Development* **1998**; *74*(1-2): 111-120.
- Baranov, P. V., Gurvich, O. L., Fayet, O., Prere, M. F., Miller, W. A., Gesteland, R. F., Atkins, J. F., and Giddings, M. C. RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Research* **2001**; *29*(1):264-267.
- Barrett, J., Jefferies, J. R., and Brophy, P. M. Parasite proteomics. *Parasitology Today* **2000**; *16*(9):400-403.
- Bedell, J. A., Korf, I., and Gish, W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **2000**; *16*(11):1040-1041.
- Ben-Dor, A., Shamir, R., and Yakhini, Z. Clustering gene expression patterns. *Journal of Computational Biology* **1999**; *6*(3-4): 281-297.
- Ben Mamoun, C., Gluzman, I. Y., Hott, C., MacMillan, S. K., Amarakone, A. S., Anderson, D. L., Carlton, J. M., Dame, J. B., Chakrabarti, D., Martin, R. K., Brownstein, B. H., Goldberg, D. E. Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis. *Molecular Microbiology* **2001**; *39*(1):26-36.
- Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A., and Wheeler, D. L. GenBank. *Nucleic Acids Research* **1999**; *27*(1):12-17.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., Wheeler, D. L. GenBank. *Nucleic Acids Research* **2000**; *28*(1):15-18.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. GenBank. *Nucleic Acids Research* **2002**; *30*(1):17-20.
- Berg, O. G., and von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology* **1987**; *193*(4):723-750.
- Berg, O. G., von Hippel, P. H. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *Journal of Molecular Biology* **1988**; *200*(4):709-723.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**; *28*(1): 235-242.
- Bini, L., Sanchez-Campillo, M., Santucci, A., Magi, B., Marzocchi, B., Comanducci, M., Christiansen, G., Birkelund, S., Cevenini, R., Vretou, E., Ratti, G., and Pallini, V. Mapping of *Chlamydia trachomatis* proteins by immobilized-polyacrylamide two-dimensional electrophoresis: spot identification by N-terminal sequencing and immunoblotting. *Electrophoresis* **1996**; *17*(1):185-190.
- Bini, L., Magi, B., Marzocchi, B., Cellesi, C., Berti, B., Raggiaschi, R., Rossolini, A., and Pallini, V. Two-dimensional electrophoretic patterns of acute-phase human serum proteins in the course of bacterial and viral diseases. *Electrophoresis* **1996**; *17*(3):612-616.
- Bini, L., Magi, B., Marzocchi, B., Arcuri, F., Tripodi, S., Cintonino, M., Sanchez, J. C., Frutiger, S., Hughes, G., Pallini, V., Hochstrasser, D. F., and Tosi, P. Protein expression profiles in human breast ductal carcinoma and histologically normal tissue. *Electrophoresis* **1997**; *18*(15):2832-2841.
- Bini, L., Heid, H., Liberatori, S., Geier, G., Pallini, V., and Zwilling, R. Two-dimensional gel electrophoresis of *Caenorhabditis elegans* homogenates and identification of protein spots by microsequencing. *Electrophoresis* **1997**; *18*(3-4):557-562.
- Birkelund, S., Bini, L., Pallini, V., Sanchez-Campillo, M., Liberatori, S., Clausen, J. D., Ostergaard, S., Holm, A., and Christiansen, G. Characterization of *Chlamydia trachomatis* 12-induced tyrosine-phosphorylated HeLa cell

- proteins by two-dimensional gel electrophoresis. *Electrophoresis* **1997**; 18(3-4):563-567.
- Birney, E., Thompson, J. D., and Gibson, T. J. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Research* **1996**; 24(14):2730-2739.
- Blake, J. A., Eppig, J. T., Richardson, J. E., Bult, C. J., and Kadin, J. A. The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. *Nucleic Acids Research* **2001**; 29(1):91-94.
- Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST – database for “expressed sequence tags”. *Nature Genetics* **1993**; 4(4): 332-333.
- Boguski, M. S., and Schuler, G. D. ESTablishing a human transcript map. *Nature Genetics* **1995**; 10 (4):369-371.
- Bolch, B. W., and Huang, C. J. Multivariate statistical methods for business and economics. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA **1974**.
- Bortoluzzi, S., d’Alessi, F., Romualdi, C., and Danieli, G. A. The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach. *Genome Research* **2000**; 10(3):344-349.
- Boucherie, H., Sagliocco, F., Joubert, R., Maillet, I., Labarre, J., and Perrot, M. Two-dimensional gel protein database of *Saccharomyces cerevisiae*. *Electrophoresis* **1996**; 17(11):1683-1699.
- Bouck, J., Yu, W., Gibbs, R., and Worley, K. Comparison of gene indexing databases. *Trends in Genetics* **1999**; 15(4):159-162.
- Brazma, A. On the importance of standardisation in life sciences. *Bioinformatics* **2001**; 17(2):113-114.
- Brazma, A., Vilo, J., and Ukkonen, E. Finding transcription factor binding site combinations in the yeast genome. In: GCB 1997, Proceedings of the German Conference on Bioinformatics, Kloster Irsee, Bavaria, Germany, September 22-24, 1997. Mewes, H. W., and Frishman, D., editors. **1997**:57-60.
- Brazma, A., and Vilo, J. Gene expression data analysis. *FEBS Letters* **2000**; 480(1): 17-24.
- Brazma, A., Robinson, A., Cameron, G., and Ashburner, M. One-stop shop for microarray data. *Nature* **2000**; 403(6771):699-700.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics* **2001**; 29(4):365-371.
- Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., and Bork, P. EST comparison indicates 38 % of human mRNAs contain possible alternative splice forms. *FEBS Letters* **2000**; 474(1):83-86.
- Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, S., Ares, M., and Haussler, D. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences of the USA* **1999**; 97(1): 262-267.
- Brune, R. M., Bard, J. B. L., Dubreuil, C., Guest, E., Hill, W., Kaufman, M. K., Stark, M., Davidson, D., and Baldock, R. A. A three-dimensional model of the mouse at embryonic day 9. *Developmental Biology* **1999**; 216(2):457-468.
- Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology* **1990**; 212(4):563-578.
- Bumann, D., Meyer, T. F., and Jungblut, P. R. Proteome analysis of the common human pathogen *Helicobacter pylori*. *Proteomics* **2001**; 1(4):473-479.
- Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research* **2000**; 10(12):2022-2029.
- Candia, A. L., Hu, J., Crosby, J., Lalle, P. A., Noden, D., Nadeau, J. H., and Wright, C. V. Mox-1 and Mox-2 define a novel homeobox subfamily and are differentially expressed during early mesodermal patterning in mouse embryos. *Development* **1992**; 116(4):1123-1136.
- Cash, P., Argo, E., and Bruce, K. D. Characterisation of *Haemophilus influenzae* proteins by two-dimensional gel electrophoresis. *Electrophoresis* **1995**; 16(1):135-148.
- Cash, P., Argo, E., Langford, P. R., and Kroll, J. S. Development of a *Haemophilus* two-di-

- mensional protein database. *Electrophoresis* **1997**; *18*(8):1472-1482.
- Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell* **2001**; *12*(2):323-337.
- Cavalcoli, J. D. Genomic and proteomic databases: large-scale analysis and integration of data. *Trends in Cardiovascular Medicine* **2001**; *11*(2):76-81.
- Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Kwee, S., Rasmussen, H. H., Nielsen, H. V., Cruger, D., Basse, B., Leffers, H., Honoré, B., Møller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M., and Van den Bulcke, M. Comprehensive, human cellular protein databases and their implication for the study of genome organization and function. *FEBS Letters* **1989a**; *244*(2):247-254.
- Celis, J. E., Gesser, B., Dejgaard, K., Honore, B., Leffers, H., Madsen, P., Andersen, A., Basse, B., Celis, A., Lauridsen, J. B., *et al.* Two dimensional gel human protein databases offer a systematic approach to the study of cell proliferation and differentiation. *International Journal of Developmental Biology* **1989b**; *33*(4):407-416.
- Celis, J. E., Honore, B., Bauw, G., and Vandekerckhove, J. Comprehensive computerized 2D gel protein databases offer a global approach to the study of the mammalian cell. *Bioessays* **1990a**; *12*(2):93-97.
- Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honore, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B., Ratz, G., *et al.* The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent, proliferating and SV40 transformed cells. *Electrophoresis* **1990b**; *11*(12):1072-1113.
- Celis, J. E., Gesser, B., Rasmussen, H. H., Madsen, P., Leffers, H., Dejgaard, K., Honore, B., Olsen, E., Ratz, G., Lauridsen, J. B., *et al.* Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* **1990c**; *11*(12):989-1071.
- Celis, J. E., Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Magnusson, N., Kiil, J., *et al.* A comprehensive two-dimensional gel protein database of noncultured unfractionated normal human epidermal keratinocytes: towards an integrated approach to the study of cell proliferation, differentiation and skin diseases. *Electrophoresis* **1991a**; *12*(11):802-872.
- Celis, J. E., Leffers, H., Rasmussen, H. H., Madsen, P., Honore, B., Gesser, B., Dejgaard, K., Olsen, E., Ratz, G. P., Lauridsen, J. B., *et al.* The master two-dimensional gel database of human AMA cell proteins: towards linking protein and genome sequence and mapping information (update 1991). *Electrophoresis* **1991b**; *12*(11):765-801.
- Celis, J. E., Rasmussen, H. H., Leffers, H., Madsen, P., Honore, B., Gesser, B., Dejgaard, K., and Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB Journal* **1991c**; *5*(8):2200-2208.
- Celis, J. E., Rasmussen, H. H., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gesser, B., Olsen, E., Gromov, P., Hoffmann, H. J., Nielsen, M., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Nielsen, M., Andersen, A. H., Walbum, E., Kjaergaard, I., Puype, M., Van Damme, J., and Vandekerckhove, J. The human keratinocyte two-dimensional gel protein database (update 1992): towards an integrated approach to the study of cell proliferation, differentiation and skin diseases. *Electrophoresis* **1992**; *13*(12):893-959.
- Celis, J. E., Rasmussen, H. H., Leffers, H., Madsen, P., Honore, B., Dejgaard, K., Gromov, P., Olsen, E., Hoffmann, H. J., Nielsen, M., Gesser, M., Puype, Van Damme, J., and Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA mapping and sequencing data: towards an integrated approach to the study of gene expression. *Genetic Engineering (New York)* **1993**; *15*:21-40.
- Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honore, B., Dejgaard, K., Gromov, P., Vorum, H., Vassilev, A., Baskin, Y., Liu, X., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Walbum, E., Kjaergaard, I., Andersen, I., Puype, M., Van Damme, J., and Vandekerckhove, J. The

- human keratinocyte two-dimensional protein database (update 1994): towards an integrated approach to the study of cell proliferation, differentiation and skin diseases. *Electrophoresis* **1994**; 15(11):1349-1458.
- Celis, J. E., Gromov, P., Ostergaard, M., Madsen, P., Honore, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., and Rasmussen, H. H. Human 2-D PAGE databases for proteome analysis in health and disease: <http://bio-base.dk/cgi-bin/celis>. *FEBS Letters* **1996**; 398(2-3):129-34.
- Celis, J. E., Ostergaard, M., Jensen, N. A., Gromova, I., Rasmussen, H. H., and Gromov, P. Human and mouse proteomic databases: novel resources in the protein universe. *FEBS Letters* **1998**; 430(1-2):64-72.
- Chakravarti, D. N., Fiske, M. J., Fletcher, L. D., and Zagursky, R. J. Mining genomes and mapping proteomes: identification and characterization of protein subunit vaccines. *Developmental Biology (Basel)* **2000**; 103:81-90.
- Chataway, T. K., Whittle, A. M., Lewis, M. D., Bindloss, C. A., Moritz, R. L., Simpson, R. J., Hopwood, J. J., and Meikle, P. J. Development of a two-dimensional gel electrophoresis database of human lysosomal proteins. *Electrophoresis* **1998**; 19(5):834-836.
- Chen, Q. K., Hertz, G. Z., and Stormo, G. D. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Computer Applications in Biosciences* **1995**; 11(5):563-566.
- Chen, X., Dress, A., Karas, H., Reuter, I., and Wingender, E. A database framework for mapping expression patterns Computer Science and Biology. In: Proceedings of the German Conference on Bioinformatics (GCB 1999). Giegerich, R., Hofestädt, R., Lengauer, T., Mewes, W., Schomburg, D., Vingron, M., and Wingender, E., editors. GBF-Braunschweig and University of Bielefeld, Hanover, Germany **1999**:174-178.
- Chen, H., Higgins, J., Kondorosi, E., Kondorosi, A., Djordjevic, M. A., Weinman, J. J., and Rolfe, B. G. Identification of nolR-regulated proteins in *Sinorhizobium meliloti* using proteome analysis. *Electrophoresis* **2000**; 21(17):3823-3832.
- Chen, H., Higgins, J., Oresnik, I. J., Hynes, M. F., Natera, S., Djordjevic, M. A., Weinman, J. J., and Rolfe, B. G. Proteome analysis demonstrates complex replicon and luteolin interactions in pSyma-cured derivatives of *Sinorhizobium meliloti* strain 2011. *Electrophoresis* **2000**; 21(17):3833-3842.
- Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J. M., and Botstein, D. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **1998**; 282(5396):2022-2028.
- Cho, R. J., Campbell, M. J., Winzler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, W., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* **1999**; 2(1):65-73.
- Christoffels, A., van Gelder, A., Greyling, G., Miller, R., Hide, T., and Hide, W. STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Research* **2001**; 29(1):234-238.
- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. The transcriptional program of sporulation in budding yeast. *Science* **1998**; 282(5389):699-705. Erratum in: *Science* **1998**; 282(5393):1421.
- Corbett, J. M., Wheeler, C. H., Baker, C. S., Yacoub, M. H., and Dunn, M. J. The human myocardial two-dimensional gel protein database: update 1994. *Electrophoresis* **1994**; 15(11):1459-1465.
- Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., Lengieza, C., Lew-Smith, J. E., Tillberg, M., and Garrels, J. I. YPDTM, PombePDTM and WormPDTM: model organism volumes of the BioKnowledgeTM Library, an integrated resource for protein information. *Nucleic Acids Research* **2001**; 29(1):75-79.
- Costello, J. F. and Plass, C. Methylation matters. *Journal of Medical Genetics* **2001**; 38(5):285-303.
- Date, C. J., and Darwen, H. A Guide to SQL standard. Addison Wesley Publishing Company, Reading, Massachusetts, USA **1996**.
- Davidson, D., and Baldock, R. Bioinformatics beyond sequence: mapping gene function in the embryo. *Nature Reviews Genetics* **2001**; 2(6):409-417.

- Davidson, D., Bard, J., Kaufman, M. H., and Baldock, R. The Mouse Atlas Database: a community resource for mouse development. *Trends in Genetics* **2001a**; 17(2):49-51.
- Davidson, S. B., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, G. C., and Stoeckert, C. J., Jr. K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal* **2001b**; 40(2):512-531.
- Davies, J. A., and Bard, J. B. L. Development of the metanephric kidney. *Current Topics in Developmental Biology* **1998**; 39:245-302.
- Davies, J. A., Hunter, D. A., Brandli, A., and Nieminen, P. Design considerations for small, special system developmental databases. *Seminars in Cell and Developmental Biology* **1997**; 8(5):519-525.
- Davuluri, R. V., Suzuki, Y., Sugano, S., and Zhang, M. Q. CART classification of human 5' UTR sequences. *Genome Research* **2000**; 10(11):1807-1816.
- Davuluri, R. V., Grosse, I., and Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nature Genetics* **2001**; 29(4):412-417.
- Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., S. Fox, Gelling, S., Green, L., Harrison, P., Hocking, R., Holloway, E., Hunt, S., Keil, S., Lijnzaad, P., Louis-Dit-Sully, C., Ma, J., Mendis, A., Miller, J., Morissette, J., Muselet, D., Nusbaum, H. C., Peck, A., Rozen, S., Simon, D., Slonim, D. K., Staples, R., Stein, L. D., Stewart, E. A., Suchard, M. A., Thangarajah, T., Vega-Czarny, N., Webber, C., Wu, X., Hudson, J., Auffray, C., Nomura, N., Sikela, J. M., Polymeropoulos, M. H., James, M. R., Lander, E. S., Hudson, T. J., Myers, R. M., Cox, D. R., Weissbach, J., Boguski, M. S., and Bentley, D. R. A physical map of 30,000 human genes. *Science* **1998**; 282(5389):744-746.
- Der, S. D., Zhou, A., Williams, B. R., and Silverman, R. H. Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA* **1988**; 95(26):15623-15628.
- DeRisi, J., Iyer, V. R., and Brown, P. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**; 278(5338):680-686.
- De Veer, M. J., Holko, M., Frevel, M., Walker, E., Der, S., Paranjape, J. M., Silverman, R. H., and Williams, B. R. Functional classification of interferon-stimulated genes identified using microarrays. *J Leukocyte Biology* **2001**; 69(6):912-920.
- Dralyuk, I., Brudno, M., Gelfand, M. S., Zorn, M., and Dubchak, I. ASDB: database of alternatively spliced genes. *Nucleic Acids Research* **2000**; 28(1):296-297.
- Dickmeis, T., Aanstad, P., Clark, M., Fischer, N., Herwig, R., Mourrain, P., Blader, P., Rosa, F., Lehrach, H., and Strahle, U. Identification of nodal signalling targets by array analysis of induced complex probes. *Developmental Dynamics* **2001**; 222(4):571-580.
- Dunn, M. J., Corbett, J. M., and Wheeler, C. H. HSC-2DPAGE and the two-dimensional gel electrophoresis database of dog heart proteins. *Electrophoresis* **1997**; 18(15):2795-2802.
- Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research* **2002**; 30(1):69-72.
- Eberini, I., Miller, I., Zancan, V., Bolego, C., Puglisi, L., Gemeiner, M., and Gianazza, E. Proteins of rat serum IV. Time-course of acute-phase protein expression and its modulation by indomethacine. *Electrophoresis* **1999a**; 20(4-5):846-853.
- Eberini, I., Miller, I., Gemeiner, M., Haynes, P., Aebersold, R., Puglisi, L., Sirtori, C. R., and Gianazza, E. A website for the rat serum protein study group. *Electrophoresis* **1999b**; 20(18):3599-3602.
- Ebrini, I., Agnello, D., Miller, I., Villa, P., Fratelli, M., Ghezzi, P., Gemeiner, M., Chan, J., Aebersold, R., and Gianazza, E. Proteins of rat serum V: adjuvant arthritis and its modulation by nonsteroidal anti-inflammatory drugs. *Electrophoresis* **2000**; 21(11):2170-2179.
- Eddy, S. R. Hidden Markov models. *Current Opinion in Structural Biology* **1996**; 6(3):361-365.
- Edgar, R., Domrachev, M., and Lash, A. E. Gene Expression Omnibus: NCBI gene ex-

- pression and hybridization array data repository. *Nucleic Acids Research* **2002**; 30(1): 207-210.
- Ehret, G. B., Reichenbach, P., Schindler, U., Horvath, C. M., Fritz, S., Nabholz, M., and Bucher, P. DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites. *Journal of Biological Chemistry* **2001**; 276(9):6675-6688.
- Eickhoff, H., Schuchhardt, J., Ivanov, I., Meier-Ewert, S., O'Brien, J., Malik, A., Tandon, N., Wolski, E. W., Rohlf, E., Nyarsik, L., Reinhardt, R., Nietfeld, W., and Lehrach, H. Tissue gene expression profiling using arrayed normalised cDNA libraries. *Genome Research* **2000**; 10(8):1230-1240.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* **1998**; 95(25): 14863-14868.
- Ericsson, C., Petho, Z., and Mehlin, H. An online two-dimensional polyacrylamide gel electrophoresis protein database of adult *Drosophila melanogaster*. *Electrophoresis* **1997**; 18(3-4):484-490.
- Ermolaeva, O., Rastogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M., and Boguski, M. S. Data management and analysis for gene expression arrays. *Nature Genetics* **1998**; 20(1):19-23.
- Etzold, T., Ulyanov, A., and Argos, P. SRS: information retrieval system for molecular biology data banks. *Methods in Enzymology* **1996**; 266:114-128.
- Evans, G., Wheeler, C. H., Corbett, J. M., and Dunn, M. J. Construction of HSC-2DPAGE: a two-dimensional gel electrophoresis database of heart proteins. *Electrophoresis* **1997**; 18(3-4):471-479.
- Ewing, B., and Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **1998**; 8(3):186-94.
- Ewing, B., and Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics* **2000**; 25(2):232-234.
- Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(17):9721-9726.
- Fickett, J. W., and Hatzigeorgiou, A. G. Eukaryotic promoter recognition. *Genome Research* **1997**; 7(9):861-878.
- Fitch, W. M. Distinguishing homologous from analogous proteins. *Systematic Zoology* **1970**; 19(2):99-113.
- Fountoulakis, M., Schuller, E., Hardmeier, R., Berndt, P., and Lubec, G. Rat brain proteins: two-dimensional protein database and variations in the expression level. *Electrophoresis* **1999**; 20(18):3572-3579.
- Fountoulakis, M., Berndt, P., Boelsterli, U. A., Crameri, F., Winter, M., Albertini, S., and Suter, L. Two-dimensional database of mouse liver proteins: changes in hepatic protein levels following treatment with acetaminophen or its nontoxic regioisomer 3-acetamidophenol. *Electrophoresis* **2000**; 21(11):2148-2161.
- Fountoulakis, M., Juranville, J. F., Berndt, P., Langen, H., and Suter, L. Two-dimensional database of mouse liver proteins. An update. *Electrophoresis* **2001**; 22(9):1747-1763.
- Frech, K., Herrmann, G., and Werner, T. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Research* **1993**; 21(7):3117-3118.
- Frech, K., Dietze, P., and Werner, T. ConsInspector 3.0: new library and enhanced functionality. *Computer Applications in Biosciences* **1997**; 13:109-110.
- Frech, K., Quandt, K., and Werner, T. Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biology* **1998**; 1(1):29-38.
- Frith, M. C., Hansen, U., and Weng, Z. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* **2001**; 17(10): 878-889.
- Gardiner-Garden, M., and Littlejohn, T. G. A comparison of microarray databases. *Briefings in Bioinformatics* **2001**; 2(2):143-158.
- Garrels, J. I., and Franza, B. R., Jr. Transformation-sensitive and growth-related changes of protein synthesis in REF52 cells. A two-dimensional gel analysis of SV40-, adenovirus-, and Kirsten murine sarcoma virus-transformed rat cells using the REF52 protein database. *Journal of Biological Chemistry* **1989a**; 264(9):5299-5312.
- Garrels, J. I., and Franza, B. R., Jr. The REF52 protein database. Methods of database construction and analysis using the QUEST system and characterizations of protein pat-

- terms from proliferating and quiescent REF52 cells. *Journal of Biological Chemistry* **1989b**; 264(9):5283-5298.
- Garrels, J. I., Franza, B. R., Chang, C., and Latter, G. Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* **1990**; 11(12):1114-1130.
- Garrels, J. I., Futcher, B., Kobayashi, R., Latter, G. I., Schwender, B., Volpe, T., Warner, J. R., and McLaughlin, C. S. Protein identifications for a *Saccharomyces cerevisiae* protein database. *Electrophoresis* **1994**; 15(11):1466-1486.
- Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **2000**; 11(12):4241-4257.
- Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J., and Brown, P. O. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell* **2001**; 12(10):2987-3003.
- Gasteiger, E., Jung, E., and Bairoch, B. SWISS-PROT: Connecting biological Knowledge via a protein database. *Current Issues in Molecular Biology* **2001**; 3(3):47-55.
- Gawantka, V., Pollet, N., Delius, H., Vingron, M., Pfister, R., Nitsch, R., Blumenstock, C., and Niehrs, C. Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mechanisms of Development* **1998**; 77(2):95-141.
- Gelfand, M. S., Dubchak, I., Dralyuk, I., and Zorn, M. ASDB: database of alternatively spliced genes. *Nucleic Acids Research* **1999**; 27(1):301-302.
- Ghosh, D. New developments of a transcription factors database. *Trends in Biochemical Sciences* **1991**; 16(11):445-447.
- Ghosh, D. TFD: the transcription factors database. *Nucleic Acids Research* **1992**; 20 (Supplement):2091-2093.
- Ghosh, D. Status of the transcription factors database (TFD). *Nucleic Acids Research* **1993**; 21(13):3117-3118.
- Ghosh, D. OOTFD (Object-Oriented Transcription Factors Database): an object-oriented successor to TFD. *Nucleic Acids Research* **1998**; 26(1):360-362.
- Ghosh, D. Object-oriented transcription factors database (ooTFD). *Nucleic Acids Research* **2000**; 28(1):308-310.
- Giometti, C. S., Taylor, J., and Tollaksen, S. L. Mouse liver protein database: a catalog of proteins detected by two-dimensional gel electrophoresis. *Electrophoresis* **1992**; 13(12):970-991.
- Giometti, C. S., Tollaksen, S. L., Chubb, C., Williams, C., and Huberman, E. Analysis of proteins from human breast epithelial cells using two-dimensional gel electrophoresis. *Electrophoresis* **1995**; 16(7):1215-1224.
- Giometti, C. S., Williams, K., and Tollaksen, S. L. A two-dimensional electrophoresis database of human breast epithelial cell proteins. *Electrophoresis* **1997**; 18(3-4):573-581.
- Gössling, E., Kel-Margoulis, O. V., Kel, A. E., and Wingender, E. MATCH – a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes. In: GCB 2001, Proceedings of the German Conference on Bioinformatics, October 7-10, 2001, GBF-Braunschweig, Braunschweig, Germany **2001**:158-160.
- Golub, T. R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**; 286(5439):531-538.
- Gordon, D., Abajian, C., and Green, P. Consed: a graphical tool for sequence finishing. *Genome Research* **1998**; 8(3):195-202.
- Grebenova, D., Halada, P., Stulik, J., Havlicek, V., and Hrkál, Z. Protein changes in HL60 leukemia cells associated with 5-aminolevulinic acid-based photodynamic therapy. Early effects on endoplasmic reticulum chaperones. *Photochemistry and Photobiology* **2000**; 72(1):16-22.
- Grillo, G., Attimonelli, M., Liuni, S., and Pesole, G. CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Computer Applications in Bioscience* **1996**; 12(1):1-8.
- Gruber, T. R. A translation approach to portable ontologies. *Knowledge Acquisition* **1993**; 5(2):199-220; <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html#1>.

- Grünenfelder, B., Rummel, G., Vohradsky, J., Röder, D., Langen, H., and Jenal, U. Proteomic analysis of the bacterial cell cycle. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(8): 4681-4686.
- Grunau, C., Schattevoy, R., Maché, N. and Rosenthal, A. MethTools – a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Research* **2000**; 28(5): 1053-1058.
- Grunau, C., Renault, E., Rosenthal, A., and Roizes, G. MethDB – a public database for DNA methylation data. *Nucleic Acids Research* **2001**; 29(1):270-274.
- GuhaThakurta, D., and Stormo, G. D. Identifying target sites for cooperatively binding factors. *Bioinformatics* **2001**; 17(7):608-621.
- Haas, S., Beissbarth, T., Rivals, E., Krause, A., and Vingron, M. GeneNest: automated generation and visualization of gene indices. *Trends in Genetics* **2000**; 16(11):521-523.
- Hanash, S. M., and Teichroew, D. Mining the human proteome: experience with the human lymphoid protein database. *Electrophoresis* **1998**; 19(11):2004-2009.
- Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J., and Bork, P. Alternative splicing of human genes: more the rule than the exception? *Trends in Genetics* **1999**; 15(10):389-390.
- Hardison, R. C., Oeltjen, J., and Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Research* **1997**; 7(10):959-966.
- Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., and Brown, P. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **2000**; 1(2):research0003.1-0003.21.
- Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. Supervised harvesting of expression trees. *Genome Biology* **2001**; 2(1):research0003.1-research0003.12.
- Hawkins, V., Doll, D., Bumgarner, R., Smith, T., Abajian, C., Hood, L., and Nelson, P. S. PEDB: the Prostate Expression Database. *Nucleic Acids Research* **1999**; 27(1):204-208.
- Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, M. R., Manzoni, C., Vignati, M., and Gianazza, E. Proteins of rat serum: I. Establishing a reference two-dimensional electrophoresis map by immunodetection and microbore high performance liquid chromatography-electrospray mass spectrometry. *Electrophoresis* **1998**; 19(8-9): 1484-1492.
- Hecksher-Sorensen, J., and Sharpe, J. 3D confocal reconstruction of gene expression in mouse. *Mechanisms of Development* **2001**; 100(1):59-63.
- Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Research* **1999**; 27(1):318-322.
- Henikoff, J. G., and Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in Biosciences* **1996**; 12(2):135-143.
- Hernychova, L., Stulik, J., Halada, P., Macela, A., Kroca, M., Johansson, T., and Malina, M. Construction of a *Francisella tularensis* two-dimensional electrophoresis protein database. *Proteomics* **2001**; 1(4):508-515.
- Herwig, R., Poustka, A., Müller, C., Bull, C., Lehrach, H., and O'Brien, J. Large-scale clustering of genetic fingerprinting data. *Genome Research* **1999**; 9(11):1093-1105.
- Hodges, P. E., McKee, A. H. Z., Davis, B. P., Payne, W. E., and Garrels, J. I. Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Research* **1999**; 27(1):69-73.
- Hodges, P. E., Carrico, P. M., Hogan, J. D., O'Neill, K. E., Owen, J. J., Mangan, M., Davis, B. P., Brooks, J. E., and Garrels, J. I. Annotating the human proteome: the Human Proteome Survey Database (HumanPSD™) and an in-depth target database for G protein-coupled receptors (GPCR-PD™) from Incyte Genomics. *Nucleic Acids Research* **2002**; 30(1):137-141.
- Holm, L., and Sander, C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Research* **1999**; 27(1):244-247.
- Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **1998**; 95(5):717-728.
- Hoogland, C., Baujard, V., Sanchez, J. C., Hochstrasser, D. F., and Appel, R. D. Make2ddb: a simple package to set up a two-dimensional electrophoresis database for the

- World Wide Web. *Electrophoresis* **1997**; 18(15):2755-2758.
- Hoogland, C., Sanchez, J.-C., Tonella, L., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. Current status of the SWISS-2DPAGE database. *Nucleic Acids Research* **1998**; 26(1):332-333.
- Hoogland, C., Sanchez, J.-C., Tonella, L., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. The SWISS-2DPAGE database: what has changed during the last year. *Nucleic Acids Research* **1999a**; 27(1):289-291.
- Hoogland, C., Sanchez, J. C., Walther, D., Baujard, V., Baujard, O., Tonella, L., Hochstrasser, D. F., and Appel, R. D. Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* **1999b**; 20(18):3568-3571.
- Hoogland, C., Sanchez, J.-C., Tonella, L., Binz, P.-A., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Research* **2000**; 28(1):286-288.
- Hope, I. A., Albertson, D. G., Martinelli, S. D., Lynch, A. S., Sonnhhammer, E., and Durbin, R. The *C. elegans* expression pattern database: a beginning. *Trends in Genetics* **1996**; 12(9):370-371.
- Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B., and Auffray, C. The Genexpress Index: a resource for gene discovery and the genetic map of the human genome. *Genome Research* **1995**; 5(3):272-304.
- Hu, G. K., Madore, S. J., Moldover, B., Jatke, T., Balaban, D., Thomas, J., and Wang, Y. Predicting splice variant from DNA chip expression data. *Genome Research* **2001**; 11(7):1237-1245.
- Huang, X., and Madan, A. CAP3: A DNA sequence assembly program. *Genome Research* **1999**; 9(9):868-77.
- Hubbard, M. J., Faught, M. J., Carlisle, B. H., and Stockwell, P. A. ToothPrint, a proteomic database for dental tissues. *Proteomics* **2001**; 1(1):132-135.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraborty, K., Simon, J., Bard, M., and Friend, S. H. Functional discovery via a compendium of expression profiles. *Cell* **2000**; 102(1):109-126.
- Hutchinson, G. B. The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Computer Applications in Biosciences* **1996**; 12(5):391-398.
- Iseli, C., Jongeneel, C. V., and Bucher, P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* **1999**: 138-148.
- James, P. Of genomes and proteomes. *Biochemical and Biophysical Research Communications* **1997**; 231(1):1-6.
- Jennings, E. G., and Young, R. A. Genome expression on the World Wide Web. *Trends in Genetics* **1999**; 15(5):202-204.
- Ji, H., Whitehead, R. H., Reid, G. E., Moritz, R. L., Ward, L. D., and Simpson, R. J. Two-dimensional electrophoretic analysis of proteins expressed by normal and cancerous human crypts: application of mass spectrometry to peptide-mass fingerprinting. *Electrophoresis* **1994**; 15(3-4):391-405.
- Ji, H., Reid, G. E., Moritz, R. L., Eddes, J. S., Burgess, A. W., and Simpson, R. J. A two-dimensional gel database of human colon carcinoma proteins. *Electrophoresis* **1997**; 18(3-4):605-613.
- Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X., and Li, Y. AsMamDB: an alternative splicing database of mammals. *Nucleic Acids Research* **2001**; 29(1):260-263.
- Jain, A. K., and Dubes, R. C. Algorithms for clustering data. Prentice-Hall, Englewood Cliffs, New Jersey, USA **1988**.
- Jefferies, J. R., Brophy, P. M., and Barrett, J. Investigation of *Fasciola hepatica* sample preparation for two-dimensional electrophoresis. *Electrophoresis* **2000**; 21(17):3724-3729.
- Jost, J. P. and Saluz, H. P. In: DNA Methylation: molecular biology and biological significance. Birkhäuser Verlag, Basel, Switzerland **1993**.
- Jung, E., Hoogland, C., Chiappe, D., Sanchez, J. C., and Hochstrasser, D. F. The establishment of a human liver nuclei two-dimensional electrophoresis reference map. *Electrophoresis* **2000**; 21(16):3483-3487.
- Jungblut, P., Otto, A., Regitz, V., Fleck, E., and Wittmann-Liebold, B. Identification of human myocard proteins separated by two-

- dimensional electrophoresis. *Electrophoresis* **1992**; 13(9-10):739-741.
- Jungblut, P., Otto, A., Zeindl-Eberhart, E., Pleissner, K.-P., Knecht, M., Regitz-Zagrosek, V., Fleck, E., and Wittmann-Liebold, B. Protein composition of the human heart: The construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis* **1994**; 15(5):685-707.
- Jungblut, P. R., Zimny-Arndt, U., Zeindl-Eberhart, E., Stulik, J., Koupilova, K., Pleissner, K. P., Otto, A., Muller, E. C., Sokolowska-Kohler, W., Grabher, G., and Stoffler, G. Proteomics in human disease: cancer, heart and infectious diseases. *Electrophoresis* **1999a**; 20(10):2100-2110.
- Jungblut, P. R., Grabher, G., and Stoffler, G. Comprehensive detection of immunorelevant *Borrelia garinii* antigens by two-dimensional electrophoresis. *Electrophoresis* **1999b**; 20(18):3611-3622.
- Jungblut, P. R., Schaible, U. E., Mollenkopf, H. J., Zimny-Arndt, U., Raupach, B., Mattow, J., Halada, P., Lamer, S., Hagens, K., Kaufmann, S. H. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Molecular Microbiology* **1999**; 33(6):1103-1117.
- Jungblut, P. R., Bumann, D., Haas, G., Zimny-Arndt, U., Holland, P., Lamer, S., Siejak, F., Aebischer, A., and Meyer, T. F. Comparative proteome analysis of *Helicobacter pylori*. *Molecular Microbiology* **2000**; 36(3):710-725.
- Junker, V., Contrino, S., Fleischmann, W., Hermjakob, H., Lang, F., Magrane, M., Martin, M. J., Mitalitonna, N., O'Donovan, C., Apweiler, R. The role SWISS-PROT and TrEMBL play in the genome research environment. *Journal of Biotechnology* **2000**; 78(3):221-234.
- Kan, Z., Rouchka, E. C., Gish, W. R., and States, D. J. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research* **2001**; 11(5):889-900.
- Karlin, S., and Macken, C. Assessment of inhomogeneities in an *E. coli* physical map. *Nucleic Acids Research* **1991**; 19(15):4241-4246.
- Karplus, K., Barrett, C., and Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **1998**; 14(10):846-856.
- Kaufman, M. H. The atlas of mouse development. Academic Press, London, United Kingdom **1999**.
- Kawamoto, S., Ohnishi, T., Kita, H., Chisaka, O., and Okubo, K. Expression profiling by iAFLP: A PCR-based method for genome-wide gene expression profiling. *Genome Research* **1999**; 9(12):1305-1312.
- Kawamoto, S., Yoshii, J., Mizuno, K., Ito, K., Miyamoto, Y., Ohnishi, T., Matoba, R., Hori, N., Matsumoto, Y., Okumura, T., Nakao, Y., Yoshii, H., Arimoto, J., Ohashi, H., Nakanishi, H., Ohno, I., Hashimoto, J., Shimizu, K., Maeda, K., Kuriyama, H., Nishida, K., Shimizu-Matsumoto, A., Adachi, W., Ito, R., Kawasaki, S., and Chae, K. S. BodyMap: a collection of 3' ESTs for analysis of human gene expression information. *Genome Research* **2000**; 10(11):1817-1827.
- Kawashima, T., Kawashima, S., Kanehisa, M., Nishida, H., and Makabe, K. W. MAGEST: Maboya gene expression patterns and sequence tags. *Nucleic Acids Research* **2000**; 28(1):133-135.
- Kawashima, T., Kawashima, S., Kohara, Y., Kanehisa, M., and Makabe, K. W. Update of MAGEST: Maboya gene expression patterns and sequence tags. *Nucleic Acids Research* **2002**; 30(1):119-120.
- Kel, A. E., Kolchanov, N. A., Kapitonov, V. V., Ponomarenko, M. P., Likhachev, E. A., Lim, H. A., and Milanesi, L. Computer analysis and recognition of functional sites on the base of oligonucleotide patterns distributions. In: Proceedings of the 2nd International Conference on Electrophoresis, Supercomputing and the Human Genome, June 1992, St. Petersburg, Florida, USA. Cantor, C. R., and Lim H. A., editors. Book News, Inc., Portland, Oregon, USA **1992**:521-544.
- Kel, A. E., Ponomarenko, M. P., Likhachev, E. A., Orlov, Yu., L., Ischenko, I. V., Milanesi, L., and Kolchanov, N. A. SITEVIDEO: A computer system for functional site analysis and recognition. Investigation of the human splice sites. *Computer Applications in Bio-sciences* **1993**; 9(6):617-627.
- Kel, A., Kondrakhin, Yu., Kolpakov, Ph., Ponomarenko, M., Wingender, E., and Kolchanov, N. Computer analysis of the structure of transcription factor binding sites. *SAMS* **1995a**; 18-19:819-822.
- Kel, A. E., Kondrakhin, Y. V., Kolpakov, Ph. A., Kel, O. V., Romashenko, A. G., Wingender, E.,

- Milanesi, L., and Kolchanov, N. A. Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. In: Proceedings of the 3rd International Conference of Intelligent Systems for Molecular Biology. Rawlings, C. J., editor. Menlo Park, California, USA **1995b**; 3:197-205.
- Kel, O. V., Romashenko, A. G., Kel, A. E., Wingender, E., and Kolchanov, N. A. A compilation of composite regulatory elements affecting transcription in vertebrates. *Nucleic Acids Research* **1995c**; 23(20):4097-4103.
- Kel, A., Kel-Margoulis, O., Babenko, V., and Wingender, E. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *Journal of Molecular Biology* **1999**; 288(3):353-376.
- Kel, A. E., Kel-Margoulis, O. V., Farnham, P. J., Bartley, S. M., Wingender, E., and Zhang, M. Q. Computer-assisted identification of cell-cycle related genes: New targets for E2F transcription factors. *Journal of Molecular Biology* **2001a**; 309(1):99-120.
- Kel, A., Kel-Margoulis, O., Ivanova, T., and Wingender, E. ClusterScan: A tool for automatic annotation of genomic regulatory sequences by searching for composite clusters. In: GCB 2001, Proceedings of the German Conference on Bioinformatics, Braunschweig, Germany, October 7-10, 2001. Wingender, E., Hofestädt, R., and Liebich, I., editors. GBF-Braunschweig, Braunschweig, Germany **2001b**:96-101.
- Kel-Margoulis, O. V., Kel, A. E., Reuter, I., Deineko, I. V., and Wingender, E. TRANS-Compel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Research* **2001a**; 30(1):332-334.
- Kel-Margoulis, O. Automatic annotation of the regulatory regions of cell cycle related genes on human chromosomes. In: Abstracts of papers presented at the 2000 Meeting on Genome Sequencing & Biology, May 10 – May 14, 2000. Boguski, M., editor. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA **2001b**.
- Kelley, S. Getting started with Acedb. *Briefings in Bioinformatics* **2000**; 1(2):131-137.
- Knecht, M., Regitz-Zagrosek, V., Pleissner, K.-P., Jungblut, P., Steffen, C., Hildebrandt, A., and Fleck, E. Characterization of myocardial protein composition in dilated cardiomyopathy by two-dimensional gel electrophoresis. *European Heart Journal* **1994a**; 15(Supplement D):37-44.
- Knecht, M., Regitz-Zagrosek, V., Pleissner, K.-P., Emig, S., Jungblut, P., Hildebrandt, A., and Fleck, E. Dilated cardiomyopathy: Computer-assisted analysis of endomyocardial biopsy protein patterns by two-dimensional gel electrophoresis. *European Journal of Clinical Chemistry and Clinical Biochemistry* **1994**; 32(8):615-624.
- Knudsen, S. Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **1999**; 15(5):356-361.
- Kohonen, T. Self-organizing maps. Springer-Verlag, Berlin, Germany **1997**.
- Kolchanov, N. A., Rogozin, I. B., Kel, A. E., Ponomarenko, M. P., Lihachov, J., and Milanesi, L. Computer analysis and recognition of human splice sites. In: Proceedings of Genome Informatics Workshop II. Kanehisa, M., editor. Universal Academic Press, Tokyo, Japan **1991**:104-107.
- Kolchanov, N. A., Ananko, E. A., Podkolodnaya, O. A., Ignatieva, E. V., Stepanenko, I. L., Kel-Margoulis, O. V., Kel, A. E., Merkulova, T. I., Goryachkovskaya, T. N., Busygina, T. V., Kolkpakov, F. A., Podkolodny, N. L., Naumochkin, A. N., and Romashchenko, A. G. Transcription Regulatory Region Database (TRRD): its status in 1999. *Nucleic Acids Research* **1999**; 27(1):303-306.
- Kondrakhin, Yu. V., Shamin, V. V., and Kolchanov, N. A. Construction of a generalized consensus matrix for recognition of vertebrate pre-mRNA 3'-terminal processing sites. *Computer Applications in Biosciences* **1994**; 10(6):597-603.
- Kondrakhin, Yu. V., Kel, A. E., Kolchanov, N. A., Romashchenko, A. G., and Milanesi, L. Eukaryotic promoter recognition by binding sites for transcription factors. *Computer Applications in Biosciences* **1995**; 11(5):477-488.
- Kovalyov, L. I., Shishkin, S. S., Efimochkin, A. S., Kovalyova, M. A., Ershova, E. S., Egorov, T. A., and Musalyamov, A. K. The major protein expression profile and two-dimensional protein database of human heart. *Electrophoresis* **1995**; 16(7):1160-1169.
- Kovarova, H., Stulik, J., Hochstrasser, D. F., Bures, J., Melichar, B., and Jandik, P. Two-dimensional electrophoretic study of normal colon mucosa and colorectal cancer. *Applied*

- and *Theoretical Electrophoresis* **1994**; 4(3): 103-106.
- Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M., and Apweiler, R. CluStr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Research* **2001a**; 29(1):33-36.
- Kriventseva, E. V., Biswas, M., and Apweiler, R. Clustering and analysis of protein families. *Current Opinion in Structural Biology* **2001b**; 11(3):334-339.
- Krizman, D. B., Wagner, L., Lash, A., Strausberg, R. L., and Emmert-Buck, M. R. The Cancer Genome Anatomy Project: EST sequencing and the genetics of cancer progression. *Neoplasia* **1999**; 1(2):101-106.
- Krufft, V., Eubel, H., Jansch, L., Werhahn, W., and Braun, H. P. Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. *Plant Physiology* **2001**; 127(4):1694-1710.
- Kumar, A., Cheung, K.-H., Ross-Macdonald, P., Coelho, P. S. R., Miller, P., and Snyder, M. TRIPLES: a database of gene function in *Saccharomyces cerevisiae*. *Nucleic Acids Research* **2000a**; 28(1):81-84.
- Kumar, A., des Etages, S. A., Coelho, P. S. R., Roeder, G. S., and Snyder, M. High-throughput methods for the large-scale analysis of gene function by transposon tagging. *Methods in Enzymology* **2000b**; 328:550-574.
- Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon R., E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. A public database for gene expression in human cancers. *Cancer Research* **1999**; 59(21):5403-5407.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Le Vine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaanty, K. D., Miner, T. L., Delehaanty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Szek, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Peltier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubinfeld, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hockamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R.,

- Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., and Szustakowski, J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**; 409(6822):860-921.
- Lang, F. SWISS-PROT + TrEMBL. *Trends in Genetics* **1997**; 13(10):417.
- Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. SAGEmap: A public gene expression resource. *Genome Research* **2000**; 10(7):1051-1060.
- Latter, G. I., Boutell, T., Monardo, P. J., Kobayashi, R., Futcher, B., McLaughlin, C. S., and Garrels, J. I. A *Saccharomyces cerevisiae* Internet protein resource now available. *Electrophoresis* **1995**; 16(7):1170-1174.
- Le Crom, S., Devaux, F., Jacq, C., and Marc, P. yMGV: helping biologists with yeast microarray data mining. *Nucleic Acids Research* **2002**; 30(1):76-79.
- Leffers, H., Dejgaard, K., Honore, B., Madsen, P., Nielsen, M. S., and Celis, J. E. cDNA expression and human two-dimensional gel protein databases: towards integrating DNA and protein information. *Electrophoresis* **1996**; 17(11):1713-1719.
- Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M. B. The I. M. A. G. E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **1996**; 33(1):151-152.
- Liberatori, S., Bini, L., De Felice, C., Magi, B., Marzocchi, B., Raggiaschi, R., Frutiger, S., Sanchez, J. C., Wilkins, M. R., Hughes, G., Hochstrasser, D. F., Bracci, R., and Pallini, V. A two-dimensional protein map of human amniotic fluid at 17 weeks' gestation. *Electrophoresis* **1997**; 18(15):2816-2822.
- Louie, A. Y., Hüber, M. M., Ahrens, T. A., Rothbacher, U., Moats, R., Jacobs, R. E., Fraser, S. E., and Meade, T. J. *In vivo* visualization of gene expression using magnetic resonance imaging. *Nature Biotechnology* **2000**; 18(3):321-325.
- Lyons, T. J., Gasch, A. P., Gaither, L. A., Botstein, D., Brown, P. O., and Eide, D. J. Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proceedings of the National Academy of Sciences of the USA* **2000**; 97(14):7957-7962.
- Maglott, D. R., and Nierman, W. C. Clone and genomic repositories at the American Type Culture Collection. *Genomics* **1990**; 8(3):601-605.
- Makabe, K. W., Kawashima, T., Kawashima, S., Minokawa, T., Adachi, A., Kawamura, H., Ishikawa, H., Yasuda, R., Yamamoto, H., Kondoh, K., Arioka, S., Sasakura, Y., Kobayashi, A., Yagi, K., Shojima, K., Kondoh, Y., Kido, S., Tsujinami, M., Nishimura, N., Takahashi, M., Nakamura, T., Kanehisa, M., Ogasawara, M., Nishikata, T., and Nishida, H. Large-scale cDNA analysis of the maternal genetic information in the egg of *Halocynthia roretzi* for a gene expression catalog of ascidian development. *Development* **2001**; 128(13):2555-2567.
- Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I., and Koonin, E. V. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Research* **1999**; 9(7):608-628.
- Manabe, T., Yamamoto, H., and Kawai, M. Studies on the procedure for the construction of cellular protein databases employing micro two-dimensional electrophoresis: an HL-60 protein database. *Electrophoresis* **1995**; 16(3):407-422.
- Marc, P., Devaux, F., and Jacq, C. yMGV: a database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Research* **2001**; 29(13):E63.
- Martinelli, S. D., Brown, C. G., and Durbin, R. Gene expression and development databases for *C. elegans*. *Seminars in Cell and Developmental Biology* **1997**; 8(5):459-467.
- Mathesius, U., Keijzers, G., Natera, S. H., Weinman, J. J., Djordjevic, M. A., and Rolfe, B. G. Establishment of a root proteome reference map for the model legume *Medicago truncatula* using the expressed sequence tag database for peptide mass fingerprinting. *Proteomics* **2001**; 1(11):1424-1440.
- Matloubian, M., David, A., Engel, S., Ryan, J. E., and Cyster, J. G. A transmembrane CXC chemokine is a ligand for HIV coreceptor Bonzo. *Nature Immunology* **2000**; 1(4):298-304.
- Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., and Gianazza, E. Proteins of rat serum: II. Influence of some biological

- parameters of the two-dimensional electrophoresis pattern. *Electrophoresis* **1998**; *19*(8-9):1493-1500.
- Miller, I., Haynes, P., Eberini, I., Gemeiner, M., Aebersold, R., and Gianazza, E. Proteins of rat serum: III. Gender-related differences in protein concentration under baseline conditions and upon experimental inflammation as evaluated by two-dimensional electrophoresis. *Electrophoresis* **1999**; *20*(4-5):836-845.
- Mironov, A. A., Fickett, J. W. and Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Research* **1999**; *9*(12):1288-1293.
- Mills, J. C., and Gordon, J. I. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Research* **2001**; *29*(15):E72.
- Minsky, M., and Papert, S. Perceptrones: An introduction to computational geometry. MIT Press, Cambridge, Massachusetts, USA **1969**.
- Mirkin, B. Mathematical classification and clustering. Kluwer Academic Publishers, Dordrecht, The Netherlands **1996**.
- Möller, S., Leser, U., Fleischmann, W., and Apweiler, R. EDITtoTREMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics* **1999**; *15*(3):219-227.
- Mollenkopf, H. J., Jungblut, P. R., Raupach, B., Mattow, J., Lamer, S., Zimny-Arndt, U., Schaible, U. E., and Kaufmann, S. H. A dynamic two-dimensional polyacrylamide gel electrophoresis database: the mycobacterial proteome via internet. *Electrophoresis* **1999**; *20*(11):2172-2180.
- Momjian, B. PostgreSQL: Introduction and Concepts, Addison Wesley Publishing Company, Boston, Massachusetts, USA **2000**.
- Müller, E.-C., Thiede, B., Zimny-Arndt, U., Scheler, C., Prehm, J., Müller-Werdan, U., Wittmann-Liebold, B., Otto, A., and Jungblut, P. High-performance human myocardial two-dimensional electrophoresis database: Edition 1996. *Electrophoresis* **1996**, *17*(11): 1700-1712.
- Munasinghe, A., Patankar, S., Cook, B. P., Madden, S. L., Martin, R. K., Kyle, D. E., Shoaibi, A., Cummings, L. M., and Wirth, D. F. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. *Molecular and Biochemical Parasitology* **2001**; *113*(1):23-24.
- Myers, T. G., Anderson, N. L., Waltham, M., Li, G., Buolamwini, J. K., Scudiero, D. A., Paull, K. D., Sausville, E. A., and Weinstein, J. N. A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **1997**; *18*(3-4):647-653.
- Nature Genetics supplement 'The Chipping Forecast'. *Nature Genetics* **1999**; *21*(Supplement 1):1-60.
- Neidhardt, F. C. Report of workshop on cellular protein databases derived from two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **1989**; *10*(2):73-75.
- Nelson, P. S., Clegg, N., Eroglu, B., Hawkins, V., Bumgarner, R., Smith, T., and Hood, L. The prostate expression database (PEDB): status and enhancements in 2000. *Nucleic Acids Research* **2000**; *28*(1):212-213.
- Nelson, P. S., Pritchard, C., Abbott, D., and Clegg, N. The human (PEDB) and mouse (mPEDB) Prostate Expression Databases. *Nucleic Acids Research* **2002**; *30*(1):218-220.
- Nieminen, P., Pekkanen, M., Aberg, T., and Thesleff, I. A graphical WWW-database on gene expression in tooth. *European Journal of Oral Science* **1998**; *106*(Supplement 1):7-11.
- Nishida, H. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue-restricted stage. *Developmental Biology* **1987**; *121*(2):526-541.
- Novina, C. D., and Roy, A. L. Core promoters and transcriptional control. *Trends in Genetics* **1996**; *12*(9):351-355.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **1996**; *24*(1):1-252.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **1997**; *25*(1):1-282.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **1998**; *26*(1):1-389.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **1999**; *27*(1):1-379.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **2000**; *28*(1):1-382.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **2001**; *29*(1):1-349.
- Nucleic Acids Research Database issue. *Nucleic Acids Research* **2002**; *30*(1):1-417.
- O'Donovan, C., Martin, M. J., Glemet, E., Codani, J. J., and Apweiler, R. Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics* **1999**; *15*(3):258-259.
- O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E., Bairoch, B., and Apweiler, R. High quality protein knowledge resource –

- SWISS-PROT and TrEMBL. *Briefings in Bioinformatics* **2002**; 3(3): in press.
- Ogawa, N., DeRisi, J., and Brown, P. O. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Molecular Biology of the Cell* **2000**; 11(12):4309-4321.
- Oh, J. M., Hanash, S. M., and Teichrow, D. Mining protein data from two-dimensional gels: tools for systematic post-planned analyses. *Electrophoresis* **1999**; 20(4-5):766-774.
- Ohler, U., Niemann, H., Liao, G., and Rubin, G. M. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* **2001**; 17(Supplement 1):199-206.
- Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics* **1992**; 2(3): 173-179.
- Otto, A., Benndorf, R., Wittmann-Liebold, B., and Jungblut, P. Identification of proteins on two-dimensional gels for the construction of a human heart 2-DE database. *Journal of Protein Chemistry* **1994**; 13(6):478-480.
- Otto, A., Thiede, B., Müller, E.-C., Scheler, C., Wittmann-Liebold, B., and Jungblut, P. Identification of human myocardial proteins separated by two-dimensional electrophoresis using an effective sample preparation for mass spectrometry *Electrophoresis* **1996**; 17(10):1643-1650.
- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research* **2001**; 29(1):159-164.
- Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research* **2000**; 28(1):141-145.
- Pardo, M., Monteoliva, L., Pla, J., Sanchez, M., Gil, C., and Nombela, C. Two-dimensional analysis of proteins secreted by *Saccharomyces cerevisiae* regenerating protoplasts: a novel approach to study the cell wall. *Yeast* **1999**; 15(6):459-472.
- Pardo, M., Ward, M., Pitarch, A., Sanchez, M., Nombela, C., Blackstock, W., and Gil, C. Cross-species identification of novel *Candida albicans* immunogenic proteins by combination of two-dimensional polyacrylamide gel electrophoresis and mass spectrometry. *Electrophoresis* **2000a**; 21(13):2651-2659.
- Pardo, M., Ward, M., Bains, S., Molina, M., Blackstock, W., Gil, C., and Nombela, C. A proteomic approach for the study of *Saccharomyces cerevisiae* cell wall biogenesis. *Electrophoresis* **2000b**; 21(16):3396-3410.
- Pasquali, C., Frutiger, S., Wilkins, M. R., Hughes, G. J., Appel, R. D., Bairoch, A., Schaller, D., Sanchez, J. C., and Hochstrasser, D. F. Two-dimensional gel electrophoresis of *Escherichia coli* homogenates: the *Escherichia coli* SWISS-2DPAGE database. *Electrophoresis* **1996**; 17(3):547-555.
- Patankar, S., Munasinghe, A., Shoaibi, A., Cummings, L. M., and Wirth, D. F. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Molecular Biology of the Cell* **2001**; 12(10):3114-3125.
- Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology* **1990**; 183:63-98.
- Pearson, W. R. Using the FASTA program to search protein and DNA sequence databases. *Methods in Molecular Biology* **1994a**; 25: 365-389.
- Pearson, W. R. Using the FASTA program to search protein and DNA sequence databases. *Methods in Molecular Biology* **1994b**; 24: 307-331.
- Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **2000**; 132:185-219.
- Perier, R. C., Junier, Th., and Bucher, P. The Eukaryotic Promoter Database EPD. *Nucleic Acids Research* **1998**; 26(1):353-357.
- Perier, R. C., Junier, T., Bonnard, C., and Bucher, P. The Eukaryotic Promoter Database (EPD): recent developments. *Nucleic Acids Research* **1999**; 27(1):307-309.
- Perier, R. C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. The eukaryotic promoter database (EPD). *Nucleic Acids Research* **2000**; 28(1):302-303.
- Perrot, M., Sagliocco, F., Mini, T., Monribot, C., Schneider, U., Shevchenko, A., Mann, M., Jenö, P., and Boucherie, H. Two-dimensional

- gel protein database of *Saccharomyces cerevisiae* (update 1999). *Electrophoresis* **1999**; 20(11):2280-2298.
- Pesole, G., Liuni, S., Grillo, G., and Saccone, C. Structural and compositional features of untranslated regions of eukaryotic mRNAs. *Gene* **1997**; 205(1-2):95-102.
- Pesole, G., Bernardi, G., and Saccone, C. Isochore specificity of AUG initiator context of human genes. *FEBS Letters* **1999**; 464(1-2):60-62.
- Pesole, G., and Liuni, S. Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. *Trends in Genetics* **1999**; 15(9):378.
- Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Larizza, A., Makalowski, W., and Saccone, C. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research* **2000a**; 28(1):193-196.
- Pesole, G., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., and Saccone, C. Analysis of oligonucleotide AUG start codon context in eukaryotic mRNAs. *Gene* **2000b**; 261(1):85-91.
- Pesole, G., Grillo, G., Larizza, A., and Liuni, S. The untranslated regions of eukaryotic mRNAs: structure, function and bioinformatic tools for their analysis. *Briefings in Bioinformatics* **2000c**; 1(3):236-249.
- Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., and Liuni, S. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*; **2001**; 276(1-2):73-81.
- Piétu, G., Mariage-Samson, R., Fayein, N. A., Matingou, C., Eveno, E., Houkgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M. D., Wirkner, U., Ansorge, W., Cox, D., Nagase, T., Nomura, N., and Auffray, C. The Genexpress IMAGE Knowledge Base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Research* **1999a**; 9(2):195-209.
- Piétu, G., Eveno, E., Soury-Segurens, B., Fayein, N. A., Mariage-Samson, R., Matingou, C., Leroy, E., Dechesne, C., Krieger, S., Ansorge, W., Reguigne-Arnould, I., Cox, D., Dehejia, A., Polymeropoulos, M. H., Devignes, M. D., and Auffray, C. The Genexpress IMAGE Knowledge Base of the human muscle transcriptome: A resource of structural, functional, and positional candidate genes for muscle physiology and pathologies. *Genome Research* **1999b**; 9(12):1313-1320.
- Pitarch, A., Pardo, M., Jimenez, A., Pla, J., Gil, C., Sanchez, M., and Nombela, C. Two-dimensional gel electrophoresis as analytical tool for identifying *Candida albicans* immunogenic proteins. *Electrophoresis* **1999**; 20(4-5):1001-1010.
- Pitarch, A., Diez-Orejias, R., Molero, G., Pardo, M., Sanchez, M., Gil, C., and Nombela, C. Analysis of the serologic response to systemic *Candida albicans* infection in a murine model. *Proteomics* **2001**; 1(4):550-559.
- Pleissner, K.-P., Sander, S., Oswald, H., Regitz-Zagrosek, V., and Fleck, E. The construction of the World Wide Web-accessible myocardial two-dimensional gel electrophoresis protein database "HEART-2DPAGE": A practical approach. *Electrophoresis* **1996**; 17(8):1386-1392.
- Pleissner, K. P., Sander, S., Oswald, H., Regitz-Zagrosek, V., and Fleck, E. Towards design and comparison of World Wide Web-accessible myocardial two-dimensional gel electrophoresis protein databases. *Electrophoresis* **1997**; 18(3-4):480-483.
- Pleissner, K.-P., Soeding, P., Sander, S., Oswald, H., Neuss, M., Regitz-Zagrosek, V., and Fleck, E. Dilated cardiomyopathy-associated proteins and their presentation in a WWW-accessible two-dimensional gel protein database. *Electrophoresis* **1997**; 18(5):802-808.
- Poirier, F., Imam, N., Pontet, M., Joubert-Caron, R., and Caron, M. The BPP (protein biochemistry and proteomics) two-dimensional electrophoresis database. *Journal of Chromatography. B, Biomedical Sciences and Applications* **2001a**; 753(1):23-28.
- Poirier, F., Pontet, M., Labas, V., le Caer, J. P., Sghiouar-Imam, N., Raphael, M., Caron, M., and Joubert-Caron, R. Two-dimensional database of a Burkitt lymphoma cell line (DG 75) proteins: protein pattern changes following treatment with 5'-azacytidine. *Electrophoresis* **2001b**; 22(9):1867-1877.
- Pollet, N., Schmidt, H. A., Gawantka, V., Vingron, M., and Niehrs, C. Axeldb: a *Xenopus laevis* database focusing on gene expression. *Nucleic Acids Research* **2000**; 28(1):139-140.
- Praz, V., Perier, R., Bonnard, C., and Bucher, P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Research* **2002**; 30(1):322-324.

- Pritchard, C. C., Hsu, L., Delrow, J., and Nelson, P. S. Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences of the USA* **2001**; 98(23): 13266-13271.
- Prestridge, D. S., and Stormo, G. SIGNAL SCAN 3.0: new database and program features. *Computer Applications in Biosciences* **1993**; 9(1):113-115.
- Prestridge, D. S. Predicting pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology* **1995**; 249(5):923-932.
- Pruitt, K. D., and Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* **2001**; 29(1):137-140.
- Ptashne, M., and Gann, A. Transcription activation by recruitment. *Nature* **1997**; 386(6625):569-577.
- Ptitsyn, A. A., Rogozin, I. B., Grigorovich, D. A., Strelets, V. B., Kel, A. E., Milanese, L., and Kolchanov, N. A. AutoGene: A computer system for nucleotide sequence analysis. *Molekuliarnaia Biologiia (Molecular Biology, Moscow)* **1996**; 30(2):258-264.
- Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. MatInd and MatInspector – New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* **1995**; 23(23):4878-4884.
- Rappsilber, J., and Mann, M. What does it mean to identify a protein in proteomics? *Trends in Biochemical Sciences* **2002**; 27(2): 74-78.
- Rasmussen, H. H., Orntoft, T. F., Wolf, H., and Celis, J. E. Towards a comprehensive database of proteins from the urine of patients with bladder cancer. *Journal of Urology* **1996**; 155(6):2113-2119.
- Rasmussen, R. K., Ji, H., Eddes, J. S., Zugaro, L. M., Reid, G. E., Simpson, R. J., and Dorow, D. S. Two-dimensional gel database of human breast carcinoma cell expressed proteins: an update. *Electrophoresis* **1998**; 19(5):818-825.
- Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics* **1997**; 13(4):163.
- Reese, M. G., and Eckman, F. H. Novel neural network prediction systems for human promoters and splice sites. In: *Proceedings of the Workshop on Gene-Finding and Gene Structure Prediction*, Pennsylvania, Philadelphia, USA. Searls, D., Fickett, J., Stormo, G., and Noordewier, M., editors. DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, Piscataway, New Jersey, USA **1995**.
- Reymond, M. A., Sanchez, J.-C., Hughes, G. J., Riese, J., Tortola, S., Peinado, M. A., Kirchner, T., Hohenberger, W., Hochstrasser, D. F., and Kockerling, F. Standardized characterization of gene expression in human colorectal epithelium by two-dimensional electrophoresis. *Electrophoresis* **1997**; 18(15):2842-2848.
- Riggins, G. J., and Strausberg, R. L. Genome and genetic resources from the Cancer Genome Anatomy Project. *Human Molecular Genetics* **2001**; 10(7):663-667.
- Ringwald, M., Baldock, R., Bard, J., Kaufman, M., Eppig, J. T., Richardson, J. E., Nadeau, J. H., and Davidson, D. A database for mouse development. *Science* **1994**; 265(5181): 2033-2034.
- Ringwald, M., Eppig, J. T., Begley, D. A., Corradi, J. P., McCright, I. J., Hayamizu, T. F., Hill, D. P., Kadin, J. A., and Richardson, J. E. The Mouse Gene Expression Database (GXD). *Nucleic Acids Research* **2001**; 29(1): 98-101.
- Rippe, K., Vonhippel, P. H., and Langowski, J. Action at a distance: DNA-looping and initiation of transcription. *Trends in Biochemical Sciences* **1995**; 20(12):500-506.
- Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C., and Friend, S. H. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **2000**; 287(5454):873-880.
- Rosenkrands, I., Weldingh, K., Jacobsen, S., Hansen, C. V., Florio, W., Gianetti, I., and Andersen, P. Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. *Electrophoresis* **2000a**; 21(5):935-948.
- Rosenkrands, I., King, A., Weldingh, K., Moniatte, M., Moertz, E., and Andersen, P. Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis* **2000**; 21(17): 3740-3756.
- Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey,

- S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalton, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **2000**; 24(3):227-235.
- Ross-MacDonald, P., Coelho, P. S. R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.-H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S., and Snyder, M. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **1999**; 402(6760):413-418.
- Roush, W. A womb with a view. *Science* **1997**; 278(5342):1397-1399.
- Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D. A. A *Drosophila* complementary DNA resource. *Science* **2000a**; 287(5461):2222-2224.
- Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vossshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., and Lewis, S. Comparative genomics of the eukaryotes. *Science* **2000b**; 287(5461):2204-2215.
- Sahnoun, I., Dehais, P., Van Montagu, M., Rossignol, M., and Rouze, P. PPMdb: a plant plasma membrane database. *Journal of Biotechnology* **2000**; 78(3):235-246.
- Sanchez, J.-C., Appel, R. D., Golaz, O., Pasquali, C., Ravier, F., Bairoch, A., and Hochstrasser, D. F. Inside SWISS-2DPAGE database. *Electrophoresis* **1995**; 16(7):1131-1151.
- Sanchez, J. C., Golaz, O., Frutiger, S., Schaller, D., Appel, R. D., Bairoch, A., Hughes, G. J., and Hochstrasser, D. F. The yeast SWISS-2DPAGE database. *Electrophoresis* **1996**; 17(3):556-565.
- Sanchez, J. C., Chiappe, D., Converset, V., Hoogland, C., Binz, P. A., Paesano, S., Appel, R. D., Wang, S., Sennitt, M., Nolan, A., Cawthorne, M. A., and Hochstrasser, D. F. The mouse SWISS-2DPAGE database: a tool for proteomics study of diabetes and obesity. *Proteomics* **2001**; 1(1):136-163.
- Santaren, J. F. Towards establishing a protein database of *Drosophila*. *Electrophoresis* **1990**; 11(3):254-267.
- Sarto, C., Marocchi, A., Sanchez, J.-C., Giannone, D., Frutiger, S., Golaz, O., Wilkins, M. R., Doro, G., Cappellano, F., Hughes, G. J., Hochstrasser, D. F., and Mocarelli, P. Renal cell carcinoma and normal kidney protein expression. *Electrophoresis* **1997**; 18(3-4):599-604.
- Sasik, R., Hwa, T., Iranfar, N., and Loomis, W. F. Percolation clustering: A novel algorithm applied to the clustering of gene expression patterns in *Dictyostelium* development. In: Pacific Symposium on Biocomputing 6. Altman, R., (ed.), World Scientific, Singapore, **2001**:335-347.
- Satoh, N. Developmental biology of ascidians. Cambridge University Press, Cambridge, United Kingdom **1994**.
- Scherf, M., Klingenhoff, A., and Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *Journal of Molecular Biology* **2000a**; 297(3):599-606.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* **2000b**; 24(3):236-244.
- Schulze-Kremer, S. Ontologies for molecular biology and bioinformatics. *In Silico Biology* **2002**; 2:0017.
- Seledtsov, I. A., Solovyev, V. V., and Merkulova, T. I. New elements of glucocorticoid-receptor binding sites of hormone-regulated genes. *Biochimica et Biophysica Acta* **1991**; 1089(3):367-376.
- Sese, J., Nikaidou, H., Kawamoto, S., Minesaki, Y., Morishita, S., and Okubo, K. BodyMap incorporated PCR-based expression profiling data and a gene ranking system. *Nucleic Acids Research* **2001**; 29(1):156-158.

- Shago, M., and Giguere, V. Isolation of a novel retinoic acid-responsive gene by selection of genomic fragments derived from CpG-island-enriched DNA. *Molecular and Cellular Biology* **1996**; 16(8):4337-4348.
- Sharan, R., and Shamir, R. CLICK: A clustering algorithm with applications to gene expression analysis. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB). Altman, R., editor. AAAI Press, Menlo Park, California, USA **2000**:307-316.
- Shaw, A. C., Røssel Larsen, M., Roepstorff, P., Holm, A., Christiansen, G., and Birkelund, S. Mapping and identification of HeLa cell proteins separated by immobilized pH-gradient two-dimensional gel electrophoresis and construction of a two-dimensional polyacrylamide gel electrophoresis database. *Electrophoresis* **1999a**; 20(4-5):977-983.
- Shaw, A. C., Røssel Larsen, M., Roepstorff, P., Justesen, J., Christiansen, G., and Birkelund, S. Mapping and identification of interferon gamma-regulated HeLa cell proteins separated by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis* **1999**; 20(4-5):984-993.
- Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., and Cherry, J. M. The Stanford Microarray Database. *Nucleic Acids Research* **2001**; 29(1):152-155.
- Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **2001**; 106(6):697-708.
- Simpson, R. J., Tsugita, A., Celis, J. E., Garrels, J. I., and Mewes, H. W. Workshop on two-dimensional gel protein databases. *Electrophoresis* **1992a**; 13(12):1055-1061.
- Simpson, R. J. Two-dimensional acrylamide gel electrophoresis inventory of cellular protein databases. *Codata Bulletin* **1992b**; 23:94-99.
- Simpson, R. J., Connolly, L. M., Eddes, J. S., Pereira, J. J., Moritz, R. L., and Reid, G. E. Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database. *Electrophoresis* **2000**; 21(9):1707-1732.
- Skolnick, J., and Fetrow, J. S. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends in Biotechnology* **2000**; 18(1):34-39.
- Smale, S. T. Core promoter architecture for eukaryotic protein-coding genes. In: Transcription: mechanisms and regulation. Conaway, R. C. and Conaway, J. W., editors. Raven Press, New York, New York, USA **1994**:63-81.
- Smale, S. T. Transcription initiation from TATA-less promoters within eukaryotic protein-encoding genes. *Biochimica et Biophysica Acta* **1997**; 1351(1-2):73-88.
- Smith, B. R. Visualizing human embryos. *Scientific American* **1999**; 280(3):76-81.
- Smith, B. R., Huff, D. S., and Johnson, G. A. Magnetic resonance imaging of embryos: an Internet resource for the study of embryonic development. *Computerized Medical Imaging and Graphics* **1999**; 23(1):33-40.
- Solovyev, V. V., and Rogozin, I. B. The program package of the context analysis of DNA, RNA and protein sequences. In: Search for homology and functional sites. Institute of Cytology and Genetics of the Union of Soviet Socialist Republics Academy of Science, Novosibirsk, Russia **1986**:1-70.
- Solovyev, V. V., and Salamov, A. A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In: Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology. Gaasterland, T., Karp P., Karpus K., Ouzounis C., Sander, C., and Valencia, A., editors. AAAI Press, Menlo Park, California, USA **1997**:294-302.
- Sorek, R., and Amitay, M. Piecing together the significance of splicing. *Nature Biotechnology* **2001**; 19(3):196.
- Sowa, J. F. Knowledge representation: logical, philosophical, and computational foundations. Brooks Cole Publishing Co., Pacific Grove, California, USA **2000**; <http://users.bestweb.net/~sowa/ontology/>.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Molecular Biology of the Cell* **1998**; 9(12):3273-3297.

- Sporle, R., and Schughart, K. Paradox segmentation along inter- and intrasomitic borderlines is followed by dysmorphology of the axial skeleton in the open brain (opb) mouse mutant. *Developmental Genetics* **1998**; 22(4):359-373.
- Sporle, R., and Schughart, K. System to identify individual somites and their derivatives in the developing mouse embryo. *Developmental Dynamics* **1997**; 210(3):216-266.
- Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H., and Schreiber, R. D. How cells respond to interferons. *Annual Review of Biochemistry* **1998**; 67:227-264.
- Stein, R., Mori, N., Matthews, K., Lo, L. C., and Anderson, D. J. The NGF-inducible SCG10 mRNA encodes a novel membrane-bound protein present in growth cones and abundant in developing neurons. *Neuron* **1988**; 1(6):463-476.
- Stein, L. D., and Thierry-Mieg, J. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Research* **1998**; 8(12):1308-1315.
- Stoeckert, C. J., Jr., Pizarro, A., Manduchi, E., Gibson, M., Brunk, B., Crabtree, J., Schug, J., Shen-Orr, S., and Overton, G. C. A relational schema for array and non-array based gene expression data. *Bioinformatics* **2001**; 17(4):300-308.
- Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P., and Tuli, M. A. The EMBL nucleotide sequence database. *Nucleic Acids Research* **2001**; 29(1):17-21.
- Stormo, G. D. Information content and free energy in DNA-protein interactions. *Journal of Theoretical Biology* **1998**; 195(1):135-137.
- Strausberg, R. L., Feingold, E. A., Klausner, R. D., and Collins, F. S. The mammalian gene collection. *Science* **1999**; 286(5439):455-457.
- Streicher, J., Weninger, W. J., and Müller, G. B. External marker-based automatic congruencing: A new method of 3D reconstruction from serial sections. *Anatomical Record* **1997**; 248(4):583-602.
- Streicher, J., Donat, M. A., Strauss, B., Sporle, R., Schughart, K., and Müller, G. B. Computer-based three-dimensional visualization of developmental gene expression. *Nature Genetics* **2000**; 25(2):147-152.
- Streicher, J., and Müller, G. B. 3D modelling of gene expression patterns. *Trends in Biotechnology* **2001**; 19(4):145-148.
- Stulik, J., Bures, J., Jandik, P., Langr, F., Kovarova, H., and Macela, A. The different expression of proteins recognized by monoclonal anti-heat shock protein 70 (hsp70) antibody in human colonic diseases. *Electrophoresis* **1997**; 18(3-4):625-628.
- Stulik, J., Kovarova, H., Macela, A., Bures, J., Jandik, P., Langr, F., Otto, A., Thiede, B., and Jungblut, P. Overexpression of calcium-binding protein calgranulin B in colonic mucosal diseases. *Clinica Chimica Acta* **1997**; 265(1):41-55.
- Stulik, J., Österreicher, J., Koupilova, K., Knizek, J., Macela, A., Bures, J., Jandik, P., Langr, F., Dedic, K., and Jungblut, P. R. The analysis of S100A9 and S100A8 expression in matched sets of macroscopically normal colon mucosa and colorectal carcinoma: the S100A9 and S100A8 positive cells underlie and invade tumor mass. *Electrophoresis* **1999**; 20(4-5):1047-1054.
- Stulik, J., Koupilova, K., Österreicher, J., Knizek, J., Macela, A., Bures, J., Jandik, P., Langr, F., Dedic, K., and Jungblut, P. R. Protein abundance alterations in matched sets of macroscopically normal colon mucosa and colorectal carcinoma. *Electrophoresis* **1999**; 20(18):3638-3646.
- Stulik, J., Österreicher, J., Koupilova, K., Knizek, J., Bures, J., Jandik, P., Langr, F., Dedic, K., Schafer, B. W., and Heizmann, C. W. Differential expression of the Ca^{2+} binding S100A6 protein in normal, preneoplastic and neoplastic colon mucosa. *European Journal of Cancer* **2000**; 36(8):1050-1059.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the USA* **1999**; 96(6):2907-2912.
- Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A genomic perspective on protein families. *Science* **1997**; 278(5338):631-637.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. The COG database: new developments in phylogenetic classifica-

- tion of proteins from complete genomes. *Nucleic Acids Research* **2001**; 29(1):22-28.
- Thalmann, I., Comegys, T. H., Liu, S. Z., Ito, Z., and Thalmann, R. Protein profiles of perilymph and endolymph of the guinea pig. *Hear Research* **1992**; 63(1-2):37-42.
- Thalmann, I., Kohut, R. I., Ryu, J., Comegys, T. H., Senarita, M., and Thalmann, R. Protein profile of human perilymph: in search of markers for the diagnosis of perilymph fistula and other inner ear disease. *Otolaryngology – Head and Neck Surgery* **1994**; 111(3 Part 1):273-280.
- Thalmann, I., Kohut, R. I., Ryu, J. H., and Thalmann, R. High resolution two-dimensional electrophoresis: technique and potential applicability to the study of inner ear disease. *American Journal of Otolaryngology* **1995**; 16(2):153-157.
- Thalmann, R., Henzl, M. T., and Thalmann, I. Specific proteins of the organ of Corti. *Acta Otolaryngologica* **1997**; 117(2):265-268.
- The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research* **2001**; 11(8):1425-1433.
- The Plasmodium Genome Database Collaborative. PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucleic Acids Research* **2001**; 29(1):66-69.
- Theiler, K. The house mouse. Springer-Verlag, Berlin **1972**; (63):168.
- Theiler, K. The house mouse: Atlas of embryonic development. Springer-Verlag, New York, USA **1989**.
- Thiede, B., Siejak, F., Dimmler, C., Jungblut, P. R., and Rudel, T. A two dimensional electrophoresis database of a human Jurkat T-cell line. *Electrophoresis* **2000**; 21(13):2713-2720.
- Thompson, C. C., and McKnight, S. L. Anatomy of an enhancer. *Trends in Genetics* **1992**; 8(7):232-236.
- Tikunov, Y., and Kel, A. Kernel method for estimation of functional site local consensi. Classification of transcription initiation sites in eukaryotic genes. In: GCB2000, Proceedings of the German Conference on Bioinformatics, October 5-7, 2000, Heidelberg. Bornberg-Bauer, E., editor. Logos-Verlag, Berlin, Germany **2000**:83-88.
- Toda, T., Kaji, K., and Kimura, N. TMIG-2DPAGE: a new concept of two-dimensional gel protein database for research on aging. *Electrophoresis* **1998a**; 19(2):344-348.
- Toda, T., Satoh, M., Sugimoto, M., Goto, M., Furuichi, Y., and Kirura, N. A comparative analysis of the proteins between the fibroblasts from Werner's syndrome patients and age-matched normal individuals using two-dimensional gel electrophoresis. *Mechanisms of Ageing and Development* **1998**; 100(2):133-143.
- Toda, T., Sugimoto, M., Omori, A., Matsuzaki, T., Furuichi, Y., and Kimura, N. Proteomic analysis of Epstein-Barr virus-transformed human B-lymphoblastoid cell lines before and after immortalization. *Electrophoresis* **2000**; 21(9):1814-1822.
- Toda, T. Proteome and proteomics for the research on protein alterations in aging. *Annals of the New York Academy of Sciences* **2001**; 928:71-78.
- Törönen, P., Kolehmainen, M., Wong, G., and Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Letters* **1999**; 451(2):142-146.
- Tonella, L., Walsh, B. J., Sanchez, J. C., Ou, K., Wilkins, M. R., Tyler, M., Frutiger, S., Gooley, A. A., Pescaru, I., Appel, R. D., Yan, J. X., Bairoch, A., Hoogland, C., Morch, F. S., Hughes, G. J., Williams, K. L., and Hochstrasser, D. F. '98 *Escherichia coli* SWISS-2DPAGE database update. *Electrophoresis* **1998**; 19(11):1960-1971.
- Tonella, L., Hoogland, C., Binz, P. A., Appel, R. D., Hochstrasser, D. F., and Sanchez, J. C. New perspectives in the *Escherichia coli* proteome investigation. *Proteomics* **2001**; 1(3):409-423.
- Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and Pontoglio, M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *Journal of Molecular Biology* **1997**; 266(2):231-245.
- Tsonis, P. A. A comparative two-dimensional gel protein database of the intact and regenerating newt limbs. *Electrophoresis* **1993**; 14(1-2):148-156.
- Valuev, V. P., Afonnikov, D. A., Ponomarenko, M. P., Milanesi, L., and Kolchanov, N. A. ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and pep-

- tides evolved *in vitro*. *Nucleic Acids Research* **2002**; 30(1):200-202.
- Van Bogelen, R. A., Hutton, M. E., and Neidhardt, F. C. Gene-protein database of *Escherichia coli* K-12: edition 3. *Electrophoresis* **1990**; 11(12):1131-1166.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **1995**; 270(5235):484-487.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu X. The sequence of the human genome. *Science* **2001**; 291(5507):1304-1351.
- Verbeek, F. J., Lawson, K. A., and Bard, J. B. Developmental bioinformatics: linking genetic data to virtual embryos. *International Journal of Developmental Biology* **1999**; 43(7):761-771.
- Vilo, J., Brazma, A., Jonassen, I., Robinson, A., and Ukkonen, E. Mining for putative regulatory elements in the yeast genome using gene expression data. In: ISMB'2000. AAAI Press, Menlo Park, California, USA **2000**: 384-394.
- Vohradsky, J., Li, X. M., and Thompson, C. J. Identification of procaryotic developmental stages by statistical analyses of two-dimensional gel patterns. *Electrophoresis* **1997**; 18(8):1418-1428.
- Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **1999**; 15(10): 776-784.

- Wait, R., Gianazza, E., Eberini, I., Sironi, L., Dunn, M. J., Gemeiner, M., and Miller, I. Proteins of rat serum, urine, and cerebrospinal fluid: VI. Further protein identifications and interstrain comparison. *Electrophoresis* **2001**; *22*(14):3043-3052.
- Ward, L. D., Hong, J., Whitehead, R. H., and Simpson, R. J. Development of a database of amino acid sequences for human colon carcinoma proteins separated by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **1990**; *11*(10):883-891.
- Ward, L. D., Ji H., and Simpson, R. J. Two-dimensional electrophoresis of proteins from cultured human colonic carcinoma cells: Development of a database of amino acid sequences. In: Techniques in protein chemistry: II. Villafranca, J. J., editor. Academic Press, San Diego, California, USA **1991**: 21-30.
- Wasserman, W. W., and Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology* **1998**; *278*(1):167-181.
- Watanabe, J., Sasaki, M., Suzuki, Y., and Sugano, S. FULL-malaria: a database for a full-length enriched cDNA library from human malaria parasite, *Plasmodium falciparum*. *Nucleic Acids Research* **2001**; *29*(1):70-71.
- Wattiez, R., Hermans, C., Bernard, A., Lesur, O., and Falmagne, P. Human bronchoalveolar lavage fluid: two-dimensional gel electrophoresis, amino acid microsequencing and identification of major proteins. *Electrophoresis* **1999**; *20*(7):1634-1645.
- Wattiez, R., Hermans, C., Cruyt, C., Bernard, A., and Falmagne, P. Human bronchoalveolar lavage fluid protein two-dimensional database: study of interstitial lung diseases. *Electrophoresis* **2000**; *21*(13):2703-2712.
- Weiller, G. F., Djordjevic, M. J., Caraux, G., Chen, H., and Weinman, J. J. A specialised proteomic database for comparing matrix-assisted laser desorption/ionization-time of flight mass spectrometry data of tryptic peptides with corresponding sequence database segments. *Proteomics* **2001**; *1*(12):1489-1494.
- Weldingh K., Hansen A., Jacobsen S., and Andersen P. High resolution electroelution of polyacrylamide gels for the purification of single proteins from *Mycobacterium tuberculosis* culture filtrate. *Scandinavian Journal of Immunology* **2000**; *51*(1):79-86.
- Welle, S., Bhatt, K., and Thornton, C. A. Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Research* **1999**; *9*(5):506-513.
- Welle, S., Bhatt, K., and Thornton, C. A. High-abundance mRNAs in human muscle: comparison between young and old. *Journal of Applied Physiology* **2000**; *89*(1):297-304.
- Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., and Somogyi, R. Large-scale temporal gene expression mapping of CNS development. *Proceedings of the National Academy of Sciences of the USA* **1998**; *95*(1):334-339.
- Werhahn, W., and Braun, H. P. Biochemical dissection of the mitochondrial proteome from *Arabidopsis thaliana* by three-dimensional gel electrophoresis. *Electrophoresis* **2002**; *23*(4):640-646.
- Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2000**; *28*(1): 10-14.
- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2001**; *29*(1):11-16.
- Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., and Rapp, B. A. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Research* **2002**; *30*(1):13-16.
- Williams, K., Chubb, C., Huberman, E., and Giometti, C. S. Analysis of differential protein expression in normal and neoplastic human breast epithelial cell lines. *Electrophoresis* **1998**; *19*(2):333-343.
- Wingender, E. Gene regulation in eukaryotes. VCH Verlagsgesellschaft mbH, Weinheim, Germany **1993**.
- Wingender, E., Karas, H., and Kneuppel, R. TRANSFAC database as a bridge between sequence data libraries and biological function. In: Pacific Symposium on Biocomputing '97 (PSB'97). Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E., editors.

- World Scientific, Toh Tuck Link, Singapore **1996**:477-485.
- Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knueppel, R., Romaschenko, A. G., and Kolchanov, N. A. TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation *Nucleic Acids Research* **1997**; *25*(1):265-268.
- Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S., and Urbach, S. The TRANSFAC system on gene expression regulation. *Nucleic Acids Research* **2001**; *29*(1):281-283.
- Wirth, P. J., Luo, L. D., Fujimoto, Y., Bisgaard, H. C., and Olson, A. D. The rat liver epithelial (RLE) cell protein database. *Electrophoresis* **1991**; *12*(11):931-954.
- Wirth, P. J., Luo, L. D., Benjamin, T., Hoang, T. N., Olson, A. D., and Parmelee, D. C. The rat liver epithelial (RLE) cell nuclear protein database. *Electrophoresis* **1993**; *14*(11): 1199-1215.
- Wolfsberg, T. G., and Landsman, D. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Research* **1997**; *25*(8):1626-1632.
- Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z. Z., Ledley, R. S., Lewis, K. C., Mewes, H. W., Orcutt, B. C., Suzek, B. E., Tsugita, A., Vinayaka, C. R., Yeh, L. S., Zhang, J., and Barker, W. C. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Research* **2002**; *30*(1):35-37.
- Wyrick, J. J., Holstege, F. C., Jennings, E. G., Causton, H. C., Shore, D., Grunstein, M., Lander, E. S., and Young, R. A. Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **1999**; *402*(6760):418-421.
- Wyrick, J. J., and Young, R. A. Deciphering gene expression regulatory networks. *Current Opinion in Genetics and Development* **2002**; *12*(2):130-136.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **2002**; *30*(1):303-305.
- Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Gooley, A. A., Hochstrasser, D. F., and Williams, K. L. The *Dic-tyostelium discoideum* proteome – the SWISS-2DPAGE database of the multicellular aggregate (slug). *Electrophoresis* **1997**; *18*(3-4):491-497.
- Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics* **2001**; *17*(4):309-318.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. *Science* **1997**; *276*(5316):1268-1272.
- Zhang, M. Q. Identification of human gene core promoters *in silico*. *Genome Research* **1998a**; *8*(3):319-326.
- Zhang, M. Q. Statistical features of human exons and their flanking regions. *Human Molecular Genetics* **1998b**; *7*(5):919-932.
- Zhang, K., and Zhao, H. Assessing reliability of gene clusters from gene expression data. *Functional Integrative Genomics* **2000**; *1*(3):156-173.

Index

a

- a *Caenorhabditis elegans* database (ACeDB) 820, 856
- a *Xenopus laevis* database (Axelddb) 856
- A-box 18
- A-site 7–8
- absorption coefficient 108
- ACeDB, *see* a *Caenorhabditis elegans* database
- acetylation 24–25, 660
- acetylation of histones 22
- β -actin 66
- γ -actin 66
- activator 21 f.
- acylation 53
 - acetyl group 53
 - acetyl-coenzyme A 53
 - myristoyl group 53
 - myristoyl-coenzyme A 53
- adapter-tagged competitive PCR (ATAC-PCR) 278 ff., 321
 - automatic DNA sequencer 280
 - cDNA digestion 280
 - characteristics 281
 - cluster analysis 283
 - DNA sequencing 281
 - electropherogram 280
 - EST sequencing 281
 - expression profiling 281
 - group classification 283
 - large-scale analysis 281
 - paramagnetic beads 280
 - principle 278
- adapter proteins 800
- ADDER, *see* amplification of double-stranded cDNA end restriction fragments
- ADGE, *see* amplified differential gene expression
- adherence chromatography 105
- affinity chromatography 105, 625, 631, 633–636, 660
 - principle 632
- affinity column chromatography 105
- affinity mass spectrometry, *see* affinity-combined mass spectrometry
- affinity-combined mass spectrometry 659, 660
- AFLP, *see* amplified fragment length polymorphism
- AFM, *see* atomic force microscopy
- alcohol precipitation 138
- Alexa 437
- alkaline phosphatase 439
- alternative processing 5
- alternative splice database of mammals (AsMamDB) 825 f.
- alternative splicing 5, 38 f., 50, 824 f., 833
- alternative splicing database (ASDB) 39, 824 f.
- alternatively spliced transcripts 824 f.
- alternative splice database of mammals 825 ff.
- ambiguous letter code 802
- American Standard Code for Information Interchange (ASCII) 791
- american type culture collection (ATCC) 98, 777
- aminoacyl tRNA 6
- aminoacyl tRNA-binding site 7
- amperometry 667
- amplification *in vitro* 99
- amplification of double-stranded cDNA end restriction fragments 267 ff.
 - adapter 269
 - ADDER display libraries 276
 - cDNA restriction fragments 268
 - digestion 269
 - downstream primer 271
 - exponential phase 275
 - false-positives 274

- gel electrophoresis 269
- ligation 269
- paramagnetic streptavidin beads 268
- PCR amplification 268–269, 275–276
- PCR efficiency 276
- polyacrylamide gel 270
- principle 268 ff.
- real-time RT-PCR 276
- restriction fragment 270, 274, 277
- RNA complexity 267
- sequence information 277
- single nucleotide polymorphism (SNP) 274
- SNP, *see* single nucleotide polymorphism
- touch-down PCR 269
- upstream primer 271
- amplified differential gene expression (ADGE) 478
 - adapter 478
 - ADGE-combined microarray 482 f.
 - control 478, 480, 482
 - CT adapter 478–479, 481–483
 - CT primer 478, 480–483
 - digestion 481
 - DNA microarray 478, 480–481
 - DNA reassociation 481–482
 - gel separation 480
 - PCR amplification 481
 - principle 478 ff.
 - quadratic magnification 481–482
 - restriction enzyme 478, 481
 - scatter plot 482–483
 - tester 478, 480, 482
 - TT adapter 478–479, 481, 483
 - TT primer 478, 480–483
- amplified fragment length polymorphism (AFLP) 177, 285
- AMPP, *see* *Arabidopsis* mitochondrial proteome project database
- analysis of variance arguments (ANOVA) 797
- anion-exchange 139
- ANOVA, *see* analysis of variance arguments
- antibodies 671 ff., 687
 - affinity 671, 673
 - antigen 671
 - antigen-binding region 672
 - bovine serum albumin 672
 - capture 674
 - computer-assisted molecular modelling 673
 - construction 673
 - direct label 682
 - epitope 671–672
 - flexible hinge region 672
 - fluorochrome-conjugated 687
 - heavy chain 672
 - heterophile 686
 - high-affinity 674
 - human anti-mouse 685
 - keyhole limpet haemocyanin (KLH) 672
 - light chain 672
 - monoclonal 672
 - paratope 671–672
 - polyclonal 671
 - signal 674
 - specificity 673
 - stability 673
 - structure 672
- antibody library 665, 673
- anticoagulant 142
- antisense RNA 45 ff.
- ANU-2DPAGE, *see* Australian National University two-dimensional polyacrylamide gel electrophoresis database
- AP-PCR (arbitrarily primed PCR) 340
- Apache 846
- apoenzyme-reactivation immunoassay (ARIS) 676
- aptamers 673
- Arabidopsis* mitochondrial proteome project database (AMPP) 861
- ARE, *see* AU-rich response element
- Argonne protein mapping group server 861
- ARIS, *see* apoenzyme-reactivation immunoassay
- Array Database at the National Human Genome Research Institute (ArrayDB) 856
- Array Designer Software 433
- Array-Plate™ multiplexed molecular profiling system 471, 475, 477
- ArrayDB, *see* array database at the National Human Genome Research Institute
- ArrayExpress 826
- artificial intelligence 804
- Artificially Selected Proteins/Peptides Database (ASPD) 853
- Asalign 825
- ASCII, *see* American Standard Code for Information Interchange
- ASDB, *see* alternative splicing database
- AsMamDB, *see* alternative splice database of mammals
- ASPD, *see* Artificially Selected Proteins/Peptides Database
- association rate constant 669 f.
- ATAC-PCR, *see* adapter-tagged competitive PCR
- ATCC, *see* American Type Culture Collection
- atomic force microscopy (AFM) 449, 667
 - AFM tip 452

- attractive mode 450
- CFM, *see* chemical force microscopy 451 ff.
- chemical force microscopy (CFM) 451 ff.
- DNA-modified tip 451–452
- fermions 450
- MeNPOC 450
- molecular interaction 451
- non-contact mode 450
- quantum mechanical exclusion principle 450
- repulsive forces 450
- resonant frequency 450
- topographic contours 450
- van der Waals interaction 450
- ATR, *see* attenuated total reflection
- attenuated total reflection (ATR) 667
- attenuation 34
- AU-rich response element (ARE) 31
- Australian National University two-dimensional polyacrylamide gel electrophoresis database (ANU-2DPAGE) 861
- auto-nucleation 448
- auto-radiography 411, 442
- auto-radiography imaging systems 438
- Autogene Promoter 808
- Axeldb, *see a Xenopus laevis* database

b

- B-box 18
- B-DNA 26
- background false-positive rate 384
- bacteria 140, 143
 - gram-negative 143
 - gram-positive 143
- BALF 2D, *see* two-dimensional polyacryl-amide gel electrophoresis database for broncho-alveolar lavage fluid proteins
- basal transcription complex 799
- basic local alignment search tool (BLAST) 772, 821, 830
- basic transcription factor 18, 20, 49
- Bayesian logic 785
- bdNA, *see* branched DNA
- bdNA-ISH, *see* branched DNA *in situ* hybridisation
- bead arrays, *see* suspension arrays
- Best Hit method, *see* BeT method (BeT)
- BeT method (BeT) 775
- BIA-MS, *see* biomolecular interaction analysis mass spectrometry
- binding of activating or removal of inhibitory factors 50
- bioinformatics template library 786
- BioKnowledge® library 866
- bioluminescence 755
- bioluminescence imaging system 755
- bioluminescence immunoassay 679
- bioluminescent reporter probe 753
 - *Discosoma* species red fluorescent protein (DsRed) 754
 - DsRed, *see Discosoma* species red fluorescent protein
 - green fluorescent protein (GFP) 753
 - luciferases 753
- bioluminescent reporter probe 753
- biomolecular interaction analysis 667
- biomolecular interaction analysis mass spectrometry (BIA-MS) 661
- BioSap 433
- biotin 436, 439, 445, 451, 471 f., 559
- Blast Integrated Oligonucleotide Selection Accelerator Package 433
- BLAST, *see* basic local alignment search tool
- BODIPY® 630/650 437
- body coils 733
- BodyMap 506, 510, 825 f.
- Boolean screening 820
- BPP-2DPAGE, *see* protein database of haematopoietic cell lines
- Brain-Sort® 487, 489
- branched DNA 164 ff.
- branched DNA *in situ* hybridisation 165 f.
- branched DNA assay 164 ff.
 - capture probe 164 f.
 - flow cytometry 166
 - signal amplification 164
- Brewster angle spectrometry 667, 669
- bromodeoxyuridine 445
- Bucher's specifications 811

c

- C/EBP, *see* CCAAT/enhancer-binding protein
- CA-IEF 626
- CAAT box 4
- caesium chloride 138 f.
- caesium chloride density-gradients 138 f.
- γ-camera 747 f.
- Cancer Genome Anatomy Project (CGAP) 551, 571, 575, 776, 844
 - CGAP library 777
- cap site 849
- capillary electrochromatography (CEC) 635
 - column 635
 - electroosmotic flow 635
 - EOF, *see* electroosmotic flow
 - silica capillary 635
 - tryptic digest 635

- capillary electrophoresis (CE) 624, 635
 - column 635
 - electroendosmotic flow 635
 - EOF, *see* electroendosmotic flow 635
 - silica capillary 635
 - tryptic digests 635
- capillary-electrophoretic separation 439
- capture ELISA 683
- capture ELISPOT 680
- capture enzyme-linked immunosorbent assay 683
- Capture® 2 technology 475
- carboxylation 57
- Catch 806
- CCAAT box transcription factor 27
- CCAAT/enhancer-binding protein 27
- CCCTC-binding factor 24
- CCD, *see* charge-coupled device
- CCLS, *see* chemical cross linking subtraction
- cDNA array 365, 368, 369
- cDNA libraries 463, 570, 575
- cDNA library screening 468
- cDNA xProfiler 97 f., 856
- cDNA-AFLP, *see* cDNA-based amplified fragment length polymorphism fingerprinting
- cDNA-based amplified-fragment length polymorphism fingerprinting (cDNA-AFLP) 285 ff., 359
 - adapter 288
 - anchor design 291
 - anchor ligation 287
 - band isolation 292
 - co-amplification 295
 - co-migration 295
 - dissociation temperature 291
 - electro-elution 292
 - enzymatic digestion 288
 - false-positives 295
 - fingerprint generation 292
 - flexibility 296
 - ligation 288
 - linkage map 295
 - melting temperature 291
 - mis-priming 295
 - PCR amplification 288
 - polymorphic transcripts 295
 - pre-amplification 287, 291
 - primer design 291
 - primer extension 296
 - principle 286
 - radioactive labelling 295
 - re-amplification 292
 - restriction enzyme digestion 287
 - restriction enzyme recognition sites 290 f.
 - RNA fingerprint 285
 - sensitivity 295
 - TDF, *see* transcript-derived fragment 285
 - template preparation 287
 - transcript-derived fragment 285, 292, 294
- cDNA-RDA, *see* representational difference analysis of cDNA
- CE, *see* capillary electrophoresis
- CE, *see* control element
- CEC, *see* capillary electrochromatography
- CEDIA, *see* cloned enzyme donor immunoassay
- cell culture models 98
- cell cycle 11–12
- cell disruption 144
 - detergent 144
 - French pressing 144
 - shearing 144
 - sonication 144
- cell extract 144
 - equilibrium sedimentation 145
 - fractionation 145
 - velocity sedimentation 145
- cell homogenate 144
- cell isolation 99 ff.
 - adherence chromatography 105
 - affinity chromatography 105 f.
 - affinity column chromatography 105
 - centrifugation 106
 - column chromatography 105 f.
 - countercurrent elutriation 103 ff.
 - density-gradient centrifugation 106 f.
 - dielectric affinity separation 105
 - differential-velocity centrifugation 103 ff.
 - equilibrium 106
 - flow cytometry 107 ff.
 - immunopurification of cells 99 ff.
 - Stoke's law 106
 - surface affinity chromatography 105 f.
- cell lines 98
- cell preparation, *see* cell isolation
- cell sorting 107 ff.
- cell transfection, *see* transfection
- cellular quality control of translation 50
- CFM, *see* chemical force microscopy
- CFP, *see* cyan fluorescent protein
- CGAP, *see* cancer genome anatomy project 551, 571, 575, 576, 776, 831, 844
- chaotropic agent 135 ff.
- charge-coupled device (CCD) 345, 446, 462, 542–543, 753
- charge-coupled device camera 345, 436
- chemical cross linking subtraction (CCLS) 180

- chemical force microscopy (CFM) 451 ff.
- chemical shift 733 f.
- chemical synthesis 439
- chemiluminescence immunoassay (LIA) 679
- chemiluminescent detection 439 ff.
 - *in vitro* transcription 439
 - 3'-end labelling 439
 - 3'-tailing 439
 - alkaline phosphatase 439
 - background signal 442
 - biotin 439
 - cDNA synthesis 439
 - chemical synthesis 439
 - CSPD® 440
 - dephosphorylation 440
 - digoxigenin 439
 - dynamic range 442
 - nick translation 439
 - PCR amplification 439
 - principle 440
 - random primed labelling 439
 - sensitivity 442
 - specificity 442
- ChIP 12
- ChIP on chip 12
- chloroform 137
- chromatin 22 ff., 25
- chromatin immunoprecipitation 12
- chromatin opening activity 22
- chromatin remodelling 21
- circularisation of mRNA 51
- Cis-element Cluster Finder, *see* Cister
- cis*-elements, *see cis*-regulatory elements
- cis*-regulatory elements 800, 802
 - closed-loop model 51
- Cister (Cis-element Cluster Finder) 807
- citrate 142
- class IIS restriction enzyme 485 ff.
- classical silencer element 26
- classification 74 ff.
- CLICK, *see* cluster identification *via* connectivity kernels
- cloned enzyme donor immunoassay (CEDIA) 676
- closed architecture 74 f., 484
- cluster affinity search technique (CAST) 793
- cluster analysis 500, 790 ff.
 - cluster structure 791
 - cluster validation 797 f.
 - data normalisation 791
 - dissimilarity measures 792
 - euclidean distance 792
 - expression profile 790
 - external cluster validation 797
 - functional categorisation 790
 - gene expression matrix 792
 - guilt-by-association concept 790
 - hamming distance 792
 - internal cluster validation 797
 - metrics 792
 - molecular classification of diseases 791
 - mutual information 792
 - network analyses 790
 - Pearson correlation 791–792
 - similarity matrix 792
 - similarity measure 792
 - Spearman's rank correlation 792
- cluster identification *via* connectivity kernels (CLICK) 793
- cluster of orthologous groups (COG) 775
- clustered expressed sequence tag 771
- clustered sets of expressed sequence tags 776 f.
- clusters of sites 806 ff.
 - genetic algorithm 807
- clusters of SWISS-PROT and TrEMBL protein database (CluSTr) 853, 861
- ClusterScan 807 f.
- CluSTr, *see* clusters of SWISS-PROT and TrEMBL protein database
- co-activator 18, 20 f.
- co-expression 819
- co-factor 57
- co-repressor 26 ff., 28
 - global 28
 - specific 28
- coding sequence tag 333
- coding strand 3 f.
- codon
 - anticodon 5 ff.
 - start codon 7
 - stop codon 5 ff.
 - termination signal 5 ff.
- COG, *see* cluster of orthologous groups
- colourimetry 667
- column chromatography 105, 631, 633
 - affinity chromatography 632–634
 - cation exchange 633
 - column 631, 633 f.
 - gel filtration 633
 - IEC, *see* ion exchange chromatography
 - ion exchange chromatography (IEC) 632–633
 - mobile phase 631, 633–634
 - normal phase chromatography 634
 - principle 631
 - reverse phase chromatography (RPC) 634
 - RPC, *see* reverse phase chromatography

- salt gradient 633
- SCX, *see* strong cation-exchange
- SEC, *see* size exclusion chromatography
- size exclusion chromatography 632
- solid phase 631
- stationary phase 633–634
- strong cation-exchange (SCX) 633
- column-liquid chromatography 439
- combinatorial matrix gene analysis 206
- combinatorial nature of transcription regulation 801
- combinatorial regulation 800, 806
- comparative EST analysis 468
- comparative expressed sequence tag analysis 770, 771
 - *In silico* subtraction 776 f.
 - alternative splice variants 770
 - alternative splicing 776
 - cDNA libraries 770, 776–777
 - CGAP library 777
 - classification of ESTs 774
 - clustering of ESTs 772
 - comparative genomics 770
 - differentially expressed transcripts 770
 - domain detection 773
 - EST grouping 773
 - EST processing 773
 - functional annotation 775
 - functional classification 773
 - gene identification 770
 - gene prediction 770
 - genome annotation 776
 - genome comparisons 775
 - homology search 772–773
 - orthologous groups 776
 - orthologue 775
 - phylogenetic studies 776
 - pre-normalised cDNA library 776
 - prediction of open reading frames 773
 - quality measurement 771
 - redundancy 771, 776
 - SAGE libraries 777, 770 ff.
- COMPEL 801
- competitive immunoassays 674 f.
- competitive RT-PCR 70, 296 ff.
 - accuracy 297
 - agarose gel 297
 - competitor 296–297, 301
 - competitor mRNA 298–299, 301
 - crossover contamination 301
 - gel electrophoresis 298, 299
 - heteroduplex 299–301
 - high-performance liquid chromatography (HPLC) 298, 299
 - HPLC, *see* high-performance liquid chromatography
 - ion paired-reverse phase-high performance liquid chromatography IP-RP-HPLC 296, 300–301
 - IP-RP-HPLC, *see* ion paired-reverse phase-high performance liquid chromatography
 - multiple tube 300
 - multiple-tube titration 299
 - PCR amplification efficiency 301
 - PCR efficiency 297 f.
 - principle 296 ff., 299
 - Q-RT-PCR program 300
 - regression plot 299
 - reverse transcription efficiency 297, 301
 - sensitivity 297
 - single tube 300
 - single-tube competitive RT-PCR 299, 301
 - wild-type transcript 298
- competitive hybridisation, *see* subtractive hybridisation
- COMPLUYEAST-2DPAGE, *see* Yeast protein database COMPLUYEAST-2DPAGE
- composite clusters 807
- composite elements 800, 805, 806
- composite regulatory elements, *see* composite elements
- Comprehensive Perl Archive Network (CPAN) 781
- composite score 806
- computed tomography (CT) 741, 752
- computer beamers 428
- Consed 772–773
- consensus sequence 3 f.
- conservation of RNA sequences 51
- ConsInspector 803–804
- contac-tip deposition printing 416, 417 ff.
 - needles 417
 - screw 417
 - tips 417
- control element 17
- control of translation 50
- controlled degradation 50
- cooperative binding 22
- core promoter 20, 799, 809
- CorePromoter Finder 809 f.
- correlation between positions 804
 - dinucleotide matrices 804
 - dinucleotide weight matrices 804
 - information measure 804
 - trinucleotide matrices 804
- correlation of mRNA and protein expression 58
- Coulter Counter® 108

- countercurrent elutriation 103 ff.
 - blood cell separation 103
 - cell elutriator 103
 - centrifugal force 103 f.
 - shear force 103 f.
- coupled PACS and SuRF 338
- μCP, *see* micro-contact printing
- CPAN, *see* Comprehensive Perl Archive Network
- CpG island 29, 800, 809–810
- CpG methylation 29
- cross contamination 425 f.
- cross-match 773
- CSPD® 440
- CT, *see* computed tomography
- CTCF, *see* CCCTC-binding factor
- CTF, *see* CCAAT box transcription factor
- μCUT, *see* laser microdissection ultraviolet cut
- Cy3™ 436–437
- Cy5™ 436–437
- cyan fluorescent protein (CFP) 689 ff.
- cycloheximide 344
- CYTOMER® gene expression database 826 ff.
- cytometry 107 ff.
- cytoplasmic RNA localisation, *see* intracellular localisation of RNA
- d**
- Dalton 639
- DAN, *see* deadenylating nuclease
- DAPI, *see* 4',6-diamidino-2-phenylindole
- data integration 788 ff.
 - cross-species expression profiling 788
 - experimental design 788 ff.
 - standard 788
 - type I data sets 788
 - type II data sets 788
- data management 775, 778
 - fields 780
 - meta information 786
 - query 781
 - records 780
 - Boolean form 781
 - calibration 782
 - communication structures 781
 - comparison 783
 - comparison algorithms 782
 - data exchange protocols 781
 - data formats 781
 - data generation 782
 - graphical user interface (GUI) 778
 - GUI, *see* graphical user interface
 - hardware demands 780
 - high-throughput experiments 779
 - information exchange 780
 - integration 778
 - internal alignment normalisation 782
 - model algorithms 782
 - normalisation procedure 783
 - operating system 780
 - processing 778
 - quality control 779
 - query languages 785
 - reduction algorithms 782
 - relational databases 780
 - representation 778
 - retrieval 778
 - software demands 780
 - software standards 779
 - statistical analysis 784
 - trans-national scientific standards 779
 - user interfaces 786
- data mining, *see* data management
- database of expressed sequence tags (dbEST) 575, 773, 777, 829
 - invertebrates 773
 - plants 773
 - vertebrates 773
- Database of Interacting Proteins (DIP™) 853
- Database of the Translational Recoding Events (RECODE) 859
- Database of Transposon-Insertion Phenotypes, Localisation, and Expression in *Saccharomyces cerevisiae* (TRIPLES) 847 f.
- Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs (UTRdb) 849
- Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs (UTRsite) 849
- dbEST, *see* database of expressed sequence tags
- dCST, *see* coding sequence tag
- DDD, *see* digital differential display
- DDRT-PCR, *see* differential display RT-PCR
- de novo* protein synthesis 50
- dead-time 114
- deadenylating nuclease 33
- decision tree 810
- deformylation, post-translational modification 57
- DEgenerate Oligo Nucleotide Design and Analysis System (DEODAS) 433
- degradation 9, 141
- densitometry 667
- density-gradient centrifugation 106
- DEODAS, *see* DEgenerate Oligo Nucleotide Design and Analysis System
- DEPC, *see* diethyl pyrocarbonate

- DEPD[®], *see* digital expression pattern display
- DEPDstat 490
- dephosphorylation 440
- Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH 98
- Deutsches Krebsforschungszentrum (DKFZ), *see* German Cancer Research Center
- Deutsches Ressourcenzentrum für Genomforschung GmbH 777
- DGED, *see* differential gene expression displayer
- DGED, *see* digital gene expression displayer
- DGTools 533
- DHGP, *see* German Human Genome Project
- 4',6-diamidino-2-phenylindole (DAPI) 688
- dielectric affinity separation 105
- diethyl pyrocarbonate (DEPC) 135
- differential display, *see* fluorescent mRNA differential display
- differential display RT-PCR 177, 218–219, 237 ff., 319, 326, 340, 343, 358, 495 ff.
- differential display RT-PCR (DDRT-PCR), *see* fluorescent mRNA differential display
- differential gene expression displayer (DGED) 98, 776
- differential splicing 38 f.
- differential subtraction chain (DSC) 185 ff.
- λ amplicon 190
 - λ DNA 189
 - amplicon 186
 - conversion 189
 - double exponential decrease 187
 - driver 186, 189
 - enrichment efficiency 187, 191
 - enrichment kinetics 190
 - enrichment speed 190
 - equalisation 189
 - genomic representational sampling 186
 - linear amplification 191
 - linear regression 189
 - negative amplification 186
 - negative enrichment 186
 - normalisation 189
 - principle 188
 - probability 189
 - tester 186, 189
- differential subtraction display (DSD) 180
- differential display RT-PCR (DDRT-PCR) 180
- differential-velocity centrifugation 103 ff.
- blood cell separation 103
 - cell elutriator 103
 - centrifugal force 103 f.
 - shear force 103 f.
- differentiation induction subtraction hybridisation (DISH) 178 ff.
- cDNA libraries 179, 182, 185
 - DISH library 184
 - driver 182
 - high-density DNA microarrays 183
 - Northern blot 183
 - OPS, *see* overlapping pathway screening
 - overlapping pathway screening 185
 - protocol 182
 - reverse Northern 182–183, 185
 - temporally-spaced cDNA libraries 179 ff.
 - tester 182
- diffuse optical tomography (DOT) 753
- 2D-DIGE, *see* two-dimensional difference gel electrophoresis
- digestion 568
- digital differential display (DDD) 830 ff.
- digital expression pattern display (DEPP[®]) 484 ff.
- adapter 485
 - adapter ligation 485
 - advantages 493
 - automated procedure 491
 - Brain-Sort[®] 487, 489
 - capillary electrophoresis 485–487, 489
 - class IIS restriction enzymes 485 ff.
 - data analysis 491
 - DEPDstat 490
 - disadvantages 493
 - drug-profiling 489
 - false-positive peaks 490
 - false-positives 490
 - Genomax 491
 - PCR amplification 485, 487
 - principle 485 ff.
 - redundancy 486
 - reliability 485
 - reproducibility 485
 - resolution 489
 - restriction enzyme digestion 485
 - sensitivity 489
 - statistical data analysis 489
 - statistical significance 490
 - variance 490
- digital gene expression displayer (DGED) 776, 857
- digital micro-mirrors 428 f.
- computer beamers 428
 - television technology 428
 - ultraviolet light 428
 - virtuous mask 428
- digoxigenin 439, 445, 471 f.

- DIP™, *see* database of interacting proteins
- direct capture enzyme immunoassay 681
- direct electrical measurements 434
- direct immunometric-enzyme immunoassay 681
- direct optical biosensors 668
- Direct Protect™ Lysate RPA 476
- directed random oligonucleotide priming (DROP) 180
- chemical cross linking subtraction (CCLS) 180
- Discosoma* species red fluorescent protein (DsRed) 754
- discovery science 58, 62
- disruption 135 f.
- disruption of cells 144
- detergent 144
 - French pressing 144
 - shearing 144
 - sonication 144
- dissociation constants 669 f.
- dissociation rate constant 669 f.
- DKFZ, *see* German Cancer Research Center
- DNA chip 410–411
- DNA methylation 839
- DNA methylation database (MethDB) 839
- DNA methyltransferases 839
- DNA mfold 353
- DNA micro-contact printing (μCP) 419
- DNA microarray 410 ff., 443, 478
- *in situ* synthesis 417, 426
 - *in vitro* transcription 412, 435–436
 - Alexa 437
 - bilayer resist procedure 430
 - biotin 413, 436
 - BODIPY® 630/650 437
 - calibration 412
 - cDNA synthesis 435
 - computer beamers 428
 - contact-tip deposition printing 416, 417 ff.
 - cross contamination 425 f.
 - CSPD® 440
 - Cy3™ 413
 - Cy5™ 413
 - dephosphorylation 440
 - digital micro-mirrors 428 f.
 - direct electrical measurements 434
 - DNA-binding dyes 436
 - electro capture 416, 423 ff.
 - FAM 437
 - FITC 437
 - fluorescent dyes 437
 - fluorescent labels 435
 - high density 411 f., 418
 - Hoechst 33258 436
 - hybridisation conditions 416
 - hybridisation kinetics 414
 - immobilisation 417
 - indirect fluorescence 437
 - inter-spot distance 418, 420
 - interferometric methods 434
 - internal standards 412
 - JOE 437
 - low-density 411
 - medium-density 411
 - MeNPOC 427
 - micro-contact printing (μCP) 416, 419
 - micro-fluidics network (μFN) 416, 419 f.
 - micro wet printing (μWP) 417, 431 f.
 - non-specific binding 415
 - normalisation 412
 - nozzle 422
 - nucleic acids detection 434 ff.
 - nucleic acids labelling 434 ff.
 - oligoethyleneglycol 419
 - oligonucleotide design 432 f.
 - PCR 412
 - PCR amplification 435
 - phosphoramidite 432
 - photolithographic masks 426 f.
 - photolithography 417, 426 ff., 428 f.
 - photoresist wet masking 429 ff.
 - photosensitive protective groups 426 f.
 - phycoerythrin 413
 - piezoelectric dispenser 420, 422
 - piezoelectric printers 420
 - piezoelectric printing 416, 420 ff.
 - piezoelectric tip 420–421
 - pin-and-ring array system 418
 - poly-L-lysine 418
 - printing cartridge 431
 - production 416 ff., 417
 - PYMOC 427
 - quality analysis 412
 - quartz crystal microbalance dissipation 434
 - radioactive labelling 438 f.
 - radioactivity 437
 - random primed labelling 435
 - reflectometric interference spectroscopy (RiFS) 434
 - reproducibility 422
 - ROX 437
 - SAW, *see* surface acoustic wave
 - Schiff-base reaction 419
 - signal amplification 437
 - silane 419

- silicon frame 431
- silver staining 443 ff.
- single resist process 429 f.
- solid-phase synthesis 429
- spot density 420
- spotters 419
- statistical methods 412
- streptavidin 413
- surface acoustic wave (SAW) 434
- surface plasmon resonance imaging 434
- SYBR® green II 436
- TAMRA 437
- target amplification 437
- television technology 428
- tetramethyl rhodamine 437
- Texas Red 437
- two-colour detection system 413
- unspecific hybridisation 415
- vaporisation 422
- variance analysis 412
- XNA on Gold™ 425 f.
- DNA-binding dyes 346 ff., 436
- DNA-binding protein 25
- dot blot 167 ff., 171
- DOT, *see* diffuse optical tomography
- double-focusing sector 641
- double restriction site primer 240
- DPF, *see* Dragon Promoter Finder
- Dragon Promoter Finder (DPF) 810
- driver 178
- DROP, *see* directed random oligonucleotide priming
- dry ice 141
- DSC, *see* differential subtraction chain
- DSD, *see* differential subtraction display
- DSMZ, *see* Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH
- DsRed, *see* *Discosoma* species red fluorescent protein
- dynamic thresholding algorithm 816
- e**
- EBI, *see* European Bioinformatics Institute
- EBP, *see* European Bacteria Project
- ECLIA, *see* electrochemiluminescence immunoassay
- ECO2DBASE 854
- Edinburgh mouse atlas 826
- Edinburgh mouse atlas gene expression database (EMAGE) 840 ff.
- Edinburgh Mouse Atlas Project (EMAP) 840 f.
- EDS, *see* enzymatic degrading subtraction
- EDTA 142
- EGAD, *see* Expressed Gene Anatomy Database
- electrical/electrochemical detection 453 ff.
 - amperometric affinity assay 454
 - capacitance 453
 - competitor probes 455
 - conductivity 455
 - current signal 453
 - cyclic voltammetry 454
 - direct detection 453
 - direct label-free electrical detection 454
 - electron-conducting redox hydrogel 454
 - gold nanoparticles 455
 - horseradish peroxidase 455
 - intercalators 453
 - magnetic beads 455
 - peroxidase 454
 - photo-electrochemical method 455
 - photo-electrochemistry 455
 - photocurrent 455
 - redox indicator 453
 - reporter deposition amplification 455
 - sensitivity 453
 - silver deposition 455
 - transducer 453
- electro capture 416, 423 ff.
 - active programmable electronic device 423
 - agarose layer 424
 - aluminium 423
 - biotin 424
 - biotinylated molecules 423
 - dielectric layer 424
 - electric silicon chip 424
 - immobilisation 423
 - permeation layer 424
 - photoresist 423
 - platinum electrodes 423–424
 - resist 423
 - silicone dioxide 424
 - streptavidin 424
 - streptavidin/biotin interaction 424
 - thermic oxidation 423
 - trisilicone tetranitride 423–424
- electroblotting 647
- electrochemical focussing 423
- electrochemical immunoassay 679, 682
- electrochemiluminescence immunoassay (ECLIA) 679, 683
- electroendoosmotic flow (EOF) 635
- electronic subtraction 206
- electrospray ionisation (ESI) 456, 640, 659
- electrospray ionisation tandem mass spectrometry (ESI-MS/MS) 647, 650
- ELISA, *see* enzyme-linked immunosorbent assay

- ELK1 28
- ellipsometry 667, 669
- elongation 3, 7
- elongation factor 7
- elongation of transcription 33 ff.
- *Drosophila* hsp70 36
 - arrest 34
 - bacteriophage λ Q protein 35
 - *c-myc* 35
 - eukaryotes 35
 - human immunodeficiency virus genes 35
 - pausing 34
 - prokaryotes 34
 - termination 34
 - tryptophan operon 34
- EMAGE, *see* Edinburgh mouse atlas gene expression database
- EMAP, *see* Edinburgh Mouse Atlas Project
- EMBL, *see* European Molecular Biology Laboratory
- embryonic stem cell database 862
- emission spectra 436
- EMIT, *see* enzyme-multiplied immunoassay technique
- EMMIA, *see* enzyme-modulator mediated immunoassay
- 3'-end labelling 439
- Endocrine Pancreas Consortium Database (EPConDB) 842, 857
- enhanceosome 25
- enhancer 21, 798, 800
- enhancer blockade 24
- enhancer-binding protein E2 22
- EnsEMBL 853
- Entrez genomes website 851
- Entrez-style 830
- enzymatic degrading subtraction (EDS) 192 ff.
- driver 193–194
 - exonuclease III 193–194
 - exonuclease VII 193
 - partial Klenow digestion 193
 - PCR amplification 193
 - PERT, *see* phenol emulsion reassociation technique
 - phenol emulsion reassociation technique 194
 - principle 194
 - restriction enzyme digestion 192
 - subtractive cDNA library 192
 - subtractive hybridisation 193
 - tester 193–194
- enzymatic digestion 137
- enzyme immunoassay 679
- enzyme-linked immunosorbent assay (ELISA) 667, 679, 683
- β -galactosidase 680
 - *p*-nitrophenyl phosphate 680
 - 2,2-azino-bis-[3-ethyl-benzthiazoline] sulphate 680
 - 4-umbelliferyl phosphate 680
 - alkaline phosphatase 679–680
 - capture ELISPOT 680
 - chemiluminescence 680
 - chlorophenol red- β -D-galactopyranoside 680
 - direct substrate 680
 - enzymatic cascade 680
 - enzyme immunoassay 679
 - fluorescence 680
 - glucose oxidase 680
 - horseradish peroxidase 680
 - immunoenzymometric assay 679
 - indirect substrate 680
 - microperoxidase 680
 - photoluminescence 680
 - principle 683
 - radioactivity 680
 - urease 680
- enzyme-modulator mediated immunoassay (EMMIA) 676
- enzyme-multiplied immunoassay 677
- enzyme-multiplied immunoassay technique (EMIT) 676 f.
- enzyme-sensing, activatable NIRF probes 756 ff.
- *Herpes simplex* virus protease 757
 - caspase 1 757
 - caspase 3 757
 - cathepsin B 757
 - cathepsin D 757
 - cathepsin K 757
 - enzymatic cleavage 756
 - fluorescence signal 756
 - human immunodeficiency virus protease 757
 - lysosomal enzymes 757
 - matrix metalloproteinases 757
 - nuclear factor κ B (NF κ B) 757
 - protease-sensing NIRF contrast agents 756
 - quenching/dequenching mechanism 756
- EOF, *see* electroendosmotic flow
- EPConDB, *see* Endocrine Pancreas Consortium Database
- EPD, *see* Eukaryotic Promoter Database
- epichlorohydrin 725
- CLIO-NH₂ 725
 - crystalline Fe₂O₃/Fe₃O₄ 725

- dextran cross-linked 725
 - epigenetics 29
 - eSAGE 556
 - ESI, *see* electrospray ionisation
 - ESI-MS/MS, *see* electrospray ionisation tandem mass spectrometry
 - EST, *see* expressed sequence tag
 - ESTscan 774
 - estwisedb (Wise2 package) 772, 773
 - ethidium bromide 138, 345
 - eukaryotic promoter 20f.
 - Eukaryotic Promoter Database (EPD) 20, 801, 808 ff.
 - European Bacteria Project (EBP) 865
 - European Bioinformatics Institute (EBI) 651, 778, 779, 785, 826
 - European Molecular Biology Laboratory (EMBL) 832, 849, 854, 869
 - evaluation process 782
 - evanescent field spectrometry 667
 - 5'-3' exonuclease activity 349 ff.
 - ExPASy, *see* Expert Protein Analysis System
 - Expert Protein Analysis System (ExPASy) 624, 651, 785, 855
 - Expressed Gene Anatomy Database (EGAD) 857
 - expressed sequence tag (EST) 220, 332, 410, 432, 469, 487, 547, 555, 571, 575, 586, 770, 771, 773, 819, 825, 829, 830, 844
 - expressed sequence tag libraries 830
 - Expressionist® 784
 - ExProfile system 828
 - Extensible Markup Language (XML) 781
 - external standard 63 ff.
- f**
- FACS, *see* fluorescence activated cell sorter
 - FAM 437
 - FASTA 773, 821
 - FastM 806
 - FDA, *see* Food and Drug Administration
 - FDD, *see* fluorescent mRNA differential display
 - Featurama 433
 - FIA, *see* fluorescence immunoassay
 - FirstEF 810
 - FISH, *see* fluorescence *in situ* hybridisation
 - Fisher exact test 831
 - FITC, *see* fluorescein isothiocyanate
 - FLAG-tag 15
 - Flanking sequences 804 f.
 - score of context 805
 - artificial intelligence 804
 - contextual motifs 805
 - perception method 804
 - FLIM, *see* fluorescence lifetime imaging
 - flow cells 109
 - closed cross flow 109
 - flow-through cuvette 109
 - jet-in-air 109
 - open flow across surface 109
 - flow cytometer, *see* flow cytometry
 - flow cytometry 107 ff., 250, 687
 - 4',6-diamidino-2-phenylindole 688
 - abort logic 118
 - cell populations 692
 - cell sorters 116, 689
 - cell sorting 107 ff., 116
 - CFP, *see* cyan fluorescent protein
 - Coulter Counter® 108
 - Cy7™ 693
 - cyan fluorescent protein (CFP) 689 ff.
 - cytometer configuration 689
 - DAPI 688
 - data visualisation 690
 - dead-time 114
 - detection table 113
 - drop delay 117
 - FACS, *see* fluorescence activated cell sorter
 - false-colour dot plot 690
 - FITC, *see* fluorescein isothiocyanate 688
 - flow cells 109
 - fluorescein isothiocyanate (FITC) 688
 - fluorescence activated cell sorter (FACS) 108
 - fluorescence compensation 690
 - forward scatter (FSC) 693
 - FSC, *see* forward scatter
 - gating 689
 - GFP, *see* green fluorescent protein
 - green fluorescent protein (GFP) 687 ff., 753
 - hydrodynamic focusing 111
 - interrogation table 113
 - photodiode 111
 - photomultipliers 112
 - phycoerythrin 688, 692, 693
 - principle 687
 - protein expression analysis 687
 - Scripps Research Institute 114
 - side scatter (SSC) 693
 - sorting logic 118 f.
 - spectral overlap 115
 - spectrum viewer 114
 - SSC, *see* side scatter
 - surface antigen expression 692
 - Texas Red 688
 - thermionic effect 112
 - flow-injection analysis 439

- Fluor X 437
- fluorescein isothiocyanate (FITC) 437, 688
- fluorescence 442
- fluorescence *in situ* hybridisation (FISH) 470
- fluorescence *in vivo* hybridisation (FIVH) 716 ff.
 - 2'-O-methyl-RNAs 716 f.
 - peptide nucleic acids 716 f.
 - phosphorothioate oligodeoxynucleotides 716 f.
 - probe selection 716
 - ribosomal RNA (rRNA) 716
 - RNase H cleavage 716
 - rRNA, *see* ribosomal RNA
 - small nuclear RNA (snRNA) 716
 - snRNA, *see* small nuclear RNA
- fluorescence activated cell sorter (FACS) 108, 366
- fluorescence immunoassay (FIA) 667, 679, 682
- fluorescence *in situ* hybridisation 470
- fluorescence lifetime assays 667
- fluorescence lifetime imaging (FLIM) 720
- fluorescence mediated tomography (FMT) 753
- fluorescence mRNA differential display 494
- fluorescence planar waveguide (FPWG) 666–667
- fluorescence polarisation immunoassay (FPIA) 667, 676 f.
- fluorescence resonance energy transfer (FRET) 349 ff., 477, 666–667, 720, 756
- fluorescent labels 435
 - *in vitro* transcription (IVT) 435–436
 - Alexa 437
 - biotin 436
 - BODIPY® 630/650 437
 - cDNA synthesis 435
 - Cy3TM 436–437
 - Cy5TM 436–437
 - DNA-binding dyes 436
 - FAM 437
 - FITC, *see* fluorescein isothiocyanate
 - Fluor X 437
 - fluorescent dyes 437
 - Hoechst 33258 436
 - indirect fluorescence 437
 - JOE 437
 - PCR amplification 435
 - radioactivity 437
 - random primed labelling 435
 - ROX 437
 - signal amplification 437
 - SYBR® green II 436
 - TAMRA 437
 - target amplification 437
 - tetramethyl rhodamine 437
 - Texas Red 437
- fluorescent microbeads, *see* suspension arrays
- fluorescent mRNA differential display (FDD) 177, 219, 237, 319, 326, 340, 343, 358, 494–495
 - arbitrary primer 495, 497–498
 - automation 495
 - band identification 497
 - capillary array gel electrophoresis 497
 - cDNA cloning 495
 - cross-hybridisation 499
 - denaturing polyacrylamide gel electrophoresis 496
 - digital data acquisition 497
 - digital data analysis 497
 - digital data presentation 497
 - digital data quantification 497
 - digital readout 495
 - DNA sequencing 495
 - electropherogram 498
 - false-positives 498
 - fluorescent image scanning 496
 - gel electrophoresis 495
 - high throughput 497
 - PCR amplification 496
 - polyacrylamide gel 497
 - principle 495
 - reproducibility 498–499
 - restriction enzyme digestion 498
 - robotic liquid dispenser 498
 - RT-PCR 495
 - sensitivity 497, 499
 - signal detection 499
- fluorimetry 667
- FMT, *see* fluorescence mediated tomography
- µFN, *see* micro-fluidics network
- Food and Drug Administration (FDA) 739
- Förster resonance energy transfer (FRET) 349 f.
- Formula Transistor (FORTRAN) 786
- FORTRAN, *see* Formula Transistor
- forward scatter (FSC) 112, 688, 693
- Fos 27 f.
- fourier transform-ion cyclotron resonance (FT-ICR) 641
- FPIA, *see* fluorescence polarisation immunoassay
- FPWG, *see* fluorescence planar waveguide
- fractionation of cell extracts 145
 - equilibrium sedimentation 145

- velocity sedimentation 145
 - framefinder 773–774
 - free induction decay 733
 - FRET, *see* fluorescence resonance energy transfer
 - FSC, *see* forward scatter
 - FT-ICR, *see* fourier transform-ion cyclotron resonance
 - functional categorisation 790
 - functional genomics 72
 - FunSite 805
 - FunSiteP 811
- g**
- G3PDH, *see* glyceraldehyde-3-phosphate dehydrogenase
 - gap-statistic 797
 - GAPDH, *see* glyceraldehyde-3-phosphate dehydrogenase
 - gapped BLAST 773
 - GC box 4
 - GC-rich DNA region 29
 - GDB, *see* Genome Database
 - GEF, *see* gene expression fingerprinting
 - gel filtration 633
 - GenAtlas 832
 - GenBank 575, 785, 830, 832, 836, 838, 840, 844, 849, 854, 858
 - Gencarta™ 833
 - gene chips 585
 - gene density 25
 - gene expression 811 ff.
 - spatio-temporal patterns 811
 - three-dimensional visualisation 811 ff.
 - gene expression database 819
 - data format 821
 - indispensable items 823
 - relational database management system 822
 - types of data 822
 - Gene Expression Database (GXD) 835, 840
 - Gene Expression Database for the Molecular Pharmacology of Cancer 857
 - gene expression fingerprinting (GEF) 302
 - 2D-GEF 307 f.
 - biotin 302
 - cDNA restriction fragments 302, 304
 - computer simulation 306–307
 - computer-assisted analysis 305
 - database 306
 - efficient resolution 304
 - first dimension 304
 - gel electrophoresis 302 f., 305
 - high resolution 307
 - immobilisation 303
 - nested PCR 303
 - PAGE, *see* polyacrylamide gel electrophoresis
 - PCR amplification 303
 - polyacrylamide 306
 - polyacrylamide gel electrophoresis 302, 304, 305
 - primary restriction digestion 303
 - principle 302 ff.
 - protocol 303
 - reproducibility 304, 307
 - restriction digestions 306–307
 - restriction endonuclease 302
 - restriction enzyme 303–305, 307
 - second dimension 304
 - secondary restriction fragment subsets 304
 - sequential liberation 302
 - sequential restriction 303
 - streptavidin 302, 303–304, 305
 - sub-database 306
 - two-dimensional coordinates 306–307
 - two-dimensional picture 305
 - two-dimensional separation 306
 - Gene Expression Omnibus (GEO) 834
 - gene expression screen (GES) 196
 - biotinylated driver cDNA 198
 - cDNA libraries 199
 - cloning 199
 - cross hybridisation 200
 - enriched cDNA libraries 199
 - enriched cDNA populations 200
 - hybridisation 197
 - linker ligation 197–198
 - PCR amplification 197–199
 - PCR-based subtraction 200
 - PCR-based subtractive hybridisation 196
 - plasmid vector 198
 - principle 198
 - restriction enzyme digestion 197–199
 - Southern blot 199
 - subtractive hybridisation 198–199
 - suppression 198
 - Gene Library Summarizer (GLS) 98, 858
 - Gene Ontology 773–774, 835, 845, 853
 - gene separation 24
 - gene shaving 796 f.
 - algorithm 797 f.
 - concept of variance 796 f.
 - gene signature 826
 - GeneCalling® 501
 - adapter ligation 502
 - adapter-specific primers 502
 - adapter 502
 - biotin 502

- capillaries 502
- capillary electrophoresis 503
- electrophoresis 502
- flexibility 505
- linkers 503
- PCR amplification 502
- peak quality 502
- peak resolution 502
- polyacrylamide gel 502
- principle 502
- reproducibility 502
- restriction endonuclease digestion 502
- restriction enzymes 503
- restriction fragment 504
- GeneCards 856
- GeneEMAC 811 ff., 814
 - automatic congruencing 812 ff.
 - computer-driven realignment 814
 - digital enlargement capability 818
 - digital image processing 815
 - drill holes 814–815
 - dynamic thresholding algorithm 816
 - dynamic threshold segmentation 812 ff.
 - embedding 814
 - expression maps 813
 - external marker 814
 - high-resolution analysis 818
 - histological sectioning 815
 - image capturing 812 ff., 815
 - image congruencing 816
 - image segmentation 816
 - internet database 819
 - landmark 814
 - macro-driven image processing 816
 - macro 814
 - quantification 819
 - realignment 815
 - reconstruction 817–818
 - reference image 815
 - resin block 814
 - resolution 818
 - serial histological section 813
 - serial section 818
 - specimen preparation 812 ff.
 - stereoscopic pairs of colour images 817
 - surface generation 812 ff.
 - theoretical mathematical model 819
 - three-dimensional model 816
 - three-dimensional visualisation 811 ff.
 - visualisation 812 ff.
 - whole-mount *in situ* hybridisation (WISH) 814
 - WISH, *see* whole-mount *in situ* hybridisation
 - whole-organism screens 817 f.
- GeneMap'99 832
- GeneNest 772 ff., 777
- GenePix 846
- general transcription factor 799
- GeneSpring™ 536, 784
- genetic algorithm 807, 809
- genetic network analyses 790
- genetic networks 819
- Genexpress IMAGE Knowledge Base of the Human Genome and Transcriptomes 832 f.
- Genexpress Index 832
- Genomax 491
- genome annotation 13
- Genome Database (GDB) 832
- GenomeNet 778
- Genomic Unified Schema (GUS) 843
- GenPept 652
- Genzyme Molecular Oncology SAGE Database 857
- GEO, *see* Gene Expression Omnibus
- German Cancer Research Center, *see* Deutsches Krebsforschungszentrum (DKFZ)
- German Collection of Microorganisms and Cell Cultures, *see* Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH
- German Human cDNA Project 575
- German Human Genome Project (DHGP) 575
- German Resource Center Primary Database, *see* Deutsches Ressourcenzentrum für Genomforschung GmbH (RZPD)
- GES, *see* gene expression screen
- GFP, *see* green fluorescent protein
- glass slides 411
- glassware 135
- global mRNA amplification 363
- GLS, *see* Gene Library Summarizer
- glyceraldehyde-3-phosphate dehydrogenase (GAPDH) 65 f.
- glycosylation 56 f., 656 f., 660
 - dolichol 57
 - glycoprotein 56
 - N-glycosidic 56
 - N-linked glycosylation 56
 - O-glycosidic 56
 - O-linked glycosylation 56
 - proteoglycan 56
- gold 419
- goodness-of-fit statistic 797
- Graphical User Interface (GUI) 778
- grating coupler 667, 669

- green fluorescent protein (GFP) 687–689, 753
- gRNA, *see* trans-acting guide RNA, trypanosome
- guanidine hydrochloride 138–141
- guanidine isothiocyanate 138–141
- GUI, *see* Graphical User Interface
- guilt-by-association concept 790
- GUS, *see* Genomic Unified Schema
- GXD, *see* Gene Expression Database
- h**
- Haemophilus* 2D protein database 862
- half-life 11
- HAT, *see* histone-acetyl transferase
- Heart High-Performance 2-DE Database (HP-2DPAGE) 862
- Heart Science Centre 2D-PAGE Database (HSC-2DPAGE) 854
- HEART-2DPAGE, *see* Human Myocardial Two-Dimensional Electrophoresis Protein Database
- heparin 142
- hetero-clusters 807
- heterogeneous assays 676 ff., 681
- heterogeneous nuclear riboprotein (hnRNP) 44
- HGMP, *see* Human Genome Mapping Project
- hidden Markov model (HMM) 804, 807
- hidden Markov model profiler (HMMER) 772, 773
- hierarchical clustering 491 f., 492, 793 f., 846
 - algorithm 793
 - average linkage 793 f.
 - complete linkage 793 f.
 - single-linkage 793 f.
 - update parameters 794
- high-density microarrays 411 f.
- high-performance liquid chromatography (HPLC) 298–299, 624, 630 ff., 649, 659
 - affinity chromatography 631
 - column packing materials 630
 - columns 630
 - HPLC column 630
 - ion exchange chromatography (IEC) 631
 - mass spectrometry 630
 - micro-columns 630
 - mobile phase 630
 - normal phase chromatography 631
 - reverse phase chromatography (RPC) 631
 - size exclusion chromatography (SEC) 631
 - solid phase 630
 - ultraviolet absorption, fluorescence 630
- high-resolution imaging 725
- high-resolution microscope 436
- high-throughput imaging 727
- histochemical techniques 713 f.
 - applications 714
 - automated systems 713
 - combined ISH and ICC 713
 - dual staining 713
 - electron microscopy 713
 - quantification 714
 - triple staining 713
 - whole embryo mounts 714
- histogram distributions 846
- histone-acetyl transferase 23
- HMM, *see* hidden Markov model
- HMMER, *see* hidden Markov model profiler
- hnRNP, *see* heterogeneous nuclear riboprotein
- Hoechst 33258 436
- homo-clusters 806 f.
- homogeneous assays 676 ff.
- homogenisation 135 f.
- homogeniser 136
 - bead mill 136
 - enzymatic 136
 - mortar and pestle 136
 - rotorstator 136
 - spin-column 136
 - vortexing 136
- HomoloGene 775
- Homology-Derived Structures of Proteins (HSSP) 853, 861
- housekeeping gene 63, 358, 412
- HP-2DPAGE, *see* Heart High-Performance 2-DE Database
- HPI, *see* Human Protein Index
- HPLC, *see* high-performance liquid chromatography
- HPLC-MS 659
- HPLC-MS/MS 649
- HSC-2DPAGE, *see* Heart Science Centre 2D-PAGE Database
- HSSP, *see* Homology-Derived Structures of Proteins
- HTML, *see* Hypertext Markup Language 781
- Hubert's I statistic 797
- Human and Mouse 2-D PAGE Databases 863
- Human Colon Carcinoma Protein Database 862
- Human Genome Mapping Project (HGMP) 777
- Human Mitochondrial Genome Database (MITOMAP) 778

- Human Myocardial Two-Dimensional Electrophoresis Protein Database (HEART-2DPAGE) 863
- Human Protein Index (HPI) 851
- HumanPSD™ 866
- hybrid selection 139
- hybridisation 710
- incubation conditions 710
 - melting temperature 710
 - stringency 710
- hybridisation *in situ* (ISH) 439, 704, 707, 820, 841
- control 710
 - control cell 710
 - control probe 710
 - control tissue 710
 - optimisation 710
 - tissue localisation 708
 - total mRNA 710
- hybridisation *in vivo* 745
- hybridisation probes 350 f.
- acceptor fluorophore 350
 - flexibility 350
 - fluorescein donor 350
 - fluorescence resonance energy transfer 350
 - head-to-tail arrangement 350
 - probe design 350
 - quencher, *see* acceptor fluorophore
 - specificity 350
- hyperacetylation 23
- Hypertext Markup Language (HTML) 781
- hypoxanthine phosphoribosyltransferase 66 f.
- i**
- iAFLP, *see* introduced amplified fragment length polymorphism
- IBDD, *see* indexing-based differential display
- ICAT™, *see* isotope-coded affinity tag
- ICC, *see* immunocytochemistry
- IEC, *see* ion exchange chromatography
- IEF, *see* isoelectric focusing
- IEF, *see* isoelectric focussing
- IGSS, *see* immunogold/silver-staining
- IHC, *see* immunohistochemistry
- IMAC, *see* immobilised metal ion affinity chromatography
- IMAGE consortium, *see* Integrated Molecular Analysis of Genomes and their Expression Consortium
- image fusion 752
- image-localised spectroscopy 734
- imaging reporter gene 742
- immobilised metal ion affinity chromatography (IMAC) 656
- immobilised pH gradient isoelectric focusing (IPG-IEF) 626
- immuno-polymerase chain reaction 684
- immunoassay 671 ff.
- *p*-aminophenyl phosphate 682
 - D-luciferin 682
 - apoenzyme-reactivation immunoassay (ARIS) 676
 - ARIS, *see* apoenzyme-reactivation immunoassay
 - automation 671
 - binding agents 671
 - bioluminescence 682
 - bioluminescence immunoassay 679
 - biotin 676
 - BODIPY® 682
 - boradiazaindacene 682
 - calibration curve 675
 - capture ELISA 683
 - capture enzyme-linked immunosorbent assay (capture ELISA) 683
 - CEDIA, *see* cloned enzyme donor immunoassay
 - chemiluminescence 678, 680, 682
 - chemiluminescence immunoassay 679
 - chimeric detection antibody 681
 - cloned enzyme donor immunoassay 676
 - competitive 674 f., 682
 - coumarins 682
 - coupled detection reaction 681
 - cross-reactivity 685
 - cyanine 682
 - detection limits 682
 - dipyrrometheneboron difluoride 682
 - direct capture enzyme 681
 - direct immunometric-enzyme 681
 - direct label 682
 - direct signal generation 681
 - ECLIA, *see* electrochemiluminescence immunoassay
 - electrochemical immunoassay 679, 682
 - electrochemiluminescence immunoassay (ECLIA) 679, 683
 - ELISA, *see* enzyme-linked immunosorbent assay
 - EMIT, *see* enzyme-multiplied immunoassay technique
 - EMMIA, *see* enzyme-modulator mediated immunoassay
 - enzymatic cleavage 682
 - enzyme-linked immunosorbent assay (ELISA) 679 f.

- enzyme-modulator mediated immunoassay (EMMIA) 676
- enzyme-multiplied immunoassay technique (EMIT) 676 f.
- false-high result 686
- false-low result 686
- false-negative 685
- false-positive 685
- FIA, *see* fluorescence immunoassay
- flow cytometry 682
- fluorescein 682
- fluorescence 680, 682
- fluorescence immunoassay (FIA) 679
- fluorescence polarisation immunoassay (FPIA) 676 f.
- fluorescence quenchers 678
- fluorescence-activated cell sorting 682
- fluorophores 678
- FPIA, *see* fluorescence polarisation immunoassay
- heterogeneous assay 676 ff., 681
- high-dose hook effect 684
- highly-specific labels 674
- homogenous assay 676 ff.
- immuno-polymerase chain reaction 684
- immunometric immunoassay 674 f.
- indirect immunometric-enzyme 681
- interference 684, 685, 686
- isotopes 678
- LAA, *see* latex agglutination assay
- latex agglutination assay (LAA) 678
- latex particles 678
- LIA, *see* luminescence immuno assay
- ligand-binding assay 679
- LILA, *see* liposome immune lysis assay
- liposome immune lysis assay (LILA) 676
- liposomes 678
- luciferase 682
- nephelometry 678
- nephelometry turbidimetry 679
- non-competitive immunoassay 674 f., 682
- open-sandwich bioluminescent-immunoassay (OS-BLIA) 676 ff.
- OS-BLIA, *see* open-sandwich bioluminescent immunoassay
- oxazine 682
- particle-enhanced 678
- photoluminescence 680
- principles 675
- radioactivity 680
- radioimmunoassay (RIA) 678–679
- reagents 671
- redox compounds 678
- resorufin 682
- rhodamine 682
- RIA, *see* radio immunoassay
- sandwich-assays 674 f.
- sensitivity 674
- SLFIA, *see* substrate-labelled-fluorescent immunoassay
- specificity 674
- streptavidin 676
- substrate-labelled-fluorescent immunoassay (SLFIA) 676
- surface plasmon resonance (SPR) 678
- turbidimetry 678
- types of label 678
- immunoblotting, *see* Western blotting
- immunocytochemistry (ICC) 704 ff., 711 ff.
 - assay kits 711
 - phage display antibody libraries 712
 - primary antibodies 711
 - principle 712
 - sensitivity 711
 - species cross-reactivity 711
 - specificity 711
- immunoenzymometric assay 679
- immunogold/silver-staining (IGSS) 443
- immunohistochemistry (IHC) 704 ff., 711 ff.
 - assay kits 711
 - phage display antibody libraries 712
 - primary antibodies 711
 - principle 712
 - sensitivity 711
 - species cross-reactivity 711
 - specificity 711
- immunometric immunoassays 674 f.
- immunoprecipitation 147 ff.
 - 2-mercaptoethanol 147
 - antibody/antigen complex 147, 149
 - dithiothreitol (DTT) 147
 - DTT, *see* dithiothreitol
 - IgG subclasses 149
 - protein A 148–149
 - protein G 148–149
 - protein G sepharose 147–148
 - SDS, *see* sodium dodecylsulphate
 - SDS-based polyacrylamide gel electrophoresis (SDS-PAGE) 147
 - SDS-PAGE, *see* SDS-based polyacrylamide gel electrophoresis
 - sepharose 147
 - sodium dodecylsulphate (SDS) 147
 - Western blotting 147
- immunoprecipitation of cells 99
- immunopurification 15
- immunopurification of cells 99 ff.
 - antibody/antigen reactions 99

- antibody/cell complex 101
- cell population 100
- chymopapain 100
- competing molecule 100
- direct method 100 f.
- disruption of a linker interaction 100
- Dynabeads 100
- enzymatic cleavage 100
- heterogenous cell population 99 ff.
- homogenous cell population 99 ff.
- immunomagnetic purification 99 ff.
- immunoprecipitation of cells 99 ff.
- indirect method 100 f.
- magnetic beads 101
- negative cell isolation 100 ff.
- negative isolation 102
- positive cell isolation 100
- positive isolation 102
- pronase 100
- trypsin 100
- immunosensors 678
- immunohistochemistry 439
- in-gel-digest 648
- indexing-based differential display (IBDD) 308 ff.
 - adapter 308
 - adapter ligation 311
 - auto-radiography 309
 - cDNA fragments 310, 312
 - cDNA population subdivision 311
 - cDNA populations 308
 - hot start ligation 311
 - hot start PCR 311
 - indexed cDNA fragments 312
 - ligation 308–309
 - molecular indexing 308
 - PCR amplification 312
 - principle 308 ff.
 - restriction enzyme 309
 - restriction fragment 309
- indirect immunometric-enzyme immunoassay 681
- inducible transcription factor 49
- initiation complex 16, 27
- initiation complex formation 17
- initiation factor 7
- initiation of RNA synthesis 50
- INRA Maize Proteome Database 863
- insulator 23 f.
- insulin 53
- integrated data 833
- Integrated Molecular Analysis of Genomes and their Expression (IMAGE) 832, 777, 830
- interactive QuickTime™ VR models 817
- interactome 14
- interferometric methods 434
- Interferon-Stimulated Gene Database (ISG Database) 836
- internal standard 63 ff.
- internet databases 819
- InterPro 773–774, 853, 861
- intracellular localisation of RNA 40 ff., 44
 - actin-dependent transport 42 f.
 - active transport 42 f.
 - *cis*-acting elements 44
 - diffusion 42
 - local anchoring 42
 - localised degradation 42
 - localised synthesis 42
 - localised transcription 42
 - multistep localisation 42 f.
 - *trans*-acting factors 44
- intracellular RNA localisation, *see* intra cellular localisation of RNA
- intravital confocal and multiphoton microscopy 753
- introduced amplified fragment length polymorphism (iAFLP) 280, 506, 826
 - adapter 506, 507
 - amplification efficiencies 506
 - amplification kinetics 506
 - automatic sequencer 506
 - competitor 506
 - cross-hybridisation 509
 - electropherogram 508
 - elongase activity 509
 - iAFLP profile 506
 - ligation 506–507
 - non-specific priming 509
 - principle 506 f.
 - relative quantification 506
 - restriction enzyme digestion 507
 - sequencing gel 507, 506 ff.
- Invader® RNA assay 471, 476
 - FRET probe 477
- ion exchange chromatography (IEC) 624, 631, 633, 635
 - principle 632
 - SCX, *see* strong cation-exchange
 - stationary phase 633
 - strong cation-exchange (SCX) 633
- ion paired-reverse phase-high performance liquid chromatography (IP-RP-HPLC) 296
- ion separation 641
- ion trap 641
- IP-RP-HPLC, *see* ion paired-reverse phase-high performance liquid chromatography

IPG-IEF, *see* immobilised pH gradient isoelectric focusing

ISG Database, *see* Interferon-Stimulated Gene Database

ISH probe 705

- amplification of the signal 711
- amplification of the target 711
- antisense oligodeoxynucleotide 705
- biotin 708
- cDNA 705
- detection methods 707 f.
- digoxigenin 708
- direct fluorescence 708
- enzymatic amplification 711
- fluorochromes 708
- labelling methods 707 f.
- non-isotopic labels 708
- PCR-amplified ISH 711
- peptide nucleic acid (PNA) 705
- PNA, *see* peptide nucleic acid 705
- probe design 708 f.
- probe label 707 f.
- radioisotope 707 f.
- ribonucleotide probe 705
- riboprobe 705
- sensitivity 705, 709, 711
- specificity 705, 708

ISH, *see in situ* hybridisation

isoelectric focusing 629

- carrier ampholytes 626
- immobilised pH gradient 626
- isoelectric point 626
- principle 625 f.

isoelectric focusing (IEF) 624 f.

isoelectric point 625–626

isolation of cells, *see* cell isolation

isolation of nucleic acids, *see* RNA isolation

isolation of proteins 133

isolation of RNA

isotope-coded affinity tag (ICAT™) 59 f., 636, 655

Italian Telethon Institute of Genetics and Medicine (TIGEM) 777

j

Jaccard coefficient 797

Jackson Laboratory 826

Java Database Connectivity (JDBC) 781

Java™ 781

JDBC, *see* Java Database Connectivity

JOE 437

JPEG 828

Jun 27 f.

k

K-means 466, 795 f.

- algorithm 796
- euclidean distance 796

K-median clustering 796

K-midranges clustering 796

KEGG 856

Kidney Development Database 837 f.

Kozak sequence 7

l

LAA, *see* latex agglutination assay

lab-on-a-chip 410

laboratory information management system (LIMS) 531

laser, *see* light amplification by stimulated emission of radiation

laser capture microdissection (LCM), *see* microdissection

laser cutting 120 f.

laser microdissection (LMD), *see* microdissection

laser microdissection ultraviolet cut

laser pressure catapulting (LPC) 121 ff., 128

laser scissors 125, 128

latex agglutination assay (LAA) 678

2D-LC, *see* two-dimensional liquid chromatography

LCM, *see* laser capture microdissection

LCR, *see* locus control region

LCS, *see* linker capture subtraction

Lead Exon Database (LEPD) 810

LEADS™ 833

LEPD, *see* Lead Exon Database

LIA, *see* luminescence immunoassay

library browser 844

ligand-binding assay 673, 679

light amplification by stimulated emission of radiation (laser) 109, 120 f.

- ablation 122
- ablative photo-decomposition 120
- laser ablation 122
- near-infrared 126
- nitrogen laser 121
- ultraviolet A 121

LILA, *see* liposome immune lysis assay

LIMS, *see* laboratory information management system

linker capture subtraction (LCS) 201 ff., 234

- clonal redundancy 205
- cloning 202
- digestion 202–203
- driver 203

- enrichment 204
 - hybridisation 202
 - linker 203
 - linker ligation 202
 - mung bean nuclease 202–203
 - PCR amplification 203
 - plasmid 203
 - relative abundance of mRNA 205
 - restriction 203
 - sequencing 202, 204
 - subtractive hybridisation 203, 205
 - subtractive library 203
 - tester 203
 - linker capture subtraction technique 234
 - linker ligation 197
 - lipidation 660
 - liposome immune lysis assay (LILA) 676
 - liquid nitrogen 141
 - lithium chloride 138
 - lithium chloride precipitation 138
 - live-cell catapulting 124, 132
 - live-cell fusions 132
 - live-cell microinjection 132
 - live-cell microsurgery 132
 - LMD, *see* laser microdissection
 - localisation of RNA 40 ff.
 - actin-dependent transport 42 f.
 - active transport 42 f.
 - *cis*-acting elements 44
 - diffusion 42
 - local anchoring 42
 - localised degradation 42
 - localised synthesis 42
 - localised transcription 42
 - multistep localisation 42 f.
 - *trans*-acting factors 44
 - localised magnetic resonance spectroscopy 735
 - locus control region (LCR) 22 f.
 - LocusLink 830–831
 - loop formation 21 f.
 - low-density microarrays 411
 - LPC, *see* laser pressure catapulting
 - luciferase glow assays 667
 - luminescence immunoassay (LIA) 679
 - luminometry 667
 - lysate ribonuclease protection assay 259 f.
- m**
- Maboya Gene Expression Patterns and Sequence Tags Database (MAGEST) 838
 - whole-mount *in situ* hybridisation (WISH) 838
 - Mach-Zehnder interferometer 667, 669
 - MAGE, *see* Microarray Gene Expression
 - MAGE-ML, *see* Microarray Gene Expression Data Markup Language
 - MAGE-OM, *see* Microarray Gene Expression Data Object Model
 - MAGEST, *see* Maboya Gene Expression Patterns and Sequence Tags Database
 - magnetic force microscopy 725
 - magnetic nanosensor 726
 - magnetic resonance 732
 - magnetic resonance imaging (MRI) 724
 - metabolites 735
 - reporter proteins 735
 - magnetic resonance reporter gene 724
 - magnetic resonance spectroscopy (MRS) 732 ff.
 - 5-fluorocytosine 736
 - 5-fluorouracil 736
 - arginine kinase 736
 - creatinine kinase 736 f.
 - crusher gradient 738
 - cytosine deaminase 736
 - human whole-body MRS 732
 - magnetic resonance-visible intracellular marker 737
 - phosphoarginine 736
 - phosphocreatinine 736 f.
 - prodrug conversion 738
 - quantification of nuclei 738
 - relaxation time 738
 - spectral editing techniques 738
 - magnetic resonance-sensitive marker 728 ff.
 - β -galactosidase 730
 - activatable enzyme-sensitive magnetic resonance probe 729 f.
 - activatable magnetic resonance contrast agent 729
 - E-selectin 730
 - enzyme-mediated polymerisation of paramagnetic substrate 729
 - galactopyranose 730
 - imaging of transgene expression 732
 - internalising receptors 731
 - smart magnetic resonance contrast agent 729
 - transferrin receptor 731
 - tyrosinase 728 f.
 - magnetic resonance-visible intracellular marker 737
 - maintenance gene 67 f.
 - major groove 27
 - MALDI, *see* matrix-assisted laser desorption/ionisation
 - MALDI-MS 650

- MALDI-TOF/MS, *see* matrix assisted laser desorption/ionisation time-of-flight mass spectroscopy
- Mammalian Gene Collection (MGC) 571, 576, 858
- manual tissue dissection 127
- MAR, *see* matrix attachment region
- marker metabolites 733
- Mascot 651
- MaskerAid 771, 773
- mass 639
- mass spectrometry (MS) 13, 456 ff., 623, 636, 638, 660, 661, 667
 - 2D-LC, *see* two-dimensional liquid chromatography 649
 - 3-hydroxypicolinic acid 456
 - amino acid masses 645
 - amino acid residues 645
 - amino-terminal ions 644
 - atomic weight 643
 - average mass 643
 - carboxy-terminal ions 644
 - collision 644
 - column packing material 649
 - Coomassie staining 647
 - data analysis tools 651
 - detection of ions 640
 - detection selection 644
 - digestion 648
 - double-focusing sector 641
 - electric field 640
 - electroblotting 647
 - electrospray ionisation (ESI) 456 ff., 640 f.
 - ESI, *see* electrospray ionisation
 - extraction 648
 - fourier transform-ion cyclotron resonance (FT-ICR) 641
 - FT-ICR, *see* fourier transform-ion cyclotron resonance
 - full width at half maximum (FWHM) 642
 - FWHM, *see* full width at half maximum 642
 - glycosylation 656 f.
 - high-performance liquid chromatography (HPLC) 649
 - history 639
 - HPLC, *see* high-performance liquid chromatography
 - HPLC-MS/MS 649
 - ICAT™, *see* isotope-coded affinity tags
 - in-gel-digest 648
 - intact proteins 659
 - ion separation 641
 - ion trap fragmentation spectra 654
 - ion trap 641
 - ionisation 640 f.
 - isotope-coded affinity tags 655
 - MALDI, *see* matrix-assisted laser-desorption/ionisation
 - MALDI-TOF/MS, *see* matrix assisted laser desorption/ionisation time-of-flight mass spectrometry
 - MassExtend™ 457
 - matrix 456, 640–641
 - matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF/MS) 456 ff.
 - matrix-assisted laser-desorption/ionisation 640 f.
 - modes of operation 644, 646
 - monoisotopic mass 642
 - nano-HPLC 649
 - natural abundance 643
 - neutral loss scan 646
 - non-covalent protein complexes 659
 - PAGE, *see* polyacrylamide gel electrophoresis
 - peptide digest 654
 - peptide fragment ions 644
 - peptide map 647, 650, 653
 - peptide sequencing 650
 - phosphorylation 655 f.
 - polyacrylamide gel electrophoresis (PAGE) 648
 - precursor ion scan 646
 - primer extension reaction 457
 - principle 639 ff., 646
 - product ion scan 646
 - protein modifications 655
 - protein tertiary structure 657
 - proteolytic agents 647
 - proteomics databases 651
 - proteomics servers 651
 - quadrupoles 641
 - quantification 655
 - resolving power 642
 - sequence tags 651
 - short tandem repeats (STRs) 456
 - silver staining 647
 - single nucleotide polymorphism (SNP) 456–457
 - SNP, *see* single nucleotide polymorphism
 - solid-phase extraction 647
 - STR, *see* short tandem repeat
 - thermosequenease 457
 - time-of-flight (TOF) 641
 - TOF, *see* time-of-flight
 - trypsin 646

- two-dimensional liquid chromatography 649
- mass unit 639
- mass-to-charge ratio 639
- MassExtend™ 457
- massively parallel signature sequencing (MPSS™) 538 ff.
 - adapter 543–546
 - anti-tag 539, 541–542
 - biotin 540
 - capture anti-tag 540
 - CCD, *see* charge-coupled device
 - cDNA fragment 539
 - cDNA library 549
 - charge-coupled device (CCD) 542–543
 - data handling 547
 - data mining 548
 - databases 547
 - decoder probes 544
 - digestion 543–546
 - digital format 538, 548
 - digital nature 548
 - encoder 545
 - EST database 548
 - exonuclease 541
 - flow-cell 542–543, 546–547
 - high-throughput 550
 - hybridisation reaction 544
 - initiating adapter 542–543, 544
 - library of microbeads 539
 - ligation 544
 - Megaclone™ 539 ff., 549
 - microbeads 539, 541–543, 547
 - microfluidics network 542
 - monolayer 542
 - one sub-unit mismatch 542
 - one-word mismatch 541
 - palindromes 543
 - perfect match 541–542
 - plasmid vector 541
 - principle 538 ff.–539 ff.
 - recognition site 544
 - sensitivity 538, 547–548
 - sequencing reaction 542–545
 - signature 545
 - signature sequence 538–539, 542, 547 f.
 - streptavidin 541
 - tag library 541
 - tags 539–542
 - two sub-unit mismatch 542
 - two-word mismatch 541
 - type II restriction enzyme 544
 - type II restriction recognition site 543
- Match 803, 807
- MatInspector 803
- matrix-assisted laser desorption/ionisation (MALDI) 640 ff., 653, 661–662
 - matrix 640
 - principle 640, 667, 669
- matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF/MS) 434 f., 456 ff., 470
- matrix attachment region (MAR) 23
- MATRIX SEARCH 803
- McPromoter V3.0 809
- medium-density microarrays 411
- Medline 854
- Megaclone™ 539 ff., 549
 - anti-tags 541–542
 - biotin 540
 - capture anti-tag 540
 - exonuclease 541
 - microbeads 541
 - one sub-unit mismatch 542
 - one-word mismatch 541
 - perfect match 541–542
 - plasmid vector 541
 - streptavidine 541
 - tag library 541
 - tag 540–542
 - two sub-unit mismatch 542
 - two-word mismatch 541
- MELK, *see* multi-epitope ligand-Kartographie
- melt depletion 216
- MeNPOC 450
- Merck Gene Index (MGI) 571, 576
- messenger RNA (mRNA) 2 ff.
- MethDB, *see* DNA methylation database
- methylation 29, 54
- 7-methylguanosine cap 5
- methylmercuric hydroxide 139
- MGC, *see* Mammalian Gene Collection
- MGD, *see* Mouse Genome Database
- MGED, *see* Microarray Gene Expression Database
- MGEIR, *see* Mouse Gene Expression Information Resource
- MGI, *see* Merck Gene Index
- MIAME, *see* Minimum Information about a Microarray Experiment
- micro wet printing (μWP) 417, 431 f.
- micro-contact printing
 - PDMS, *see* polydimethylsilane
 - polydimethylsilane (PDMS) 419
- micro-contact printing (μCP) 416
- micro-fluidics network (μFN) 416, 419 f.
 - antibodies 419
 - glass 419

- N-hydroxysuccinimidyl esters 419
- polystyrene 419
- silicone/silicone dioxide 419
- micro-gravimetry 667
- micro-refractometry 667
- microarray database 846
- Microarray Gene Expression (MAGE) 842
- Microarray Gene Expression Data Markup Language (MAGE-ML) 789
- Microarray Gene Expression Data Object Model (MAGE-OM) 789
- Microarray Gene Expression Database (MGED) 788 ff., 842, 847
- microdissection
 - μ CUT 125
 - ablative photo-decomposition 120
 - laser capture microdissection (LCM) 126, 128
 - laser microdissection (LMD) 125 f., 128
 - laser microdissection ultraviolet cut 125, 128
 - laser pressure catapulting (LPC) 121 ff., 128
 - laser scissors 125, 128
 - LCM, *see* laser capture microdissection
 - live-cell catapulting 124, 132
 - live-cell fusions 132
 - live-cell microinjection 132
 - live-cell microsurgery 132
 - LMD, *see* laser microdissection
 - LPC, *see* laser pressure catapulting
 - manual tissue dissection 127
 - oscillating needle 126
- microPET 743 f.
- microSAGE 556, 557 ff., 568, 570
 - anchoring enzyme 560
 - concatemerisation 562
 - concatemers 560–561, 563
 - concatenation 560
 - digestion 561, 567
 - ditag amplification 562, 567
 - ditag isolation 562
 - ditag 560–563, 567
 - ethanol precipitation 559
 - gel purification 560
 - immobilisation 560
 - ligase 561
 - ligation 560
 - linear amplification 567
 - linker remnants 565
 - linkers 560–561, 563
 - microSAGE profile 558
 - miniaturisation 565
 - PCR amplification 560, 563, 567
 - PCR artefacts 567
 - phenol/chloroform extraction 559
 - polyacrylamide gel 563
 - principle 559 ff.
 - SAGE profile 558, 564
 - sequence analysis 560
 - single-tube procedure 559
 - streptavidin 559, 561
 - tagging enzyme 560–561, 567
 - tag 567
 - template-switching oligonucleotide 565 f., 567
- microtechnology 410
- Minimal Information about a Microarray Experiment (MIAME) 789, 842, 847
- miniSAGE 565, 568 ff., 577
 - anchoring restriction enzyme 569
 - biotin 568–569
 - concatemers 569
 - digestion 569
 - ditag amplification 568
 - ditag 568–569
 - interference 568
 - ligation 568
 - linker 568–569
 - magnetic bead 568
 - modifications 570
 - mRNA capture kit 568
 - PCR amplification 568–570
 - phase lock gel system 568
 - phenol extraction 568
 - principle 568 f.
 - SAGE libraries 569
 - sequencing 569
 - streptavidin 568
 - tagging enzyme 568
 - tagging restriction enzyme 569
- mitochondrial proteome database (MitoPick) 863
- MITOMAP, *see* Human Mitochondrial Genome Database
- MitoPick, *see* mitochondrial proteome database
- molecular beacon 350, 352, 720, 755
 - alternate conformation 350
 - conformational transition 350
 - fluorescence resonance energy transfer (FRET) 756
 - fluorescent marker 350
 - FRET, *see* fluorescence resonance energy transfer
 - hairpin structure 350
 - hybridisation probe 350
 - primer extension 352
 - quencher 350, 720
 - random coil 352

- stem-and-loop structure 350
 - stem-loop motif 720
 - wavelength-shifting molecular beacon 755
 - molecular classification of diseases 791
 - molecular imaging 724
 - molecular indexing 308, 312 ff.
 - adapter 318
 - anchored oligo(dT) primer 313
 - class II restriction enzyme 315
 - class IIS restriction enzyme 313, 315
 - differential hybridisation 317
 - DNA ligase 314–315
 - electropherogram 318
 - indexing adapter 315
 - paramagnetic bead 313
 - principle 313
 - streptavidin 313
 - T4 DNA ligase 315
 - molecular probe 724
 - molecular weight 639
 - Mouse Gene Expression Database 97
 - Mouse Gene Expression Information Resource (MGEIR) 842
 - Mouse Genome Database (MGD) 835
 - Mouse Genome Informatics Project 97
 - MPSS™, *see* massively parallel signature sequencing
 - MRI, *see* magnetic resonance imaging
 - mRNA, *see* messenger RNA
 - MRS, *see* magnetic resonance spectroscopy
 - MS, *see* mass spectrometry
 - MS/MS, *see* tandem mass spectrometry
 - MSP-PCR, *see* multiplex PCR using colour-tagged modul-shuffling primers
 - multi-dimensional separation 629, 635
 - multi-epitope ligand-Kartographie (MELK) 637 f.
 - multi-PACS 338
 - multidimensional separation 625
 - multimerisation 57
 - multiple gain 39
 - multiple jet oligonucleotide synthesis 422 f.
 - multiple loss 39
 - multiple tissue Northern blot 171
 - multiplex PCR using colour-tagged module-shuffling primers (MSP-PCR) 321 ff.
 - adapter ligation 323
 - digestion restriction enzyme 323
 - electrophoresis 323
 - module-shuffling primer 323, 328
 - module 321–322
 - multiplex PCR 326
 - PAGE, *see* polyacrylamide gel electrophoresis
 - PCR amplification 323
 - polyacrylamide gel electrophoresis 324
 - principle 321
 - mutation 456
 - MycoPathDB™ 866
 - MySQL 820
 - MZI, *see* Mach-Zehnder interferometer
- n**
- nano-HPLC 649
 - nanotechnology 410
 - National Cancer Institute (NCI) 830, 844
 - National Center for Biotechnology Information (NCBI) 551, 624, 775, 779, 830, 844, 846, 851, 652
 - National Institutes of Health (NIH) 858
 - NCBI, *see* National Center for Biotechnology Information
 - NCI 2DWG image meta-database 852
 - NCI, *see* National Cancer Institute
 - near-infrared fluorescence 753
 - near-infrared fluorescent imaging (NIRF) 755–758
 - enzymatic cleavage 756
 - DNA/RNA sensing beacons 755
 - enzyme activatable NIRF sensors 755
 - fluorescence signal 756
 - intravenous-injectable fluorescent dyes 755
 - protease-sensing NIRF contrast agents 756
 - quenching/dequenching mechanism 756
 - near-infrared fluorescent optical reporter probe 753
 - near-infrared window 754
 - negative regulatory element 26
 - negative supercoiling 26
 - Nematode Expression Pattern Database (NextDB) 858
 - Nematode Genome Project 820
 - nephelometry 678
 - NEST, *see* nuclear expressed sequence tag analysis
 - network analysis 790
 - networks 819
 - Neural Network Promoter Prediction (NNPP) 809
 - neural networks 804, 809
 - neutral loss scan 645
 - NextDB, *see* Nematode Expression Pattern Database
 - nick translation 439
 - NIH, *see* National Institutes of Health
 - NIH-Image 814
 - NIRF, *see* near-infrared fluorescent imaging
 - NMR, *see* nuclear magnetic resonance

- NMRS, *see* nuclear magnetic resonance spectroscopy
- NNPP, *see* Neural Network Promoter Prediction
- non-competitive assays 674 f.
- Non-Redundant Reference Protein Database (PIR-NREF) 652
- normal phase chromatography 631, 634
- normalisation 69
- Normalisation working group 789
- Northern blot 167 ff., 411, 439–440
- *in vitro* transcription 169
 - 3'-end labelling 169
 - biotin 169
 - capillary blotting 169
 - cDNA synthesis 169
 - chemical synthesis 169
 - chemiluminescence 169
 - cross-linking 169
 - denaturing conditions 167
 - densitometry 171
 - detection limit 172
 - digoxigenin 169
 - electroblotting 169
 - formamide 173
 - gel electrophoresis 167
 - housekeeping gene 171
 - methylene blue 169
 - nick translation 169
 - nitrocellulose 172
 - normalisation 171
 - nylon 172
 - PCR amplification 169
 - probe 169
 - re-probing 173
 - stringency 173
 - stripping 173
- NRE, *see* negative regulatory element
- nuclear expressed sequence tag analysis 177
- nuclear magnetic resonance (NMR) 657
- nuclear magnetic resonance spectroscopy (NMRS) 657, 732
- nuclear run-off assay, *see* nuclear run-on assay
- nuclear run-on assay 173 ff.
- *in vitro* transcription 177
 - acceptor fluorophore 349 f., 355
 - auto-radiography 176
 - biotin 177
 - chemiluminescence 176
 - digoxigenin 176
 - flow cytometric sorting 177
 - housekeeping gene 176
 - hybridisation 175
 - incubation of nuclei 175
 - isolation of nuclei 173, 175
 - nascent nuclear mRNA chain elongation 173
 - principle 174
 - purification of transcripts 175
 - RNA isolation 175
 - RT-PCR 177
 - 5'-nuclease assay 347 ff., 353
 - 5'-3' exonuclease activity 349
 - 5'-nuclease activity 349
 - acceptor fluorophore 349
 - background fluorescence 347
 - cleavage 347, 348
 - fluorescence resonance energy transfer (FRET) 349
 - PCR amplification 347
 - polymerisation 348
 - processivity 349
 - quencher, *see* acceptor fluorophore
 - strand displacement 348
 - TaqMan™ 349
 - threshold cycle 347
 - 5'-nuclease assay (TaqMan™) 347, 349, 353
 - agarose gel 372
 - gel electrophoresis 371, 372
 - principle 370 ff.
 - targeted display primer 370, 372
- nucleic acid isolation 133
- nucleic acid preparation 133
- nucleic acids detection 7 ff., 434 ff.
- AFM, *see* atomic force microscopy
 - atomic force microscopy (AFM) 449
 - alkaline phosphatases 439
 - atomic force microscopy (AFM) 449 ff.
 - biotin 439
 - cDNA synthesis 439
 - chemical modification 435
 - chemical synthesis 439
 - chemiluminescent detection 439 ff.
 - digoxigenin 439
 - direct electrical measurements 434
 - electrical/electrochemical detection 453 ff.
 - electrospray ionisation (ESI) 456 ff.
 - 3'-end labelling 439
 - enzymatic incorporation 435
 - enzymatic modification 435
 - ESI, *see* electrospray ionisation
 - fluorescent labels 435
 - interferometric methods 434
 - *in vitro* transcription 439

- MALDI-TOF/MS, *see* matrix-assisted laser desorption/ionisation time-of-flight mass spectroscopy
- mass spectrometry (MS) 456 ff.
- matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF/MS) 434 f., 456 ff.
- nick translation 439
- PCR amplification 439
- primer labelling 435
- QCM, *see* quartz crystal microbalance
- quantum dots 435
- quartz crystal microbalance dissipation 434
- quartz crystal microbalance (QCM) 459
- radioactive labelling 438 f.
- random primed labelling 439
- reflectometric interference spectroscopy (RIFS) 434, 460 f.
- RIFS, *see* reflectometric interference spectroscopy
- SAW, *see* surface acoustic wave
- silver staining 443 ff.
- SPR, *see* surface plasmon resonance
- surface acoustic wave 434
- surface plasmon resonance imaging 434
- surface plasmon resonance (SPR) 461 ff.
- tailing 435
- 3'-tailing 439
- target amplification 434
- up-converting phosphorus 435
- nucleic acids labelling, *see* nucleic acids detection
- nucleocytoplasmic transport 40
- nucleolus 16
- nucleosome 25
- nucleotide distribution matrix 803
- nylon membrane 411
- o**
- object database 820
- Object-Oriented Transcription Factors Database (ooTFD) 801
- ODBC, *see* Open Database Connectivity
- ODD, *see* ordered differential display
- OFP, *see* oligonucleotide fingerprinting
- OligoArray 433
- oligonucleotide array, *see* oligonucleotide microarray
- oligonucleotide design 432 f.
 - artefacts 432
 - chimeric sequences 432
 - intron contamination 432
 - sequencing errors 432
- oligonucleotide fingerprinting 463
 - applications 468
 - cDNA library 467
 - cluster 464
 - cluster analysis 465, 467
 - clustering procedure 466
 - co-clustering 468
 - data normalisation 466
 - double-ranking method 466
 - grid finding 466
 - high-density DNA arrays 465
 - image analysis 466
 - library construction 464
 - mismatch discrimination 470
 - non-normalised library 469
 - normalisation 468–469
 - normalised cDNA libraries 468
 - oligonucleotide fingerprint 464, 466
 - peptide nucleic acid (PNA) 470 f.
 - pre-selection 469
 - principle 464–465
 - probe selection 464
 - procedure 464
 - random selection 469
 - reassociation kinetics 469
 - representative clone 465, 467
 - Shannon entropy 464
 - technology development 470
 - verification of cluster analysis 467
- oligonucleotide fingerprinting (OFP) 463 ff.
- oligonucleotide microarray 414 f.
 - hybridisation conditions 416
 - mismatch probes 414
 - perfect match probe 414
 - production 416 ff.
 - re-sequencing by hybridisation 415 f.
- oligonucleotide probe sequence design
 - software for genetic analysis 433
- Oligos4Array 433
- OMIM 832, 835, 840, 854
- ontology 827
- ooTFD, *see* Object-Oriented Transcription Factors Database
- open architecture system 74 f., 484, 536
- Open Database Connectivity (ODBC) 781
- open reading frame (ORF) 570
- open reading frame expressed sequence tags (ORESTES) 570 ff.
 - cDNA libraries 574
 - cloning 573
 - computational analysis 573
 - contigs 572
 - DNA contamination 572
 - low-stringency PCR conditions 574

- normalisation effect 574
- open reading frame 572, 573
- poly(A)⁺ tail 573
- polyacrylamide 572
- principle 572 f.
- quality control 573
- selective amplification 572
- sequencing 573
- untranslated region 573
- open-column chromatography 631
- open sandwich bioluminescent-immunoassay (OS-BLIA) 676 f., 677
- open-source 846
- open system, *see* open architecture system
- OPS, *see* overlapping pathway screening
- optical *in vivo* imaging 752 ff.
- Oracle™ 820, 842, 846
- ordered differential display (ODD) 326 ff.
 - 3'-restriction fragments 328, 332
 - adapter 327
 - adapter ligation 329
 - adapter-specific extended primers 329
 - adaptor ligation 327
 - cDNA fragments 328
 - co-migration 330
 - detection limit 330
 - differential *in situ* hybridisation 330
 - exhaustive search 330
 - exponential amplification 327
 - false-positive 328, 330
 - InE-primers, *see* adapter-specific extended primers
 - intentional mismatch trick 329
 - mRNA complexity 329
 - PCR suppression effect 328
 - principle 327 ff.
 - protocol 327
 - representation 330
 - reproducibility 328, 330
 - restriction endonuclease 327
 - selectivity 329
 - suppression of amplification 327
 - technical data 331
- ORESTES, *see* open reading frame expressed sequence tags
- ORF, *see* open reading frame
- organic extraction 137
- organic solvent 135 ff.
- orthologous gene sets 788
- OS-BLIA, *see* open sandwich bioluminescent-immunoassay
- oscillating magnetic field reader 725
- oscillating needle 126
- overlapping pathway screening (OPS) 185

p

- P-site 6 f., 8
- PACS, *see* preferential amplification of coding sequences
- PACS-SuRF, *see* coupled PACS and SuRF
- PAGE, *see* polyacrylamide gel electrophoresis
- 2D-PAGE, *see* two-dimensional polyacrylamide gel electrophoresis
- PAN, *see* poly(A) nuclease
- paper chromatography 631
- paramagnetic lanthanides 727
- paramagnetic metal cation 724
- Parasite Host Cell Interaction 2D-PAGE Database (PHCI-2DPAGE) 854, 863
- parasite proteome maps 864
- parent ion scan 645
- Partial List of Web 2D Electrophoretic Gel Databases 853
- particle-based assays 667
- PatSearch 802
- pattern search 802 f.
 - ambiguous letter code 802
 - consensuses 802
 - false-negative rate 803
 - false-positive rate 803
 - pattern search tools 802
- Pc-G/trx-G, *see* polycomb/trithorax group
- PCR amplification 435, 438–439
- PCR *in situ* 742
- PCR-Lite 568, 570
- PCR-SAGE 556, 565 f., 568, 570, 577 ff.
 - distortions 579
 - exponential phase 579
 - over-amplification 579
 - PCR amplification 578–579
 - PCR distortion 580
 - PCR kinetics 579
 - principle 577 f.
 - restriction enzyme 578
 - SAGE catalogue 577, 579–580
 - SAGE tags 580
 - strand switching 578, 580
- PDB, *see* Protein Data Bank
- PDD, *see* Protein Disease Database
- PEDB, *see* Prostate Expression Database
- peptide map 647
- peptide nucleic acid (PNA) 470 f.
- peptidyl transferase 7
- peptidyl tRNA 8
- peptidyl tRNA-binding site 6
- perceptron 804
- percolation clustering 793
- Perl 781, 846

- PERT, *see* phenol emulsion reassociation technique
- PET reporter gene 743
- virus thymidine kinase 743
- PET reporter probe 743
- PET, *see* positron emission tomography
- phagemid subtractive hybridisation 181
- phagemid vector 181
- PHCI-2DPAGE, *see* parasite host cell interaction 2D-PAGE database
- phenol 137
- phenol emulsion reassociation technique (PERT) 194
- PhosphoProtein Database (PPDB) 864
- phosphorylation 54, 655 f., 660
- kinase 54
 - phosphatase 54
- photolithography 417, 426 ff., 428 f.
- β -elimination 428
 - *in situ* synthesis 426
 - chrome/glass masks 426
 - MeNPOC 427–428
 - NPPOC 428
 - photo-activation 426
 - photolithographic masks 428
 - photosensitive protective layer 426
 - PYMOC 427–428
 - ultraviolet light 426
- photoresist wet masking 429 ff.
- bilayer resist procedure 430
 - phosphoramidite 432
 - photolithographic mask 429
 - photoresist 431
 - photoresist layer 429
 - polymer layer 430
 - protective polymer layer 431
 - resist layer 430
 - solid-phase synthesis 429
- photostability 436
- photostimulable storage phosphorus imaging plate 438–439
- PHP, *see* Professional Homepages 781
- Phrap 772–773
- Phred 771 ff., 773
- Phred/Phrap 820
- Phred/Phrap/Consed 821
- phycoerythrin 688
- piezoelectric dispenser 420–422
- piezoelectric jet 421
- piezoelectric printers, *see* piezoelectric printing 420
- piezoelectric printing 416, 420 ff.
- de-tritylation reagent 423
 - inter-spot distances 420
 - liquid droplet 421
 - multiple jet oligonucleotide synthesis 422 f.
 - needle 421
 - nozzle 422
 - phosphoramidite procedure 423
 - piezoelectric dispenser 420–422
 - piezoelectric jet 421
 - piezoelectric tip 420
 - polyacrylamide cushions 422
 - single jet oligonucleotide synthesis 422 f.
 - spot density 420
 - vaporisation 422
- piezoelectric tip 420
- pin-and-ring array system 418
- PIR, *see* Protein Information Resource
- PIR-International Protein Sequence Database (PIR-PSD) 652
- PIR-NREF, *see* Non-Redundant Reference Protein Database
- PIR-PSD, *see* PIR-International Protein Sequence Database 652
- Plant Plasma Membrane Database (PPMdb) 864
- plant RNAs 133 ff.
- PlasmoDB, *see* *Plasmodium* Genome Resource Database
- Plasmodium* Genome Resource Database (PlasmoDB) 859
- plasticware 135 f.
- PMMA-2DPAGE 854
- PMMA-2DPAGE, *see* Purkyne Military Medical Academy 2D-PAGE Database
- PNAs, *see* peptide nucleic acid
- polarisation 666
- poly(A) nuclease 33
- poly(A) tail 3 ff., 5
- poly(A)-PCR, *see* SIP RT-PCR
- polyacrylamide cushions 422
- polyacrylamide gel electrophoresis (PAGE) 302–305, 324, 380, 553, 628, 648, 861
- polyadenylate tail 3 ff.
- polyadenylation 9
- polycomb/trithorax group 29
- polyvinylidene fluoride (PVDF) 627
- PombeDB™ 866
- positional weight matrix (PWM) 803
- positron emission tomography (PET) 739
- 2-[¹⁸F]-fluoro-2-deoxy-D-glucose (FDG) 742
 - 3'-deoxy-3'-[¹⁸F]-fluorothymidine (FLT) 743
 - 3-(2'-[¹⁸F]-fluoroethyl)spiperone (FESP) 744–745
 - 8-[¹⁸F]-fluoroganciclovir (FGCV) 743
 - 8-[¹⁸F]-fluoropenciclovir (FPCV) 745

- annihilation radiation 739
- coincidence event 740
- FDG, *see* 2-¹⁸F-fluoro-2-deoxy-D-glucose
- FESP, *see* 3-(2'-¹⁸F-fluoroethyl)spiperone
- FGCV, *see* 8-¹⁸F-fluoroganciclovir
- FIAU, *see* [¹²⁴I]-2'-deoxy-2'-fluoro-5-iodo-1β-D-arabinofuranosyluracil
- FLT, *see* 3'-deoxy-3'-¹⁸F-fluorothymidine
- FPCV, *see* 8-¹⁸F-fluoropenciclovir
- genetic tagging 746
- genotypic interaction 746
- half-lives of the nuclides 746
- molecular imaging 741
- nuclear imaging 740
- PET reporter genes 746
- positron annihilation 740
- principle 741
- quantification 743
- radionuclides 740
- spatial resolution 740
- thymidine kinase 743–744
- virus thymidine kinase 743
- [¹²⁴I]-2'-deoxy-2'-fluoro-5-iodo-1β-D-arabinofuranosyluracil (FIAU) 743–744
- positron emission tomography (PET) 739 ff.
- post-hybridisation 710
 - non-hybridised probe 710
 - ribonuclease 710
- post-transcriptional gene silencing 45 ff.
- post-transcriptional modification 45, 851
 - antisense RNA 45
 - polyadenylation 9, 45 ff.
 - RNA editing 47, 174
 - sulphation 54
- post-transcriptional regulation 30 ff.
 - alternative splicing 38 f.
 - differential splicing 38 f.
 - intracellular localisation of RNA 40 ff.
 - intracellular RNA localisation 40 ff.
 - localisation of RNA 40 ff.
 - RNA degradation 31 ff.
 - RNA localisation 40 ff.
 - RNA stability 31 ff.
 - RNA transport 40
 - trans-RNA splicing 39
 - transcription elongation 33 ff.
- post-translational modification 9, 51, 53, 57, 623, 660, 851
 - acetylation 2–5, 600
 - acylation 53
 - carboxylation 57
 - cofactor 57
 - glycosylation 56 ff., 656–600
 - hydroxylation 57
- PostgreSQL 820
- POWER_SAGE 556
- PPDB, *see* PhosphoProtein Database
- PPMdb, *see* Plant Plasma Membrane Database
- pre-initiation 16
 - pre-initiation complex 17–18
- pre-mRNA maturation 39
- pre-mRNA trans-splicing 39
 - discontinuous group II intron form 39
 - spliced leader type 39
- preferential amplification of coding sequences (PACS) 238, 332
 - 'ATG'-seeking primers 335
 - auto-radiography 334, 336
 - coding sequence tag 333
 - dCST, *see* coding sequence tag
 - double restriction site primer (DRSP) 334–337
 - DRSP, *see* double restriction site primer
 - electrophoresis 336
 - initiation codon 335
 - multi-PACS 338
 - open reading frame 335
 - ORF, *see* open reading frame
 - PACS-SuRF, *see* coupled PACS and SuRF
 - primer design 335
 - principle 333
 - random primers 333
 - rationale 334
- preliminary RNA 2
- prenylation 53
 - farnesyl group 53
 - geranylgeranyl group 53
 - post-translational modification 53
- preparation of cells, *see* cell isolation
- preparation of nucleic acids, *see* RNA isolation
- preparation of proteins, *see* protein preparation
- preparation of RNA, *see* RNA isolation
- preRNA, *see* preliminary RNA
- PRG, *see* PET reporter gene
- probe design 708
 - GC/AT ratio 709
 - intron/exon boundaries 709
 - secondary structure 709
 - sensitivity 709
 - specificity 708
- probe labelling 722
 - 5-bromo uridine 5'-triphosphate 722
 - direct labelling 722
 - indirect labelling 722
- probe labels 707 f.
 - biotin 708
 - digoxigenin 708
 - radioisotopes 707 f.

- Probe Picher 433
- Professional Homepages (PHP) 781
- prokaryotic RNAs 133 ff.
- PromFind 809
- promoter 3 ff., 17, 800
 - enhancers 802
- promoter 2.0 809
- promoter finding 798 ff.
 - *in silico* predictions 798 ff.
 - sequence-based approach 808
 - signal-based approach 808
- promoter recognition, *see* promoter finding
- promoter region 12
 - DNA-binding site 12
- promoter remodelling 23
- promoter structure 799
- PromoterInspector 810
- PromoterScan 810
- PROSITE 854
- Prostate Expression Database (PEDB) 859
- prosthetic group, post-translational modification 57
- protein interaction 14
- protein A 148
- protein abundance 59
- Protein Data Bank (PDB) 652, 657, 854, 861
- Protein Database of Haematopoietic Cell Lines (BPP-2DPAGE) 864
- Protein Database of the Statens Serum Institute (SSI-2DPAGE) 864
- protein databases 852
- Protein Disease Database (PDD) 865
- protein domains 774
- protein expression analysis 623 ff., 687
- Protein Expression Database for the Molecular Pharmacology of Cancer 865
- protein G 148
- protein G sepharose 148
- Protein Information Resource (PIR) 652, 853
- protein isolation 133
- protein kinase-mediated phosphorylation 50
- protein libraries 664
- protein localisation 42
 - backup mechanism 42
- protein microarray 660, 664
 - binding curve 668
 - biomolecular interaction 670
 - co-ordinate covalent bonding 661
 - electrostatic interaction 661
 - enzyme-linked reaction 668
 - hydrogen bonding 661
 - hydrophobic interaction 661
 - immobilisation matrix 663
 - isotopic label 668
 - kinetic data 669
 - label-free detection 670
 - matrix polymers 663
 - molecular weight 666
 - molecule interaction 670
 - polarisation shift 669
 - principle 661–662
 - radioactivity 668
 - spectral shift 669
 - substrate 663
- protein preparation 133, 144 ff.
 - cell cultures 144 ff.
 - concentration of the sample 147
 - glassware 149
 - membrane-associated proteins 147
 - reducing agents 145
 - soluble proteins 149
 - surfactants 145
 - tissues 144 ff., 149
 - urea 145
- Protein Prospector 651
- protein sample preparation, *see* protein preparation
- protein separation 144, 624 ff.
 - affinity chromatography 625, 631, 633–636
 - capillary electrochromatography (CEC) 635
 - capillary electrophoresis (CE) 624, 635
 - CE, *see* capillary electrophoresis
 - CEC, *see* capillary electrochromatography
 - column chromatography 631, 633
 - 2D-DIGE, *see* two-dimensional difference gel electrophoresis
 - EDTA 144
 - EDTA and phenylmethylsulfonyl fluoride 144
 - gel filtration 633
 - high-performance liquid chromatography (HPLC) 624, 630, 635
 - HPLC, *see* high-performance liquid chromatography
 - ICAT™, *see* isotope-coded affinity tag method
 - IEC, *see* ion exchange chromatography
 - IEF, *see* isoelectric focussing
 - ion exchange chromatography (IEC) 624, 631, 633, 635
 - isoelectric focusing (IEC) 624 f.
 - isotope-coded affinity tag method (ICAT™) 636
 - MELK, *see* multi-epitope ligand-Kartographie
 - multi-dimensional separation 625, 635
 - multi-epitope ligand-Kartographie (MELK) 637 f.
 - normal phase chromatography 631, 634
 - open-column chromatography 631

- PAGE, *see* polyacrylamide gel electrophoresis
- 2D-PAGE, *see* two-dimensional polyacrylamide gel electrophoresis
- paper chromatography 631
- phenylmethylsulfonyl fluoride
- principle 636
- proteinase inhibitor 144
- reverse phase chromatography (RPC) 631, 634
- RPC, *see* reverse phase chromatography
- SDS-PAGE, *see* sodium dodecyl sulphate polyacrylamide gel electrophoresis
- SEC, *see* size exclusion chromatography
- size exclusion chromatography (SEC) 624, 631, 633 f., 635
- slab gel electrophoresis 624, 625 ff., 635
- sodium dodecyl sulphate polyacrylamide gel electrophoresis SDS-PAGE 624, 626 ff.
- thin-layer chromatography 631
- two-dimensional difference gel electrophoresis (2D-DIGE) 630
- two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) 629
- Western blotting 626 ff.
- protein/protein interactions 800
- proteolysis 660
- proteolytic cleavage 52
 - endoplasmic reticulum 52
 - post-translational modulation 52
 - preproprotein 52
 - preprotein 52
 - signal peptidase 52
 - signal peptide 52
- proteome 9, 13
- Proteome 2D-PAGE Database 865
- Proteome Analysis Database 853 ff.
- Proteome Database for Aging Research (TMIG-2DPAGE) 866
- Proteome®'s BioKnowledge® library 866
- proteomics databases 651
- proteomics servers 651
- PROWL 651
- PRP, *see* PET reporter probe
- pseudogene 64
- PubMed 835, 840
- Purkyne Military Medical Academy 2D-PAGE Database (PMMA-2DPAGE) 867
- PVDF, *see* polyvinylidene difluoride
- PWM, *see* positional weight matrix
- q**
- QCM, *see* quartz crystal microbalance 459
- quadrupoles 641
- QuantiGene™ expression assay 471
- QuantiGene™ expression kit 475
- Quantikine® mRNA assay 471
 - 96-well microplate format 471
 - amplification 472
 - biotin 471 f.
 - biotin/streptavidin interaction 472
 - calibrator curve 473
 - calibrator mRNA 473
 - capture probe 471
 - cell lysate samples 471, 476–477
 - cell lysate 471
 - colourimetric detection 471
 - detection probes 471
 - digoxigenin 471 f.
 - ELISA-like format 471
 - high-throughput screening 477
 - linearity 474
 - microplate reader 473
 - microplate-based mRNA detection assay 471
 - multiple analytes 477
 - poly(A)⁺ mRNA 471
 - precision 474 f.
 - principle 471 f.
 - semi-quantitative determination 476
 - sensitivity 477
 - signal amplification 476–477
 - standard curve 473
 - streptavidin 472 f.
 - total RNA 471
- quantum dots 435
- quantum efficiency 108
- quartz crystal microbalance (QCR) 459
 - change in frequency 459
 - crystal's oscillating frequency 459
 - indicator-free detection 459
 - mass-measuring detection 459
 - micro-gravimetric device 459
 - oscillator 459
 - piezoelectric principle 459
 - quartz crystal 459
 - Sauerbrey relationship 459
- quartz crystal microbalance dissipation 434
- QuickTime Virtual Reality™ movies 817, 828
- QuickTime™, *see* QuickTime Virtual Reality™ movies
- r**
- r, *see* r resources 786
- R-Plus® 787
- RACE, *see* rapid amplification of cDNA ends
- RAD, *see* RNA Abundance Database
- Radiation Hybrid Consortium 832
- Radiation Hybrid Database 832

- radioactive labelling 438 f.
 - 32-phosphorus 438
 - 33-phosphorus 438–439
 - 35-sulphur 438–439
 - auto-radiography imaging systems 438
 - PCR amplification 438
 - photostimulable storage phosphorus imaging plates 438–439
 - radioactively-labelled oligonucleotides 438
 - random primed labelling 438
 - tritium 439
 - X-ray films 438
- radioactively-labelled oligonucleotides 438
- radioactivity 439
- radioimmunoassay (RIA) 667, 678, 679
- radioisotope 749
- radiolabelled antisense oligodeoxynucleotide (RASON) 744 f., 749
- RAGE, *see* rapid analysis of gene expression
- random primed labelling 435, 438–439
- random sequencing 826
- RAP-PCR, *see* RNA arbitrarily primed PCR fingerprinting
- rapid amplification of cDNA ends (RACE) 516, 570
- rapid analysis of gene expression (RAGE) 511 ff.
 - capillary electrophoresis 517
 - coefficients of variation 517
 - EST database 516
 - expense 517
 - flexibility 516
 - gene discovery 516
 - high-throughput 518
 - immobilised cDNAs 512
 - linker 512
 - magnetic bead 512
 - maintenance 517
 - mis-priming events 514
 - multi-channel fluorescent detection 517
 - overhangs 514
 - PCR amplification 512
 - principle 511 ff.
 - RAGE PCR 515
 - RAGE primers 513
 - RAGETag library 511, 514–516
 - recognition site 514
 - restriction enzyme 512, 514
 - restriction site 512
 - robotic liquid handling equipment 517
 - sequencing 514
 - statistical significance 517
 - streptavidin 512
 - variability 517
- rapid subtraction hybridisation (RaSH) 206 ff.
 - cDNA library 211
 - driver 207
 - Northern blot 207, 209 f.
 - phage library 209
 - plasmid library 209
 - protocol 208
 - RaSH libraries 212 f.
 - reverse Northern 207, 209 f.
 - Southern blot 211
 - subtracted cDNA library 207–208, 209 ff., 212
 - tester 207
- RaSH, *see* rapid subtraction hybridisation
- RASON, *see* radiolabelled antisense oligodeoxynucleotide
- Rat Serum Protein Database 867
- Rat Serum Protein Study Group Database 867
- RC4D, *see* restriction fragment length polymorphism-coupled domain-directed differential display
- RD-PCR, *see* restriction display polymerase chain reaction
- RDA, *see* representational difference analysis
- rDNA 16
- re-sequencing by hybridisation 415 f.
- READS™, *see* restriction enzyme analysis of differential sequences
- real-time reverse transcription polymerase chain reaction, *see* real-time RT-PCR
- real-time RT-PCR 69 ff., 124, 345 ff., 365, 369–370
 - *in situ* RT-PCR 358
 - 5'-3' exonuclease activity 345, 349, 349 ff.
 - 5'-nuclease activity 349
 - 5'-nuclease assay 347 ff., 353
 - absolute copy numbers 357
 - acceptor fluorophore 349–350, 355
 - amplicon 347
 - amplification efficiency 347, 355–356
 - amplification plot 355, 356, 357
 - background fluorescence 346–347
 - cleavage 348
 - cleavage of the fluorogenic probe 347
 - DNA-binding dyes 346 ff.
 - dynamic range 355
 - elongation 347
 - ethidium bromide 345
 - extension phase 355
 - external standard 358
 - flexibility 350
 - fluorescein donor 350

- fluorescence detection 346
- fluorescence resonance energy transfer (FRET) 349–350
- fluorescence signal 346
- fluorogenic probes 347
- Förster resonance energy transfer (FRET) 349 f.
- FRET, *see* Förster resonance energy transfer or fluorescence resonance energy transfer
- head-to-tail arrangement 351
- housekeeping gene 358
- hybridisation 351
- hybridisation probe 350 f.
- intercalator 345
- intra-molecular interaction 355
- melting curve 347
- molecular beacon 350, 352
- nucleotide composition 347
- PCR amplification 347
- PCR stopper 355
- polymerisation 346, 348–349, 351–352, 355
- post-PCR melting curve 345
- post-PCR processing 345
- precision 355
- primer extension 352
- principle 345 ff.
- probe cleavage 349
- probe design 350
- processivity 349
- quencher, *see* acceptor fluorophore
- reliability 357
- reproducibility 355
- run-to-run variation 355
- Scorpion™ 355 ff.
- sensitivity 356
- specificity 350
- standard curve 355–356, 358
- steady-state mRNA level 345, 357
- strand displacement 348
- SYBR® Green 345 ff.
- TaqMan™ 349, 353
- thin layer chromatography 345
- threshold cycle 346–347, 356–357
- receiver coil 733
- receptor assay 673
- receptor signal 27
 - internal 27
 - receptor-ligand 27
 - steroid 27
- receptor-targeted NIRF-probe 758
- reciprocal subtraction differential RNA display (RSDD) 219, 238
 - DDRT-PCR, *see* differential display RT-PCR
 - differential display RT-PCR (DDRT-PCR) 219, 223
 - driver 220
 - expressed sequence tag (EST) 220
 - Northern blot 223
 - PCR display 222
 - principle 220
 - protocol 221
 - reciprocal subtraction 221
 - reverse Northern 221, 223
 - subtracted cDNA 223
 - subtracted cDNA library 219, 222–223
 - subtractive hybridisation 220
 - tester 220
- RECODE, *see* Database of the Translational Recoding Events
- reflectometric interference spectroscopy (RIFS) 434, 460 ff.
 - angle of incidence 460
 - interference 460
 - interference spectrum 460
 - molecule layer 460
 - monochromatic light 460
 - refractive index 460, 460 f.
- reflectometric interferometry 667, 669
- reflectometry 667
- refractive index 668
- RefSeq 576, 652
- region of increased gene expression (RIDGE) 24, 555
- regulatory promoter 20 f.
- regulatory region 802
- relational database 820
- relational database management system (RDMS) 821–822, 835, 842
- release factor 7
- RepeatMasker 771, 773
- repetitive element 849
- reporter gene 753
 - *Discosoma* species red fluorescent protein (DsRed) 754
 - DsRed, *see* *Discosoma* species red fluorescent protein
 - luciferases 753
- representational difference analysis (RDA) 214 ff., 234
 - adapter 215, 217
 - background noise 218
 - driver 214
 - driver representation 217
 - enzymatic digestion 217
 - exponential amplification 217
 - false-positive 218
 - kinetic enrichment 214

- linear amplification 217
- linker 215
- linker-ligated cDNA 216
- melt depletion 216
- PCR amplification 217–218
- PCR-based amplification 214
- principle 216–217
- representation 215
- Southern blot 216
- subtractive cloning 214
- subtractive hybridisation 214
- tester 214, 217
- repressor 26 ff., 28
 - global 28
 - specific 28
- repressor protein 799
- reproducibility 846
- research genetics 777
- resonance frequency 733
- resonant mirror 667, 669
- restriction display polymerase chain reaction (RD-PCR) 358 ff., 365
 - adapter 359–360
 - auto-radiograph 360
 - gel electrophoresis 359–360
 - ligation 359
 - ligation of adapter 360
 - low stringency 359, 361
 - PCR amplification 359
 - polyacrylamide gel 360
 - reproducibility 359, 361
 - restriction digestion 360, 359
- restriction enzyme analysis of differential sequences (READS™) 518 ff.
 - 3'-end fragments 518
 - 3'-end restriction fragment 522
 - 5'-overhang 520
 - adapter 518, 523, 527
 - adapter formation 521
 - auto-radiography 522
 - cDNA fragment 520, 522
 - co-migrating bands 527
 - correlation 527
 - database 526
 - digestion 520
 - discrepancies 525
 - fly adapter 520, 522
 - gel electrophoresis 526
 - heel 527
 - methylene blue 526
 - modification 527
 - PCR amplification 518–519, 521–523
 - polyacrylamide gel 522
 - primer annealing 529
 - primer 521
 - principle 519 ff.
 - procedure 519
 - re-cutting 523, 527–528
 - READS™ PCR 518
 - redundancy 528
 - refinement 527
 - reproducibility 523
 - restriction endonuclease 518
 - restriction endonuclease digestion 520
 - restriction enzyme 518, 520, 523, 527
 - self-annealing 529
 - semi-quantitative RT-PCR 524
 - Y-shaped adapter 520
 - Y-shaped double-stranded adapter 520
 - Y-shaped fly adapter 522, 518 ff.
- restriction fragment differential display (RFDD) 339, 359, 365
 - adapter 365
 - amplification 364
 - cDNA restriction fragments 359
 - downstream primer 364
 - EP adapter 364–365
 - extension protection group 365
 - gel electrophoresis 364
 - ligation 364–365
 - restriction digestion 364
 - restriction fragment 365
 - upstream primer 364
- restriction fragment length polymorphism (RFLP) 498 f., 359, 363 f.
 - adapter 362–363
 - auto-radiograph 362
 - auto-radiography 363
 - family-specific domain (FSD) 363
 - FSD primer 362
 - FSD, *see* family-specific domain
 - gel electrophoresis 362
 - ligation 362
 - linker 362
 - polyacrylamide gel 363
 - principle 359
 - restriction enzyme digestion 362
- restriction fragment length polymorphism-coupled domain directed differential display (RC4D) 359, 363
- restriction landmark cDNA scanning (RLCS) 373 ff.
 - anchor primer 375–376
 - auto-radiograph 375
 - biotin 374
 - blocking 374
 - genetic mapping 373
 - landmark 373

- magnetic beads 374
- mutation analysis 373
- principle 376 ff.
- restriction enzyme 375–376, 379
- restriction enzyme digestion 374
- restriction enzyme site 376
- spot cloning 376
- streptavidin 374
- two-dimensional gel electrophoresis 375–377
- restriction landmark genomic scanning (RLGS) 373
 - DNA imprinting 374
 - DNA methylation 374
- reverse phase chromatography (RPC) 631, 634
- reversed-phase high-performance liquid chromatography (RP-HPLC) 657
- RFDD, *see* restriction fragment differential display
- RFLP, *see* restriction fragment length polymorphism
- RIA, *see* radioimmunoassay
- ribonuclease protection assay (RPA) 255 ff., 477
 - *in vitro* transcription 255, 257
 - alternative transcription initiation 255
 - auto-radiography 257
 - bacteriophage promoter 256
 - bacteriophage RNA polymerase 256
 - biotin 257, 259
 - gel electrophoresis 259
 - lysate ribonuclease protection assay 259 f.
 - non-isotopic format 259
 - nuclease digestion 259
 - nylon membrane 259
 - PAGE, *see* polyacrylamide gel electrophoresis 257
 - polyacrylamide gel electrophoresis (PAGE) 257
 - polyadenylation 255
 - principle 256 ff.
 - ribonuclease A 257
 - ribonuclease T1 257
 - riboprobe 257
 - RNA polymerase 255 f., 257
 - S1 nuclease mapping 255
 - SP6 RNA polymerase 255, 258
 - T3 RNA polymerase 255, 258
 - T7 RNA polymerase 255, 258
- ribonuclease 134 ff.
- ribonuclease protection assay 255
- ribonucleoprotein 16
- ribosomal RNA 3 ff., 11, 67
- ribosome 3 ff., 6 ff., 8
- RIDGE, *see* region of increased gene expression
- RlfS, *see* reflectometric interference spectroscopy
- RIKEN Mouse Gene Encyclopaedia Project 576
- RLCS, *see* restriction landmark cDNA scanning
- RLGS, *see* restriction landmark genomic scanning
- RNA 133 ff.
 - cleanup 136
 - cytoplasmic RNA 136
 - low-molecular weight 136, 139
 - messenger RNA (mRNA) 136
 - poly(A)⁺ mRNA 139
 - precipitation 138
 - prokaryotic mRNA 140 f.
 - purification 134 ff.
 - ribosomal RNA (rRNA) 136, 138, 140
 - rRNA, *see* ribosomal RNA
 - small RNA 138 ff.
 - stabilisation 134, 140 ff.
 - total 136
 - transfer 138, 140
 - tRNA, *see* transfer RNA
- RNA abundance database (RAD) 842 f., 857, 859
- RNA arbitrarily primed PCR fingerprinting (RAP-PCR) 339, 340 ff.
 - arbitrary primer 340, 343
 - electrophoresis 340
 - false-positive 343
 - high stringency 340
 - low stringency 340
 - principle 340 ff.
- RNA degradation 31 ff., 141
 - deadenylation-dependent pathway 31 ff., 33
 - deadenylation-independent pathway 31 ff., 33
- RNA distribution pattern 716
- RNA editing 47 ff., 174
 - kinetoplast-mitochondria 49
 - trypanosome 47
- RNA fingerprinting by arbitrarily primed PCR 218
- RNA interference 47
- RNA isolation 33 ff., 175
 - alcohol precipitation 138
 - anion-exchange 139
 - caesium chloride density-gradient 138 ff.
 - chaotropic agent 135
 - chloroform 137

- DEPC, *see* diethyl pyrocarbonate
- diethyl pyrocarbonate (DEPC) 135
- disruption 135
- enzymatic digestion 137
- ethidium bromide 138 f.
- general handling 134
- glassware 135
- guanidine hydrochloride 138, 139 f.
- guanidine isothiocyanate 138, 139 f.
- homogenisation 135
- hybrid selection 139
- lithium chloride precipitation 138
- organic extraction 137
- organic solvent 135
- phenol 137
- plasticware 135
- poly(A)⁺ mRNA 139
- precipitation 138
- ribonuclease (RNase) 134
- RNase inhibitor 135
- RNase 134
- silica-based methods 139 f.
- sodium iodide 139 f.
- sodium perchlorate 139 f.
- spin column 140
- stabilisation 134
- strong denaturant 138
- sucrose density-gradient 138 f.
- RNA localisation 40 ff.
- actin-dependent transport 42 f.
- active transport 42 f.
- *cis*-acting elements 44
- diffusion 42
- local anchoring 42
- localised degradation 42
- localised synthesis 42
- localised transcription 42
- multistep localisation 42 f.
- *trans*-acting factors 44
- RNA masking 51
- RNA patterns method 379 ff.
- background 382
- dideoxynucleotide 380
- exponential amplification 379
- fluorescent labelling 381
- linear amplification 380
- PAGE, *see* polyacrylamide gel electrophoresis
- polyacrylamide gel electrophoresis (PAGE) 380, 381
- principle 380 ff.
- radioactive labelling 381
- snake venom phosphodiesterase 382
- termination of the chain elongation 380
- terminator 382
- RNA polyadenylation 45 f.
- RNA polymerase 3 ff., 17
 - catalytic site 17
 - subunits 17
- RNA polymerase arrest 34
- RNA polymerase I 11, 16 f., 27
- RNA polymerase II 14 ff., 16 f., 19, 21, 27 f., 799
 - carboxy-terminal domain (CTD) 19
 - CTD, *see* carboxy-terminal domain
- RNA polymerase II holoenzyme 27 f.
- RNA polymerase III 11, 15 ff., 16 f., 27 f.
- RNA polymerase termination 34
- RNA preparation, *see* RNA isolation
- RNA processing 9
- RNA stabilisation 134, 140 f.
 - anticoagulant 142
 - citrate 142
 - degradation 141
 - denaturing reagent 141
 - dry ice 141, 143
 - EDTA 142
 - freezing 141
 - guanidine hydrochloride 141
 - heparin 142
 - liquid nitrogen 141, 143
 - reagents for bacteria 143
 - reagents for blood 143
 - reagents for tissues 143
- RNA stability 9, 31 ff., 50
- RNA transport 9, 40, 50
- RNase inhibitor 135
- RNase, *see* ribonuclease
- Rochester Muscle Database 843 f.
- ROX 437
- RP-HPLC, *see* reversed-phase high-performance liquid chromatography
- RPA, *see* ribonuclease protection assay
- RPC, *see* reverse phase chromatography
- rRNA, *see* ribosomal RNA
- r resources 786
- RSDD, *see* reciprocal subtraction differential display
- RZPD, *see* German Resource Center Primary Database

S

- S-Plus® 786–787
- S1 nuclease mapping, *see* ribonuclease protection assay
- SABRE, *see* selective amplification *via* biotin and restriction-mediated enrichment
- Saccharomyces* Genome Database (SGD) 845, 848, 850

- SADE, *see* SAGE adaption for downsized extracts
- SAGE adaption for downsized extracts (SADE) 581 ff.
- anchoring enzyme 565, 568, 581, 583, 585
- cloning 582
- computational analysis 582
- concatemer 582
- concatenation 581
- data acquisition 581
- ditag 582
- ligation 582
- linker 581–582
- magnetic bead 582
- PCR amplification 581–582, 585
- plasmid 582
- principle 581 f.
- ribonuclease H activity 585
- SAGE library 581, 585
- SAGE tag library 583
- sequencing 582
- streptavidin 585
- tag 581–582, 585
- SAGE gene to tag mapping 98
- SAGE homepage 557
- SAGE library 844
- SAGE, *see* serial analysis of gene expression
- SAGE-lite 556, 565 f., 577 ff.
- distortion 579
- exponential phase 579
- over-amplification 579
- PCR amplification 578–579
- PCR distortion 580
- PCR kinetic 579
- principle 578
- restriction enzyme 578
- SAGE catalogue 577, 579–580
- SAGE tag 580
- strand switching 578, 580
- SAGEmap 576, 591, 777, 844
- gene to tag mapper 844
- tag to gene mapper 844
- SAGEmap virtual Northern 859
- SAGEmap xProfiler 859
- SAGEmap, *see* tag to gene mapping database
- SAM, *see* Sequence Alignment and Modeling Software System
- sample preparation 97 ff., 709 f.
- digestion 710
- pre-treatment 709
- proteinase K 710
- sample preservation 709
- acquisition 709
- cross-linking 709
- fixation 709
- formalin 709
- glutaraldehyde 709
- liquid nitrogen 709
- paraformaldehyde 709
- snap freezing 709
- sample separation in proteomics, *see* protein separation
- sandwich type hybridisation assay 445
- sandwich-assay 674 f.
- Sanger Institute 779
- Sauerbrey relationship 459
- SAW, *see* surface acoustic wave
- Scanalyze 846
- scanning tunnelling microscope (STM) 449 f.
- scintigraphy 667, 749, 751
- ScorpionTM 353 ff.
- hexaethylene glycol 353
- SDD, *see* subtracted differential display
- SDS-based polyacrylamide gel electrophoresis (SDS-PAGE) 147, 624, 626, 629, 647
- Coomassie Blue 627
- discontinuous electrophoresis 627
- immunoblotting 628
- immunochemical detection 628
- polyacrylamide gel electrophoresis (PAGE) 628
- polyvinylidene fluoride (PVDF) 627
- principle 627
- PVDF, *see* polyvinylidene fluoride
- SDS, *see* sodium dodecyl sulphate
- separation gel 627
- silver 627
- stacking gel 627
- staining 627
- Western blotting 628
- SDS-PAGE, *see* SDS-based polyacrylamide gel electrophoresis
- SEC, *see* size exclusion chromatography
- SELDI, *see* surface-enhanced laser-desorption/ionisation
- selective amplification *via* biotin and restriction-mediated enrichment (SABRE) 227 ff.
- biotin 229
- competitive hybridisation 227 f.
- differential library screening 229
- driver 228
- enrichment 228–229, 231
- magnetic bead 229
- PCR amplification 229
- PCR tagging 229
- plasmid vector 229
- post-hybridisation distribution 228

- principle 227 ff.
- restriction enzyme 229
- streptavidin 229
- tester 228
- self-organising data 500
- self-organising map (SOM) 794 f., 846
 - algorithm 795 f.
- separation of proteins, *see* protein separation
- separation of samples *see* protein separation
- sepharose 147
- Sequence Alignment and Modeling Software System (SAM) 772, 773
- Sequence Retrieval System (SRS) 849
- Sequence Tag Alignment and Consensus Knowledgebase (STACK) 772
- sequencing by hybridisation 464
- sequential activation 23
- serial analysis of gene expression (SAGE) 551, 557, 559, 568, 570, 577, 581, 586, 844
 - amplification 552, 557
 - anchoring restriction enzyme 551–553, 557
 - cleavage specificity 553
 - cloning 552
 - computer sampling 555
 - concatemer 552, 553, 556
 - concatenation 552, 556
 - digestion 553
 - ditag 552, 556–557
 - DNA sequencing 551
 - experimental process 555
 - high-throughput 551
 - ligation 552, 557
 - linker 552, 557
 - PAGE, *see* polyacrylamide gel electrophoresis
 - polyacrylamide gel electrophoresis (PAGE) 553
 - principle 551 ff.
 - SAGE database 551, 553, 556–557
 - SAGE libraries 551, 554–556
 - SAGE tag 551, 553–557
 - sequencing 552
 - statistical analysis 553
 - streptavidin 552
 - tagging restriction enzyme 552–553, 556–557
- serial dilution 70 f.
- serum-responsive element (SRE) 28
- serum-responsive factor (SRF) 28
- SGD, *see* *Saccharomyces* Genome Database
- short tandem repeat (STR) 456
- shotgun sequencing 181, 469
- SIB, *see* Swiss Institute of Bioinformatics
- sibling cell analysis 368
- side scatter (SSC) 112, 688, 693
- Siena-2DPAGE 854, 867
- signal amplification 476
- SIGNAL SCAN 802
- SII-type elongation factor 36
- silencer 26, 28
 - active repression mechanism 26
 - classical silencer element 26
 - negative regulatory element (NRE) 26
 - NRE, *see* negative regulatory element
 - passive repression mechanism 26
 - silencer element 26
- silica-based method 139
- silver staining 443 ff., 629
 - *in vitro* transcription 444
 - advantages 446
 - applications 449
 - ArrayTube® 447
 - auto-nucleation 448
 - biotin 445
 - bleaching 447
 - bromodeoxyuridine 445
 - concentration 448
 - digoxigenin 445
 - direct labelling 443 f.
 - disadvantages 448
 - dynamic detection 447
 - gold label 444
 - indirect labelling 443 f.
 - limits 448
 - principle 443
 - quenching 447
 - reporter group 444
 - resolution 447
 - sandwich type hybridisation assay 445
 - saturation 447
 - sensitivity 446
 - signal-to-noise ratio 447
 - single nucleotide polymorphism (SNP) 444
 - SNP, *see* single nucleotide polymorphism
 - storage 448
 - streptavidin 445, 447
- single cell PCR 124
- single jet oligonucleotide synthesis 422 f.
- single nucleotide polymorphism (SNP) 251, 415, 444, 456, 457, 665, 770
- single photon emission computed tomography (SPECT) 747
 - β -emitter 750
 - ^{99m}Tc -P829 751
 - ^{188}Re -P829 751
 - 2'-deoxy-5-iodo-1- β -D-arabinofuranosyluracil (FIAU) 751

- 2'-fluoro-5-methyl-1- β -D-arabinofuranosyl-uracil (FMASU) 751
- collimator 748
- direct imaging 749
- FIAU, *see* 2'-deoxy-5-iodo-1- β -D-arabinofuranosyluracil
- FMAU, *see* 2'-fluoro-5-methyl-1- β -D-arabinofuranosyluracil
- indirect imaging 750
- noninvasive *in vivo* detection 749
- principle 747, 750
- radioisotopes 748–749
- radionuclide 748
- thymidine kinase 751, 747 ff.
- single resist process, *see* photoresist wet masking
- single-pass sequencing 570
- SIP RT-PCR 363 ff.
 - poly(A) tailing 366 f.
 - principle 365 f.
 - sequence-independent primer (SIP) 366
 - spiking experiment 369
 - virtual Northern blot 367
- SITE 802
- SITEVIDEO 804 f.
- size exclusion chromatography (SEC) 624, 631, 633, 635
 - column 633 f.
 - principle 632, 633 f.
- slab gel electrophoresis 624, 625 ff., 635
 - 2D-DIGE, *see* two-dimensional difference gel electrophoresis
 - 2D-PAGE, *see* two-dimensional polyacrylamide gel electrophoresis
 - Coomassie Blue 629
 - immobilised pH gradient 629
 - immunoblotting 629
 - SDS-PAGE, *see* sodium dodecyl sulphate polyacrylamide gel electrophoresis
 - silver staining 629
 - sodium dodecyl sulphate polyacrylamide gel electrophoresis 626 ff.
 - staining 629
 - SYPRO 630
 - two-dimensional difference gel electrophoresis (2D-DIGE) 630
 - two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) 629
 - Western blotting 626 ff.
- SLFIA, *see* substrate-labelled-fluorescent immunoassay
- slot blot 167 ff., 171
- small nuclear RNA (snRNA) 3, 16
- SMART II cDNA synthesis system 577
- SMD, *see* Stanford Microarray Database
- SNP, *see* single nucleotide polymorphism
- snRNA, *see* small nuclear RNA
- sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) 629
 - β -mercaptoethanol 626
 - dodecyl sulphate 626
 - electrophoresis 626
 - SDS 626, 624, 626 ff.
- sodium iodide 139 f.
- sodium perchlorate 139 f.
- solid-phase extraction 647
- SOM, *see* self-organising map
- source code 846
- SOURCE, *see* Stanford Online Universal Resource for Clones and ESTs
- Southern blot 411, 439–440
- SP1 22, 27
- spatial distribution 747
- SPECT, *see* single photon emission computed tomography
- spectroscopy *in vivo* 734
- splicing 5
- spot-metric plotting 846
- SPR, *see* surface plasmon resonance
- SQL, *see* Structured Query Language
- SRE, *see* Serum-Responsive Element
- SRF, *see* Serum-Responsive Factor
- SRS, *see* Sequence Retrieval System
- SSC, *see* side scatter
- SSH, *see* suppressive subtractive hybridisation
- SSI-2DPAGE, *see* Protein Database of the Statens Serum Institute
- stabilisation of RNA, *see* RNA stabilisation
- STACK, *see* Sequence Tag Alignment and Consensus Knowledgebase 772, 777
- standard 787 ff.
- standard curve 69 f.
- Stanford Microarray Database (SMD) 845 ff.
- Stanford Online Universal Resource for Clones and ESTs (SOURCE) 847
- Stanford Yeast Database 790
- steady state concentration 174
- steady state mRNA level 355
- steady-state transcription level 591
- STM, *see* scanning tunnelling microscope
- Stoke's law 106
- Stoke's shift 108
- strand switching 578, 580
- streptavidin 445, 450 f., 472 f.
- streptavidin/biotin interaction 424
- Streptomyces coelicolor* reference gel 867
- strong cation exchange (SCX) 633

- strong denaturant 138
- STRs, *see* short tandem repeats
- Structured Query Language (SQL) 785–786
 - syntax 786
- subsets of sites 805
 - combinatorial nature 805
 - kernel function estimation 805
- substrate-labelled-fluorescent immunoassay (SLFIA) 676
- subtracted cDNA libraries 219
- subtracted differential display (SDD) 181
 - differential display RT-PCR 181
- subtracted RNA fingerprinting (SURF) 237 ff., 333–334, 338
 - auto-radiography 240–241
 - cDNA 238
 - competitor 242
 - dCST, *see* differentially expressed coding sequence tag
 - DDRT-PCR, *see* differential display RT-PCR
 - differential display RT-PCR (DDRT-PCR) 237, 239, 241–242
 - differentially expressed coding sequence tag (dCST) 242
 - double restriction site primer (DRSP) 240, 242 f.
 - driver 238
 - DRSP, *see* double restriction site primer
 - electrophoresis 240
 - false-positive 238, 241–242
 - global screening of expressed genes 238
 - magnetic bead 241
 - mRNA fingerprinting 240
 - PACS, *see* preferential amplification of coding sequences
 - paramagnetic bead 239
 - poly(dA/dT) hybrid 243
 - preferential amplification of coding sequences (PACS) 238, 242
 - primer 240
 - principle 238 ff.
 - random annealing 243
 - reciprocal subtraction differential display (RSDD) 238
 - redundant identification of genes 238
 - reproducibility 238
 - RSDD, *see* reciprocal subtraction differential display
 - subtracted libraries 238
 - subtractive hybridisation 237, 239
 - tester 238, 242
 - vector 240
- subtraction hybridisation, *see* subtractive hybridisation
- subtraction *in silico* 776 f.
- subtractive cDNA library 192
- subtractive hybridisation 178 ff., 181, 344
 - enrichment 178
 - heterohybrid 178
 - homohybrid 178
 - post-hybridisation distribution 178
 - principle 178
 - shot-gun sequencing 181
- subtractive hybridisation of transcriptionally active genomic DNA fragments 233 ff.
 - advantages 237
 - chromatin 233
 - driver 234–235
 - genomic DNA 234
 - ligation 233
 - limitations 237
 - linker 233–234
 - mercury affinity chromatography 234
 - mung bean nuclease 235
 - nuclei 233
 - nucleosome 233
 - organomercurial agarose column 234
 - principle 233
 - purification of transcriptionally active genomic DNA 234
 - RDA, *see* representative difference analysis
 - representative difference analysis (RDA) 234
 - restriction digestion 235
 - subtraction library 235
 - subtractive hybridisation 233–235
 - tester 234–235
 - transcriptionally active 234
 - transcriptionally active chromatin 237
 - visualisation 236
- sucrose 138 f.
 - methylmercuric hydroxide 139 f.
 - sucrose density-gradient 138 f.
- sucrose density-gradient 138 f.
- sulphation 54
- supercoiling 26
- superparamagnetic nanoparticle 724–725
- superparamagnetic iron oxide nanoparticle 727
- supervised classification approach 793
- support vector machine 793
- suppressive subtractive hybridisation (SSM) 244
 - adapter 246
 - adapter ligation 245
 - alternative approach 248
 - avidin/biotin-based modification 248
 - cDNA library 246

- digestion 245
 - driver 244
 - exponential amplification 245–246
 - false-positive 249
 - hybridisation 246
 - intramolecular hybridisation 245
 - kinetic nature of the subtraction 249
 - kinetics of hybridisation 246
 - library screening 248
 - linear amplification 245
 - magnetic bead 248
 - modifications 248
 - pan-like structure 246
 - PCR amplification 244, 248
 - phenol/chloroform extraction 248
 - primer annealing site 246
 - principle 244 ff.
 - procedure 244 ff.
 - re-annealing 246
 - restriction enzyme digestion 244
 - subtractive hybridisation 244, 246, 249
 - suppression PCR effect 246
 - tester 244
 - SuRF, *see* subtracted RNA fingerprinting
 - surface acoustic wave (SAW) 434, 668
 - surface affinity chromatography 105 f.
 - surface plasmon resonance (SPR) 461, 661 ff., 663–664, 667–669, 678
 - adsorption 461
 - evanescent field 668
 - index of refraction 461
 - interference filter 462
 - label-free optical technique 461
 - metal interfaces 461
 - principle 668
 - reflectivity imaging 461
 - surface plasmon resonance imaging 434
 - surface-enhanced laser-desorption/ionisation 669, 667
 - suspension arrays
 - bead technology 250 f.
 - capture oligonucleotide 252
 - competitive assay 251–252
 - competitor 252
 - direct assay 252
 - expression monitoring 253
 - flow cytometer 250
 - fluid phase hybridisation 251
 - fluorescent microbead 249 ff.
 - general procedure 254
 - mutation detection 251
 - principle 250 f.
 - reaction kinetics 251
 - reporter oligonucleotide 251–252
 - SNP detection 251
 - ZipCode™ 251–253
 - SWI/SNF-like protein complex 25
 - SWICZ, *see* Swiss-Czech Proteomics Server 868
 - Swiss Institute of Bioinformatics (SIB) 651, 786
 - SWISS-2DPAGE 853–854, 868
 - Swiss-Czech Proteomics Server (SWICZ) 868
 - SWISS-PROT 651, 774, 832, 853, 853 f., 854, 861, 868
 - Sybase™ 835, 842
 - SYBR® Green 345 ff., 436
 - systematic differential display approach 326
 - systems approach 62
 - systems biology 538, 548
- t**
- tag to gene mapping database (SAGEmap) 844
 - 3'-tailing 169, 439
 - TALEST, *see* tandem arrayed ligation of expressed sequence tags
 - TAMRA 437
 - tandem affinity purification 15
 - tandem arrayed ligation of expressed sequence tags (TALEST) 585 ff.
 - *in vitro* amplification 586
 - adapter 588–589
 - artefacts 590
 - bacterial amplification 589
 - Bayesian statistical analysis 592
 - cDNA libraries 586
 - concatemer 587–588
 - digestion 588
 - ditag 586
 - DNA sequencing 586
 - false-negative 591
 - false-positive 591
 - gel electrophoresis 588
 - ligation 589
 - magnetic bead 587
 - PCR amplification 586
 - plasmid DNA 587
 - plasmid 588
 - principle 587 ff.
 - punctuating enzyme 588
 - restriction enzyme 588, 590
 - restriction enzyme digestion 589
 - statistical significance 592
 - streptavidin 587–588
 - tag libraries 590
 - tag-to-gene assignments 589
 - tag 588

- TALEST libraries 590
- threshold 592
- type II restriction endonucleases 586
- type II restriction enzyme 589
- type II restriction site 588, 585 ff.
- tandem mass spectrometry (MS/MS) 636, 644
- principle 644
- TaqMan™, *see* 5′-nuclease assay
- targeted display 370 ff.
- targeted imaging 724
- TATA box 4
- TATA box-binding protein (TBP) 17, 21, 27
- TBP, *see* TATA box-binding protein
- TCEA1, *see* transcription elongation factor A1
- television technology 428
- template-switching oligonucleotide 565 f.
- termination factor 7
- tester 178
- tetramethyl rhodamine 437
- Texas Red 437, 688
- TFD, *see* Transcription Factor Database
- TFIID 27
- TFIIH 28
- TFIIS 36
- thermodynamic considerations 803
- thin-layer chromatography 631
- three-dimensional visualisation 811 ff., 814, 828 f., 814 ff.
 - computer-driven realignment 814
 - database 828
 - digital enlargement capability 818
 - digital image processing 815
 - drill hole 814–815
 - dynamic thresholding algorithm 816
 - embedding 814
 - expression map 813
 - gene-oriented queries 828
 - high-resolution analysis 818
 - histological sectioning 815
 - image capturing 815
 - image congruencing 816
 - image segmentation 816
 - internet database 819
 - internet platform 828
 - landmarks (external markers) 814
 - macro-driven image processing 816
 - macro 814
 - quantification 819
 - realignment 815
 - reconstruction 817–818
 - reference image 815
 - resin block 814
 - resolution 818
 - serial histological sections 813
 - serial sections 818
 - stereoscopic pairs of colour images 817
 - text-oriented gene database 828
 - theoretical mathematical model 819
 - three-dimensional model 816
 - visual gene expression database 828
 - whole-mount *in situ* hybridisation (WISH) 814
 - whole-organism screen 817 f.
- threshold cycle 69 f., 347
- TIGEM, *see* Italian Telethon Institute of Genetics and Medicine
- TIGR assembler 772–773
- TIGR gene indices 772, 777
- TIGR indices 832
- time-of-flight 641
- tissue microdissection 120 ff.
- tissue-specific Northern blot 97
- tissues 143
- TMIG-2DPAGE, *see* Proteome Database for Aging Research
- TOF, *see* time-of-flight
- TOGA portal 532, 533 ff.
- TOGA, *see* total gene expression analysis
- Tooth Development Database 860
- ToothPrint 869
- total gene expression analysis (TOGA) 520, 529
 - *in vitro* transcription 530
 - automation 531
 - bar-coded trays 532
 - capillary electrophoresis 531–532, 537
 - cleavage site 531
 - DGTools 533
 - digestion 530
 - digital addressing 531
 - digital sequence tag (DST) 531
 - DST, *see* digital sequence tag
 - electropherogram 535
 - laboratory information management system (LIMS) 531
 - LIMS, *see* laboratory information management system
 - method 530
 - plate-sealing device 532
 - principle 530
 - restriction endonuclease 530–531
 - sensitivity 536
 - sequence analysis 533
 - thermocycler 532
 - TOGA address 531–535
 - TOGA portal 532, 533 ff.
 - virtual TOGA 532

- Toxoplasma gondii* proteome map 869
- tracer 178
- trans activator 20
- trans-acting guide RNA 47 f.
 - trypanosome 47 f.
- trans-RNA splicing 39
 - spliced leader type 39
- TRANSCompel 801, 806
- transcript abundance 59
- transcript elongation 9
- transcript profile 72
- transcription 2 ff., 17, 173–174
 - elongation 17
 - elongation factor 17
 - primary rate 174
 - relative rate 173
 - termination 17
- transcription activator 15
- transcription elongation 33 ff.
 - *Drosophila* hsp70 36
 - arrest 34
 - bacteriophage λ Q protein 35
 - *c-myc* 35
 - eukaryotes 35
 - human immunodeficiency virus genes 35
 - pausing 34
 - prokaryotes 34
 - termination 34
 - tryptophan operon 34
- transcription elongation factor A1 36
- transcription factor 12, 16–19, 26, 27, 798, 802
 - basal 17
 - BRF, *see* TFIIB-related factor 18
 - cleavage factor 20
 - co-activator 16
 - conditional transcription factor 27
 - CTD-binding complex 19
 - developmental 27
 - DNA-binding site 12
 - elongation factor 19
 - mediator of transcription activation 19
 - regulatory transcription factor 27
 - signal-dependent transcription factor 27
 - specific 27
 - TBP 18
 - TFIIA 18
 - TFIIB 18–19
 - TFIIB-related factor 18
 - TFIID 18–19
 - TFIIE 18–19
 - TFIIF 18
 - TFIIH 18–19
 - TFIIIA 18
 - TFIIB 18
 - TFIIC 18
- Transcription Factor Database (TFD) 801, 810
- transcription initiation 9, 16, 174
 - open complex 16
- transcription *in vitro* (IVT) 412, 435–436, 439, 444
- transcription rate 177
- Transcription Regulatory Region Database (TRRD) 801
- transcription regulatory region 800
- transcription start site (TSS) 4, 800
- transcriptional activation 21
- transcriptional activator 799
- transcriptionally active nuclei 176
- transcriptome 9, 11
- transducer 667
- TRANSFAC 801 f., 803, 810, 828
- transfection 719
 - electroporation 720
 - glass bead loading 719
 - lipofusion 719
 - liposome 719
 - microinjection 719
 - phagocytosis 719
 - pinocytosis 719
 - receptor-mediated 719
 - transient permeability 720
- transfer RNA (tRNA) 3 ff., 11, 16, 138, 140
- translation 2 ff., 5 ff., 9, 11
 - elongation 11
 - initiation 11
- translation recoding 51
- TRANSPLODER 803
- transposon 847
- transposon mutagenesis 847
- tree harvesting technique 793
- TrEMBL 652, 853 f., 861, 869
- TRIPLES, *see* Database of Transposon-Insertion Phenotypes, Localisation, and Expression in *Saccharomyces cerevisiae*
- tRNA, *see* transfer RNA
- TRRD, *see* Transcription Regulatory Region Database 801
- tryptic digest 635
- TSS, *see* transcription start site
- TSSG 810
- TSSW 810
- α -tubulin 66
- β -tubulin 66
- turbidimetry 678
- two plasmid expression 722 f.
 - green fluorescent reporter protein 722
 - nuclear localisation signal sequence 722

- two-plasmid expression system 716
 - two-colour detection system 413
 - two-dimensional difference gel electrophoresis (2D-DIGE) 630
 - two-dimensional gel electrophoresis 373, 660
 - two-dimensional liquid chromatography (2D-LC) 649
 - Two-dimensional Polyacryl-Amide Gel Electrophoresis Database for Broncho-Alveolar Lavage Fluid Proteins (BALF 2D) 869
 - two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) 851, 854
 - Coomassie Blue 629
 - immobilised pH gradient 629
 - isoelectric focusing (IEF) 629
 - principle 629
 - resolution 630
 - staining 629
 - SYPRO 630, 629, 623
 - two-hybrid assay 660
- u**
- U6 snRNA 18
 - UAS, *see* upstream activator sequence
 - UBF, *see* upstream binding factor
 - ubiquitination 55
 - proteasome 55
 - UCE, *see* upstream control element
 - under-representation of rare species 384
 - unified atomic mass unit 639
 - unified library database 830
 - UniGene 468, 576, 777, 830–831, 836, 844
 - UniGene index 772, 832
 - UniLib 830
 - 3'-untranslated region 770
 - untranslated region (UTR) 849
 - up-converting phosphorus 435
 - upstream activator sequence (UAS) 17, 21
 - upstream binding factor 27
 - upstream control element (UCE) 17
 - upstream regulatory regions 802
 - upstream regulatory sequence elements 790
 - USAGE 556
 - UTR, *see* untranslated region 849
 - UTRblast 849
 - UTRdb, *see* Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs 849
 - UTRscan 849
 - UTRsite, *see* Databases of Sequences and Functional Elements of 5' and 3'-Untranslated Regions of Eukaryotic mRNAs 849
- v**
- VERSANT® RNA assay 471, 475
 - virtual Northern blot 260 ff., 365, 369–370, 776 f., 844
 - cap-finder 262, 264
 - cDNA population 261
 - conventional Northern blot 266
 - homopolymeric tailing 262
 - LA-PCR, *see* long and accurate PCR
 - ligation 262
 - long and accurate PCR (LA-PCR) 263–265
 - PCR amplification 265
 - performance 266
 - principle 262
 - RLM, *see* RNA ligase-mediated
 - RNA ligase 262
 - RNA ligase-mediated (RLM) 264
 - TAP, *see* tobacco acid pyrophosphatase
 - TdT tailing 262
 - terminal deoxynucleotidyl transferase (TdT) 262
 - tobacco acid pyrophosphatase (TAP) 265
 - visual basic 781
- w**
- W²H 778
 - W3C, *see* World Wide Web Consortium
 - Washington University Inner Ear Protein Database 869
 - weight matrix method 803
 - consensus 803
 - forbidden nucleotide 803
 - hash table 803
 - information theory 803
 - pseudo-count 803
 - thermodynamic consideration 803
 - transcription factor-binding site 803
 - Western blot, *see* Western blotting
 - Western blotting 147, 626 ff., 628, 629
 - principle 628
 - Whitehead Institute 832
 - whole-mount *in situ* hybridisation (WISH) 813–814, 838 f.
 - whole-organism screens 817 f.
 - WISH, *see* whole-mount *in situ* hybridisation
 - World Wide Web Consortium (W3C) 781
 - WORLD-2DPAGE 624, 861
 - WormPD™ 866
 - μWP, *see* micro wet printing
 - Wu-BLAST 773
- x**
- Xenopus laevis* database (Axelddb) 856
 - X-ray crystallography 657

X-ray film 438
XML, *see* Extensible Markup Language
XNA on Gold™ 425 f.
– auto-radiography 426
– biological interface 425
– biotin 425
– chemical interface 425
– chemiluminescence 426
– cross-contamination 425 f.
– fluorescence 426
– immobilisation 426
– probe 426
– streptavidin 425
– Teflon™ 426
– thioalkanes 425
– vapour deposition 425
Xpress-Screen® 475
Xpress-Screen® kit 476 f., 477
Xpress-Screen® mRNA detection assay 471
xProfiler 844
Xylella 2D database 870

y

Yeast 2D-PAGE Protein Database
(YEAST 2D-PAGE) 870
YEAST 2D-PAGE, *see* Yeast 2D-PAGE protein
database
Yeast Genome-wide Expression Database 860
Yeast Microarray Global Viewer (yMGV) 850
Yeast protein database COMPLUYEAST-
2DPAGE (COMPLUYEAST-2DPAGE
database) 870
Yeast Protein Map (YPM) 870
Yeast Proteome Database (YPD) 848
yeast two-hybrid 14
YEPD 854
yMGV, *see* yeast microarray global viewer
YPD, *see* Yeast Proteome Database
YPM, *see* Yeast Protein Map

z

Z-DNA 26
zinc finger 24, 27