ADVANCES IN BIOCHEMICAL ENGINEERING/BIOTECHNOLOGY

110

Series Editor T. Scheper

Volume Editors M. Werther · H. Seitz

Protein–Protein Interaction

Springer

110
Advances in Biochemical Engineering/Biotechnology

Series Editor: T. Scheper

Advances in Biochemical Engineering/Biotechnology

Series Editor: T. Scheper

Recently Published and Forthcoming Volumes

Bioreactor Systems for Tissue Engineering
Volume Editors: Kasper, C., van Griensven, M., Poertner, R.
Vol. 112, 2008

Food Biotechnology
Volume Editors: Stahl, U., Donalies, U. E. B., Nevoigt, E.
Vol. 111, 2008

Protein – Protein Interaction
Volume Editors: Seitz, H., Werther, M.
Vol. 110, 2008

Biosensing for the 21st Century
Volume Editors: Renneberg, R., Lisdat, F.
Vol. 109, 2007

Biofuels
Volume Editor: Olsson, L.
Vol. 108, 2007

Green Gene Technology
Research in an Area of Social Conflict
Volume Editors: Fiechter, A., Sautter, C.
Vol. 107, 2007

White Biotechnology
Volume Editors: Ulber, R., Sell, D.
Vol. 105, 2007

Analytics of Protein-DNA Interactions
Volume Editor: Seitz, H.
Vol. 104, 2007

Tissue Engineering II
Basics of Tissue Engineering and Tissue Applications
Volume Editors: Lee, K., Kaplan, D.
Vol. 103, 2007

Tissue Engineering I
Scaffold Systems for Tissue Engineering
Volume Editors: Lee, K., Kaplan, D.
Vol. 102, 2006

Cell Culture Engineering
Volume Editor: Hu, W.-S.
Vol. 101, 2006

Biotechnology for the Future
Volume Editor: Nielsen, J.
Vol. 100, 2005

Gene Therapy and Gene Delivery Systems
Volume Editors: Schaffer, D. V., Zhou, W.
Vol. 99, 2005

Sterile Filtration
Volume Editor: Jornitz, M. W.
Vol. 98, 2006

Marine Biotechnology II
Volume Editors: Le Gal, Y., Ulber, R.
Vol. 97, 2005

Marine Biotechnology I
Volume Editors: Le Gal, Y., Ulber, R.
Vol. 96, 2005

Microscopy Techniques
Volume Editor: Rietdorf, J.
Vol. 95, 2005

Regenerative Medicine II
Clinical and Preclinical Applications
Volume Editor: Yannas, I. V.
Vol. 94, 2005

Regenerative Medicine I
Theories, Models and Methods
Volume Editor: Yannas, I. V.
Vol. 93, 2005

Protein – Protein Interaction

Volume Editors: Meike Werther · Harald Seitz

With contributions by
M. Abu-Farha · M. Alamgir · T. Beissbarth · U. Beutling
W. M. Brown · C. S. Chan · T. G. Chappell · Y. Chen · K.-H. Cho
S.-M. Choo · D. Colon · F. Dehne · M. Dumontier · F. Elisma
J.-L. Faulon · D. Figeys · R. Frank · A. Golshani · P. N. Gray
J. R. Green · H. Guan · F. Henjes · E. Kiss-Toth · U. Korf
C. Löbke · H. Mannsperger · J. Martin · S. Martin · P. Parekh
S. Pitre · A. Poustka · B. E. Power · C. Schmidt · S.-Y. Shin
K. Städing · T. Stradal · W. Tan · A. Tresch · R. J. Turner
H. Wang · T. M. L. Winstone · S.-H. Woo · Y. Zhu

Springer

Advances in Biochemical Engineering/Biotechnology reviews actual trends in modern biotechnology. Its aim is to cover all aspects of this interdisciplinary technology where knowledge, methods and expertise are required for chemistry, biochemistry, micro-biology, genetics, chemical engineering and computer science. Special volumes are dedicated to selected topics which focus on new biotechnological products and new processes for their synthesis and purification. They give the state-of-the-art of a topic in a comprehensive way thus being a valuable source for the next 3–5 years. It also discusses new discoveries and applications. Special volumes are edited by well known guest editors who invite reputed authors for the review articles in their volumes.
In references *Advances in Biochemical Engineering/Biotechnology* is abbeviated *Adv Biochem Engin/Biotechnol* and is cited as a journal.

Springer WWW home page: springer.com
Visit the ABE content at springerlink.com

ISBN 978-3-540-68817-4 ISBN 978-3-540-68820-4 (eBook)
DOI 10.1007/978-3-540-68820-4

Advances in Biochemical Engineering/Biotechnology ISSN 0724-6145

Library of Congress Control Number: 2008933498

Cover design: WMXDesign GmbH, Heidelberg
Typesetting and Production: le-tex publishing services oHG, Leipzig

Printed on acid-free paper

9 8 7 6 5 4 3 2 1 0

springer.com

Series Editor

Volume Editors

Dr. Meike Werther

Privatinstitut für
Biomedizinische Forschung
Friedrich-Ebert-Str. 33
34117 Kassel, Germany
meike.werther@gmx.de

Dr. Harald Seitz

MPI for Molecular Genetics
Dept. Vertebrate Genomics
Ihnestraße 73
14195 Berlin, Germany
seitz@molgen.mpg.de

Editorial Board

Honorary Editors

Advances in Biochemical Engineering/Biotechnology Also Available Electronically

For all customers who have a standing order to Advances in Biochemical Engineering/Biotechnology, we offer the electronic version via SpringerLink free of charge. Please contact your librarian who can receive a password or free access to the full articles by registering at:

springerlink.com

If you do not have a subscription, you can still view the tables of contents of the volumes and the abstract of each article by going to the SpringerLink Homepage, clicking on "Browse by Online Libraries", then "Chemical Sciences", and finally choose Advances in Biochemical Engineering/Biotechnology.

You will find information about the

- Editorial Board
- Aims and Scope
- Instructions for Authors
- Sample Contribution

at springer.com using the search function.

Color figures are published in full color within the electronic version on SpringerLink.

Attention all Users of the "Springer Handbook of Enzymes"

Preface

Individual organisms are defined by their genetic code. During development and as a response to external stimuli the genetic information is translated into a well-defined answer resulting in the expression and modification of proteins. The processes that control protein–protein interactions (PPI) are presently mostly described in terms of individual protein–protein interactions. In vivo such interactions are part of complex molecular interaction networks that are highly dynamic in time and space. On the basis of quantitative experiments, it would be possible to understand such complex biological systems leading to an unraveling of these networks and allowing them to be caught in quantitative and predictive models.

This textbook illustrates the rise of a relatively new area of biology. The shifting of research from the structural assembly of cells and whole organisms to metabolic diversity led to the beginning of interactomics. This field has arisen from the increasing importance of molecular biology and biochemistry in basic research as well as in prognostics and prevention of diseases in connection with biomarker development.

The behavior, morphology, and response to stimuli in biological systems are predetermined by the interactions between their components. These interactions, as we observe them now, are therefore shaped by genetic variations and selective pressure. With the understanding of molecular interactions the biology is getting easier to survey. The characterization of protein interactions can contribute to the understanding of many processes in nature.

Knowledge of the different types of biological macromolecules and increasing numbers of whole genomic studies facilitate the elucidation of cellular processes. Whether it is genomics, transcriptomics, proteomics, interactomics, or metabolomics, the full complement of genomic information at different levels can be compared between different organisms to reveal similarities or differences and even to provide consensus models.

A protein's role is reflected in its interaction with others. Much of the function of novel proteins can be predicted from identifying its interaction partners, and from characterizing its localization within the cell. There is a need to employ a combination of approaches to overcome this deficit in understanding of gene function. A number of experimental strategies have been developed and applied at a large scale with the aim to decipher gene function through identifying proteins interacting with the gene product of interest.

Protein–protein interactions are key elements for normal functioning of a living cell. A detailed description of the protein interactomics field is given in this book. We first give an introduction to the different large-scale experimental approaches used to discover protein–protein interactions. Single PPI validation techniques such as co-immunoprecipitation or fluorescence methods are then presented because they are becoming more and more integrated in a global PPI discovery strategy.

Understanding gene function at the molecular level requires characterization of protein–protein interactions. Formation of multiprotein complexes is a dynamic process, where the composition of the complex is altered to allow physiological interactions to take place. To obtain a deep understanding of such processes it is necessary to combine a number of experimental approaches. Most of these processes are driven by technologies and assays allowing automatization and parallelization of the experiments. In addition, we believe that a detailed analysis is still necessary to gain a real understanding of how proteins interact and the way they exchange their information.

Proteins are modified with a wide variety of diverse chemical groups, such as phosphate or amino groups. Stimuli of cells results in an altered modification of proteins. Every modification adds information to the proteins. That information is transformed via protein–protein interaction through the cell and results in a specific response. This means that of each protein a number of differently modified forms exist with clearly defined interaction partners and duties.

In contrast to DNA and RNA, proteins cannot be amplified to achieve sufficient amounts of material to analyze. The success of high-throughput DNA sequencing projects followed by techniques like DNA microarrays and second-generation sequencing approaches is determined by the physical and chemical behavior of DNA.

The wide spectrum of techniques dealing with protein–protein interaction provided us with the opportunity to choose the most interesting and relevant ones. Therefore, the book focuses on certain aspects ranging from sample extraction to different methods of interaction measurements. With this background we asked researchers from different universities and departments to write about their experiences in handling and achieving their aims.

In part, this book has been written as a recruiting guide as new generations of researchers are needed to move interactomics forward. An increased knowledge in this field would help in fighting the main diseases by confirmation of diagnosis and specifically targeted drug delivery.

The book owes its existence to the contributions of many people. We would like to thank Springer-Verlag for their interest in this topic and for continued support and help during the preparation of this volume, especially Dr. Marion Hertel, Chemistry Editor, and Ms. Ulrike Kreusel, Chemistry Desk Editor. We would also like to thank the scientists who spent their time in preparing up-to-date chapters allowing an in-depth view into the individual techniques. Finally, we would like to acknowledge all colleagues who made this book possible.

Habent sua fata libelli!

In this spirit, we hope that much of what you will learn from this book will be useful in understanding many aspects of bioanalytics.

Meike Werther
Harald Seitz

Contents

Adv Biochem Engin/Biotechnol (2008) 110: 1–24
DOI 10.1007/10_2007_092

Published online: 25 January 2008

Advanced Technologies for Studies on Protein Interactomes

Hongtao Guan · Endre Kiss-Toth (✉)

Cardiovascular Research Unit, University of Sheffield, Royal Hallamshire Hospital,
Glossop road, Sheffield S10 2JF, UK
E.Kiss-Toth@sheffield.ac.uk

Abstract One of the key challenges of biology in the post-genomic era is to assign function to the many genes revealed by large-scale sequencing programmes, since only a small fraction of gene function can be directly inferred from the coding sequence. Identifying interactions between proteins is a substantial part in understanding their function. The main technologies for investigating protein–protein interactions and assigning functions to proteins include direct detection intermolecular interactions through protein microarray, yeast two-hybrid system, mass spectrometry fluorescent techniques to visualize protein complexes or pull-down assays, as well as technologies detecting functional interactions between genes, such as RNAi knock down or functional screening of cDNA libraries. Over recent years, considerable advances have been made in the above techniques. In this review, we discuss some recent developments and their impact on the gene function annotation.

Keywords cDNA library screening · Protein–protein interactions · Proteomics

1 Introduction

Following the success of the high-throughput sequencing projects, one of the first being the Human Genome Programme, followed by a host of other initiatives, such as the expressed Sequence Tag (EST) sequencing efforts of RIKEN and the IMAGE consortia, an unprecedented quantity of primary DNA sequences from a variety of species became available. A major task of the post-sequencing phase is to identify candidate genes with a particular property/function. However, despite extensive bioinformatics efforts to predict the function of genes through knowledge accumulated on orthologues, the biological function of the majority of the genes in man are still unknown. It was estimated that only 30% of predicted human genes have any function assigned (Kiss-Toth et al. 2004b). Therefore, there is a need to employ a combination of approaches to overcome this deficit in our understanding of gene function, in general, and mammalian gene function, in particular. A number of experimental strategies have been developed and applied at a large scale with the aim to decipher gene function through identifying proteins interacting with the gene product of interest. Since a novel protein's role is reflected in its interactions with others, much of the function can be predicted from identifying its interacting partners, and from characterizing its localization within the cells (Blackstock and Weir 1999). Given the shear volume of novel sequence information, hence a large number of proteins of unknown function, such characterization of protein function is increasingly done on a global scale, using high-throughput tools. These exercises are often referred to as proteomics studies.

The term "proteome" was coined for the first time in 1995 to describe the protein complement of a genome (Wasinger et al. 1995). Proteomics can be divided into two areas: "*expression proteomics*, which studies global changes in protein expression, and *cell-map proteomics*, the systematic study of protein–protein interactions through the isolation of protein complexes" (Blackstock and Weir 1999). There is a great emphasis in research in the post-genomics era on understanding the questions how do proteins interact and what are their functions, at a global scale? Since most drug targets are proteins, studying protein–protein interactions at a system level is key in attempts to get closer to these goals. In this review, we will discuss some of the recent advances in key methods studying protein–protein interactions.

2 Protein (Antibody) Microarray/Protein Chips

Recent advances in high-throughput technologies enable researchers to map cellular networks at the protein level. A widely adapted platform for such studies is the protein microarray (an analogue of the DNA microarray, also commonly know as the DNA chip, or gene chip).

One version of this array technology makes use of antibodies raised against a defined group of proteins of interest. These antibodies are anchored onto the surface of a support, such as a glass slide, and then a mixture of protein samples is run over the surface to allow the antigens present to bind to their cognate antibodies. This technique has been used to study changes in protein expression levels after cells were treated with a range of chemicals and/or extracellular stimuli of interest. Sreekumar et al. (2001) reported their study using an antibody array to identify protein expression levels in cancer cells exposed to ionizing radiation, and identified both known radiation-regulated proteins (p53 and DR5) and a number of novel proteins (DFF40/CAD and CEA). Another example was reported by Haab et al. (2001) where they used a robotic device to spot specific antibody or antigen solutions on the surface of microscope slides. The experimental protein samples were fluorescently labeled in this study. The sensitivity and accuracy of the antigen–antibody interaction detected by this system were concluded to be satisfactory for both clinical and research applications to enable characterization of concentration for a large number of proteins in complex solutions.

Other modified protein arrays use different immobilized probes on the slide surface. The probes can be short peptides, aptamers (oligonucleotide/peptide), polysaccharides, allergens, or small synthetic molecules. The protein microarray has been proven to be a practical and sensitive tool for studying protein–protein interactions, protein–nucleic acid, protein–small molecule and protein drug interactions as well (review Zhu et al. 2003). In addition to identifying novel proteins interacting with the target protein, the protein microarray can also be used to study the kinetics of protein–protein interactions through real-time detection methods. This concept was adopted by Sapsford et al. (2001) who studied the kinetics of antigen–antibody interactions.

3
Yeast Two-Hybrid System

The yeast two-hybrid system was originally developed by Fields and Song (1989) to study protein–protein interactions, by exploiting the yeast *Saccharomyces cerevisiae* transcriptional activator GAL4. The strategy in this study made use of the GAL4 protein, which is composed of two separable and functionally distinct domains: an N-terminal domain which binds to specific DNA sequences (binding domain/BD), and a C-terminal domain containing acidic regions, necessary to activate transcription (activator domain/AD). Therefore, a system of two hybrid proteins, where the GAL4 DNA-binding domain (BD) was fused to a protein "X" (bait) and a GAL4 activating region (AD) to a protein "Y" (prey), was set up. If proteins X and Y can form a protein–protein complex so that the complex brings the two domains of GAL4 into proximity, transcription of the reporter gene regulated by GAL4 is initiated (Fields and Song 1989).

Over the years, scientists modified the traditional two-hybrid system technique. Whilst these derivatives all use similar principles, the various modifications enabled "fine-tuning" of the approach for specific purposes.

1. A one-hybrid system was developed to study protein–DNA interaction, where the protein binds to AD and the DNA fragment of interest is cloned upstream of the reporter gene. If the protein binds to the specific DNA (promoter) of interest, transcription of the reporter gene is initiated (Meng and Wolfe 2006).
2. In order to characterize protein function, sometimes it is desirable to select for disruption of a particular interaction by introduction of point mutations, deletions or through the use of protein or pharmacological inhibitors. Thus, reverse and counter-selection two-hybrid systems were developed. In these systems, the wild-type protein which can interact with the bait, causes sensitivity to selection reagents (e.g. cytotoxic compounds) and the cells die. Only cells expressing proteins harboring the mutations disrupting the interaction will survive (Shih et al. 1996; Vidal et al. 1996). Therefore, this approach can be used to identify novel mutants of the prey that no longer bind to the other protein of interest (the bait). Similarly, chemical compound libraries can also be screened for entities with the ability to block particular protein–protein interactions.
3. One of the major limitations of the traditional two-hybrid system has always been that some proteins, especially components of signal transduction pathways in higher eukaryotes, need post-translational modification for their ability to interact with others and thus they require modifying enzymes which are not present in yeast. The three-hybrid system was developed and applied to facilitate the post-translational modification of the prey proteins, by co-transfecting the necessary enzyme into the system, to ensure that prey proteins are functional in the host. This technique enables rapid mapping of modifications required for a certain inter-protein interaction. The most common such modification is protein phosphorylation, which is most often induced by specific stimuli and requires the expression of particular kinases. An example of the application of the Y3H strategy was published by Osborne et al. in their library screening study to identify proteins which specifically interact with an immunoreceptor tyrosine-based activation motif (ITAMs)-containing IgE receptor-derived, phosphorylated bait (Osborne et al. 1995). In the bait, the gamma subunit of the high-affinity IgE receptor, FcεRI, was used to isolate a novel SH2-containing family member (interactions between FcεRI cytoplasmic tail and the Syk or Lyn SH2 domains), which requires the phosphorylation

of the ITAMs by tyrosine kinases. A plasmid encoding the tyrosine kinase was introduced together with the "bait" and the "prey". This method was initially used for characterizing interactions that are mediated by tyrosine phosphorylation, but can be adapted to other post-translational modifications.

Another use of the three-hybrid system is detection of weak interactions between multiple proteins. In most cases, proteins bind to a number of other proteins and form large, multi-component complexes containing both weak and strong interactions. In order to identify novel proteins that weakly interact with a protein of interest, a known interacting protein can be co-expressed and this may supply a bridge and thus strengthen the interaction for the novel proteins with lower affinity interaction (Tirode et al. 1997; Tomashek et al. 1996).

4. In addition to the classical *Saccharomyces cerevisiae* system, other hosts, such as *E. coli*, have also been used. This was proposed to have multiple advantages over the yeast system, such as fast growth, higher transformation efficiency, nuclear localization not required, domains with eukaryotic activation domains do not activate *E. coli* transcription, and fewer indirect interactions involving bridging by endogenous proteins (review Hu 2001). A recent study exploiting the *E. coli* two-hybrid system was based on the twin-arginine translocation (Tat) pathway to identify interacting proteins in this pathway. Two reporter systems via the Tat pathway were used, one based on growth on selective media (maltose based) and the other enzymatic assay using a chromogenic substrate. Compared to other studies, the development of this *E. coli* two-hybrid system improved the accuracy of proteome-wide two-hybrid analyses (Strauch and Georgiou 2007).
5. In the original two-hybrid system, interacting fusion proteins (both prey and bait) need to be transported into the nucleus in order to activate the transcription of the reporter. This limits the interactions detected by the system. For instance, using full-length trans-membrane proteins as baits is problematic, due to misfolding or lack of localization in the nucleus. Some strategies to allow the two-hybrid system to take place in the cytoplasm and membrane were developed to circumvent this pitfall. Instead of using nuclear transcription factors to construct two fusion proteins, β-galactosidase was split into two fragments and re-constituted through a bait/prey interaction, and the β-galactosidase activity served as a measurement of the strength of the interaction of the bait and the prey. Thus the protein–protein interaction could be studied in the cytoplasmic milieu (Rossi et al. 1997). In another approach, the Ras-controlled signaling cascade on the *Saccharomyces cerevisiae* plasma membrane, which harbors temperature-sensitive Ras Guanine Exchange factors (GEF) Cdc25-2, was used. GEFs stimulate the transition of Ras between an inactive GDP-bound form and an active GTP-bound form. Sos protein is one of the

mammalian GEFs. If the Sos is recruited to the membrane, it will stimulate the transition of the Ras, hence the initiation of the signaling cascade, and allow growth at the non-permissive temperature (37 °C). Therefore, a bait (X) fused with Sos can identify membrane proteins (Y, prey) which interact with the bait. The readout of this system is based on the survival of *Saccharomyces cerevisiae* and the ability to grow when the temperature is increased from 25 °C to 37 °C (Aronheim et al. 1994; Aronheim et al. 1997).

6. A conceptually similar strategy to the yeast two-hybrid system for detecting protein–protein interaction studies in mammalian cells was described by Tavernier et al. (2002). In their studies, they established a system called MAPPIT, where they genetically modified the JAK/STAT pathway at the type I cytokine receptor level. They used the leptin receptor, containing a Y1138F mutation, which eliminates the recruitment of STAT transcription factor to the receptor, and hence the ligand-bound receptor cannot signal. To complement this signaling deficiency, bait and prey vectors were introduced into the system, where the bait is engineered on the C-terminal of the Leptin receptor, and the prey is fused with the C-terminal part of gp130, which contains four functional STAT3 recruitment sites. Once the prey binds to the bait, the C-terminal part of gp130 can act as a harbor for recruiting STAT3. Subsequently, STAT3 is phosphorylated and activated. The active transcription factor can then translocate into the nucleus and initiate the expression of a reporter, e.g. luciferase.

To identify novel proteins which interact with a protein of interest, a cDNA library can be constructed where proteins are expressed in fusion with the GAL4 AD, whilst the test protein (bait) is fused to the BD domain. However, large-scale screenings for systematic detection of interacting proteins against a number of baits (or increasingly whole signaling pathways) often require automation. Two main approaches have been developed to enable the deployment of robotic platforms, as described below (review Fernandes 1998; Knudsen et al. 2002). These include the matrix (or array) approach, and the library screening approach. In the matrix format, a defined set of yeast clones containing cDNA inserts as fusions with the BD or AD are screened against each other in a grid format by mating. This approach was used to study the interactions between *Drosophila* cell cycle regulators (Finley and Brent 1994), and protein–protein interactions in *Saccharomyces cerevisiae* (Ito et al. 2001; Uetz et al. 2000). Interestingly, the two studies performed by Ito et al. and Uetz et al. did not result in a fully overlapping dataset, but both greatly contributed to the establishment of the yeast Protein Interaction Map which leads to a more thorough understanding of the gene function in a single cell system.

In the second approach, a random library or a library of pooled open reading frames are screened against another library, for example, to decipher protein interaction maps in microbes, such as *Helicobacter pylori* (Rain et al. 2001). Work to catalogue all possible protein–protein interactions for a given genome in worms and yeast is ongoing although this requires a very substantial effort (Hudson et al. 1997). However, in more complex organisms (such as mammals) which contain a lot more complex genomes, and a range of specialized cell types, it is more demanding, laborious and difficult to exhaustively complete the protein–protein interaction map. Therefore, a more realistic strategy to characterize a protein–protein interaction network is to identify novel interacting proteins in a discrete cellular signaling pathway or cellular process. For example, identifying a complete set of proteins involved in spliceosome function has given us useful knowledge on pre-mRNA splicing, and has also provided a model platform for studying other cellular machines (Fromont-Racine et al. 1997).

In summary, the two-hybrid system is a proven, invaluable tool of cell biology. It has a high sensitivity, enabling detection of weak and transient interactions. The experimental setup is relatively straightforward and enables simultaneous detection and characterization of protein–protein interaction using a single protocol. It can be used to identify novel proteins interacting with the bait protein, without the need of any a priori knowledge. Since the experiment is undertaken in an in vivo system (e.g. yeast, *E. coli*), proteins of interest are likely to fold appropriately and thus the screen is more likely to detect genuine interacting proteins.

However, the two-hybrid system also has its limitations. Both false-positives and false-negatives can occur and are considered to be the most serious technical problems. False-positivity can arise for a variety of reasons: proteins interacting with the DNA upstream of the reporter gene or with proteins that interact with the promoter sequence are prone to detection as false-positive in the two-hybrid system. For example, the original two-hybrid system was designed to activate transcription through RNA polymerase II. Therefore, it is problematic to identify novel proteins which interact with the RNA polymerase II activators (as bait). False-negativity can also be caused by a range of reasons: some proteins fused with BD or AD cannot localize in the yeast nucleus, cannot fold properly, are not functional when expressed as a fusion protein, are toxic to the host, sometimes the appropriate post-translational modification does not take place, or the potential interacting protein is not sufficiently represented in the library. Therefore, other independent methods (e.g. techniques based on biological function) should be used in parallel to confirm and verify the hits detected by the two-hybrid system.

4
Two-Dimensional Gel Electrophoresis

At the early phase of the proteomics era, two-dimensional gel electrophoresis was the main tool for obtaining a general translation profile of a genome. One of the major strengths of the approach lies in its high resolution. In a standard setting, a protein mixture is firstly separated by isoelectric focusing and then in the orthogonal direction by molecular weight as the normal SDS-PAGE (O'Farrell et al. 1977). Although it is a relatively established technique, 2D electrophoresis still remains a powerful tool in proteomics. Interacting partners to a protein of interest can be identified by immunoprecipitation of that protein, followed by 2D electrophoresis. In order to clarify the identity of the various dots on the gel, two-dimensional gel electrophoresis is frequently coupled to affinity chromatography and mass spectrometry.

5
Mass Spectrometry

Mass spectrometry (also known as mass spectroscopy, mass-spec or MS) is an analytical technique used to measure the mass-to-charge ratio of ions, by generating a mass spectrum, which is a specific trait of a physical sample.

Mass spectrometry has recently become a widely used method for the characterization of proteins. A typical mass spectrometer consists of at least three components: an ionization device, a mass separator, and a detector. It can only separate molecules that are charged in the gas phase, and can only separate either positively or negatively charged molecules at a time. The two primary methods for ionization of whole proteins are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). ESI is able to transfer samples from a liquid phase to a gas phase under atmospheric pressure. MALDI ionization is a technique where samples of interest (peptides and protein complex) are co-crystallized with an acidified matrix (a small molecule). Since electrospray provides weak ionization, non-covalent interactions between molecules can be preserved. Not only simple dimers but also large assemblies such as virus capsids and heterogeneous, asymmetric ribosomes can therefore be ionized and maintained intact (Hanson et al. 2003; Rostom et al. 2000).

The matrix absorbs in the ultraviolet wavelength range and dissipates the absorbed energy thermally. The end-result of this dissipation is that a number of charged proteins/peptides of interest are present in the gas phase (Nelson et al. 1994). Mass analysis of proteolytic peptides is a much more popular method of protein characterization, as less sophisticated instrument designs can be used, resulting in more affordable setups. Additionally, sample preparation is easier once whole proteins have been digested into smaller peptide fragments. The most widely used instrument for peptide mass analysis is the quadruple ion trap. Multiple stage quadruple-time-of-flight and MALDI time-of-flight instruments also find use in this application.

To identify proteins within a complex or to study protein–protein interaction, MS is often coupled to other protein separation techniques, such as affinity purification, HPLC or two-dimensional SDS-PAGE. A frequently used method of exploiting affinity purification to study protein–protein interaction is to conjugate an epitope such as FLAG (sequence DYKDDDDK) to the N-terminal end of a gene of interest (for identifying novel proteins interacting with the product of this gene, the bait protein). The FLAG tagged gene is inserted into a vector and transfected into a mammalian cell line (such as HEK293 cells), and the cell extract is run over a column containing covalently attached antibody (in this example, anti-FLAG). Thus, the proteins interacting with the bait protein are recovered specifically.

MS is used to identify proteins in two major systems: peptide mass fingerprinting and tandem MS (MS/MS). Peptide mass fingerprinting uses the masses of proteolytic peptides as an input to search a database of predicted masses that would arise from an in silico digestion of a computer generated list of known proteins by the same enzyme (usually trypsin). In this approach, protein entries in the database are ranked according to the number of peptide masses which match their predicted trypsin digestion pattern. MS/MS is becoming a more popular experimental method for identifying proteins, by matching the observed fragment masses with a database of predicted masses for one of many given peptide sequences. In addition, MS/MS has also been applied to detect high molecular weight complexes, thus can be used for studying protein–protein interactions. In the context of multiprotein complexes, including mega-Dalton particles (Ilag et al. 2005), MS/MS can provide insight into their subunit stoichiometry and composition as well as their overall architecture. In addition, a peptide with post-translational modification, for example, phosphorylation or covalent tagging with a polysaccharide, the mass/charge ratio is different from the un-modified peptide. This difference can be detected by MS/MS followed by analysis using computer software, specifically designed to detect, identify and locate the modified sites on a peptide. Thus, MS/MS has become an indispensable tool in proteomics, to identify proteins and characterize post-translational modifications (Blackstock and Weir 1999).

Many important biological processes in cells are mediated by assemblies of ten or more proteins (Alberts 1998). The MS technique complements other traditional structural biological platforms such as NMR (nuclear magnetic resonance) and X-ray crystallography very well. In addition to decipher the composition of multi-protein complexes, MS/MS can also be used to identify substrates (or ligands) binding in large heterogeneous assemblies (van Duijn et al. 2005). The advantage of MS over other techniques in studying protein–protein interactions lies in its unambiguous identification of proteins and the accurate measurement of peptide and protein masses (Figeys et al. 2001).

Here we give an example of exploiting MS/MS to identify novel proteins in signaling networks. By 1997, researchers had characterized members of the NF-κB transcription factor family (p50/NF-κB1 and p65/RelA), which regulate the expression of many inflammatory genes and the family of inhibitors for these transcription factors, called IκB. It was also known that the phosphorylation and subsequent degradation of IκB initiates the critical step of activation of NF-κB which enables p65/RelA to be transported into the nucleus. However, the IκB kinases had not been isolated and characterized. Mercurio et al. (1997) identified IKK-1 and IKK-2 as the IκB kinases, by exploiting MS, in combination with chromatography and bioinformatics-searching of EST databases. From a TNF-α stimulated HeLa whole cell extract, a protein complex, which contains the TNF-α inducible IκB kinase activity (also called "IKK signalsome"), was fractionated by gel-filtration chromatography. After the chromatography, the active fraction was separated by SDS-PAGE, and two bands (85- and 87-kD) were excised, digested by trypsin, and then analyzed by high mass accuracy MALDI peptide mass mapping. The peptide sequence analysis revealed IKK-1 and IKK-2 as the relevant kinases, which phosphorylate IκB.

6 Protein Engineering

Protein engineering is a term for an experimental strategy to modify proteins in order to optimize their specific properties. These could include binding affinity to other proteins, catalytic activity under "non-physiological" conditions, etc. Such efforts require expertise from many disciplines, such as bioinformatics, mathematics, in silico protein design, genetics and protein biology. Two main strategies are in use for protein engineering.

The first approach is rational design, based on the detailed knowledge of the structure and function of the target protein in order to make the desired changes. Site-directed mutagenesis techniques are exploited to introduce defined structural alterations for the protein of interest. Computational protein design algorithms have also been developed and are in use to help design a mutated novel protein. However, when the high-resolution model of protein structure is not available, or in the case of recently identified proteins, this strategy may not be the optimal one. An example in the use of rational design was reported recently by Liu et al. (2007). In their study, in order to test the feasibility of the computational redesign of protein–protein interactions, the authors transferred the functional epitopes from one protein to another, after computational algorithm modeling. They took the erythropoietin receptor system as their target of protein–protein interface redesign, and engineered a mutant rat pleckstrin homology domain of phospholipase C-δ1 which could bind to the erythropoietin receptor in a cell-based assay. The affinity of binding and the biological functions (using luciferase as a reporter readout) of the non-natural protein–protein interface were tested and proved to be successful. Their study demonstrated that computational redesign is a useful tool for engineering non-natural protein–protein interaction pairs.

The second strategy is known as "directed evolution". In this approach, random mutagenesis is used to generate a pool of protein mutants, and selection pressure is applied, which gives an advantage to the desired mutants. When necessary, further rounds of selection for more mutants can be applied. This strategy is amenable for high-throughput studies for protein–protein interaction work. One application of this approach is to make novel, high-affinity antibodies and enzymes (for enzyme engineering and biocatalysis) from protein/peptide libraries (reviews Fujii 2007; Kaur and Sharma 2006; Woycechowsky et al. 2007).

Additional approaches often referred to as protein engineering involve the synthesis of small molecule leads (Arkin and Wells 2004), cross-linked interfacial peptides (Walensky et al. 2004), and α-helix mimetics (Kritzer et al. 2004), to select for inhibitors of particular protein–protein interactions. Walensky et al. reported their chemical strategy, termed hydrocarbon stapling, to generate a modified amphipathic α-helical BH3 segment (BH3 peptide), an important death domain of the Bcl-2 members. The resulting "stabilized α-helix BH3 peptides" were demonstrated to be helical, protease-resistant, and cell permeable molecules, binding with increased affinity to multidomain Bcl-2 member pockets. These chemically modified peptides may provide a useful tool for studying modulations of protein–protein interactions. Similarly, Kritzer et al. chemically substituted a residue on the p53 trans-activation domain to make the α-peptide more stable and proteolysis resistant. After the modification, the peptide could activate apoptosis. These techniques greatly enrich our knowledge of the nature of protein–protein interactions.

7
In Vitro Pull-Down Assays

A range of different methods make use of the same principle: co-precipitation of protein X with its interacting proteins. Thus, in vitro pull down is used in various designs, as discussed below.

Phage display is a test to screen for protein–protein interactions by integrating multiple genes from a "gene bank/library" into bacteriophages. The principle of phage display is based on a protein of interest (X), which is coated onto the surface of a plastic dish. A gene bank from an organism's genome is expressed in a library as fusions with the coat-protein of the bacteriophage, so that they both are displayed on the surface of the viral particle. The phage display library is then added to the coated dish. The phage particles displaying proteins, which are interacting with the protein X, remain attached to the dish, whereas all others can be washed away. In addition, the phage particles also contain the encoding gene. Therefore, a physical linkage between genotype and phenotype is established. Sequences of the interacting peptides can be determined through sequencing the encapsulated DNA. There are two ways of displaying library members (e.g. peptides) to the coat proteins of M13 phage. The first choice is to display on the major coat protein (protein-8), and this gives polyvalent display. The second choice is to display on the minor coat protein (protein-3), and with optimized engineering, this can give one display in each phage particle. One example of using phage display to identify/discover high affinity ligands/peptides to a targeted protein was demonstrated by Deshayes et al. (2002). In their study, they selected insulin-like growth factor (a 70-residue peptide hormone) as the target to investigate various epitopes of it. They demonstrated that using the improved phage display method they had developed, recognizable motifs of the peptides, which were responsible for binding to the insulin-like growth factor receptor, were identified from a large peptide library.

8
FRET, Protein Fragment Complementation Assay (PCA)

Fluorescence resonance energy transfer (FRET) is a widely used method to study protein–protein interactions based on energy transfer between two chromophores, where the emission energy of one (donor) chromophore overlaps with the excitation energy of a second molecule (acceptor). When the donor chromophore is excited at its specific fluorescence excitation wavelength, some of the excited energy is transferred to the second molecule. Therefore, to apply this principle into protein–protein interaction studies (Tsien et al. 1993), the protein of interest (X) is tagged by the donor molecule, and the test protein (or a cDNA library) is tagged by the acceptor. If the proteins under investigation interact with each other, this brings the donor and acceptor into a close proximity (1–10 nm), and the fluorescence emitted from the acceptor can be detected by fluorescent microscopy or flow cytometry. FRET is used for studying protein–protein, protein-DNA interactions and protein conformational changes. Green fluorescent protein (GFP) has been used as an indicator of the cellular physiology, mainly in studies of intracellular proteins. The most popular FRET pair for biological use is a cyan fluorescent protein (CFP)-yellow fluorescent protein (YFP) pair, as this system enables the detection of protein–protein interactions in real time, using live cells. Both CFP and YFP are specific variants of the green fluorescent protein (GFP). Another frequently utilized protein pair includes the blue fluorescent protein and enhanced green fluorescent proteins (eGFP), where BFP and eGFP are also variants of GFP. However, the BFP is only weakly fluorescent, and hence is not suitable for applications other than fluorescent microscopy and flow cytometry (review Pollok and Heim 1999).

Since FRET requires external illumination to initiate the fluorescence transfer, this can lead to a substantial background noise in the results from direct excitation of the acceptor, or photo-bleaching. To circumvent some of these limitations, an alternative approach, the bioluminescence resonance energy transfer (BRET), has been developed by using a bioluminescent luciferase (typically the luciferase from *Renilla reniformis*) to produce an initial photon emission (as an energy donor), which is then able to excite a GFP protein variant, e.g. YFP (as an energy acceptor). Since the donor in BRET produces energy through chemiluminescence, it is more amenable to small animal imaging, and hence results in greater sensitivity in live experimental objects, compared to FRET systems (De and Gambhir 2005).

Protein fragment complementation assay (PCA) is a novel method developed recently to study protein–protein interactions. PCA uses two fragments derived from a single fluorescent protein (e.g. GFP, YFP), based on the notion that some fluorescent proteins or transcription factors are modular and that their activity can be re-constituted when the two domains of the protein come to a close proximity to each other. The fluorescent protein (most often the YFP) is divided into two fragments, each of which is fused to a protein of interest. If the proteins interact with each other, this will enable the two YFP fragments to form a functional fluorophore (Michnick 2003). Not only can PCA be used to identify novel protein–protein interactions, but it is also useful to study perturbations of interactions by addition of other agents into the system. We and others have exploited the PCA to study protein function and map protein–protein interactions. In our group, we have adopted PCA in the study of the interactions between MKKs and Tribbles both in epithelial cell lines and in primary smooth muscle cells (Sung et al. 2007). Remy et al. has reported the application of PCA in a cDNA library screen with the aim to map PKB signaling networks (Remy and Michnick 2004a,b). These studies led to the characterization of novel components of the PKB activated signaling systems.

Whilst both the FRET and PCA technology requires similar instrumentation, they have different strengths and limitations, therefore complement each other's utility. FRET allows high spatial resolution assays of protein–protein interactions in living cells. The distance between the two fluorophores (donor and acceptor) needs to be less than 10 nm for FRET to occur, and direct protein–protein interaction occurs on a similar spatial scale. However, FRET is relatively insensitive in some settings, and has a very narrow dynamic range. The two fluorophores need to have similar brightness of fluorescence. As a consequence, FRET cannot easily be used for large-scale studies (in the absence of a very significant automation), since the two components for FRET need to be expressed in the same cells at optimal levels. The donor/acceptor expression levels need to be in the range of 10 : 1 to 1 : 10 (Chen et al. 2006). For example, in a cDNA library screen, it is difficult to make all the individual cDNA expressed at optimal levels to the bait fusion protein. The FRET signals can be undetectable if the two fluorophores are not aligned, or if they are not within 10 Angstrom. This can result in false-negatives, even if the two proteins of interest bind to each other but the two fluorophores are not in the right conformation within the complex (Piston and Kremers 2007). Further, to prevent the occurrence of false-positive signals, the two fluorophores used should not interact with each other. When acquiring FRET images, since the emission spectrums of both donor and acceptor overlap, there is a signal cross-talk between the donor and acceptor fluorophores. Thus, careful optimization of the genetic modification of the fluorophores (e.g. generating optimal GFP variants) and operation of the imaging acquisition (including fine-tuning of the fluorescent microscope) are required (Piston and Kremers 2007).

By contrast, PCA is easier to scale up and the protein expression levels do not need to be optimized for the two fragments to form an active 3D-structure. In addition, since the principle of PCA is based on the folding of the fluorescent protein structure, the dynamic range of a fluorescent signal is maximal (Michnick 2004). However, a signal from PCA does not discriminate between direct or indirect protein–protein interactions, since a third protein can act as a bridge, thus bringing the two proteins into sufficient proximity. However, the technique is extremely useful in protein interactome assembly studies. Consequently, other methods such as co-immunoprecipitation may need to be used to validate and give further information on the protein–protein interactions observed by PCA.

9 RNAi Knock-Down

RNA interference (RNAi) is a relatively recently described mechanism for knocking down gene expression and is found to be evolutionally conserved in most eukaryotic organisms to protect the host's genome. The phenomenon of RNAi was first observed in 1986 in transgenic plants where the transfected antisense RNA inhibited the transcription of the homologous mRNA (Ecker and Davis). Similar effects of downregulating both exogenous and endogenous RNA expression were observed in plants and fungus during the early 1990s (Napoli et al. 1990; Romano and Macino 1992). The first RNA inhibition phenomenon in animal cells was reported by Fire et al. (1998) in the nematode *C. elegans* and from this observation they coined the term "RNAi". They found that after introducing long double-stranded RNA (dsRNA) into *C. elegans* the homologous mRNA was selectively degraded.

Determining the molecular mechanism of RNAi has become one of the most exciting areas in biological science during the last decade. It has been established that the effector molecules that trigger mRNA degradation are small dsRNA with a length of 21–25 nucleotides, hence the term small interfering RNA (siRNA). siRNAs are derived from long dsRNAs by cleavage of a protein complex, called DICER, or from exogenous synthesis. In addition to the long dsRNA, RNA degradation can also be initiated by introducing synthetic siRNA into the cells. Both exogenous and endogenous siRNAs are recognized by the RNA-induced silencing complex (RISC). Once bound to RISC, the siRNA targets a specific mRNA, triggers its cleavage and thus prevents it from being translated. Several small RNA species exist naturally, including siRNAs, microRNAs (miRNAs), and repeat-associated siRNAs (rasiRNAs). The different small RNAs mature through specialized pathways and they also have different targets for degradation or repression (Doench et al. 2003; Lim et al. 2003). In different organisms, such as the nematode worm *C. elegans*, the fruit fly *Drosophila melanogaster*, and the flowering plant *Arabidopsis thaliana*, whilst the generic inhibitory mechanisms are similar, the RISC complex and the enzyme dicer are all distinct. Comprehensive reviews on the cellular mechanism of RNAi have been published recently (e.g. Dykxhoorn and Lieberman 2005), thus we will focus here on the application of RNAi in protein–protein interaction studies.

Genome-wide synthetic RNAi libraries have been used to study gene function both in several animal model organisms such as *C. elegans* (Kamath and Ahringer 2003; Kamath et al. 2003; Lettre et al. 2004) and fruit fly *Drosophila* (Boutros et al. 2004), and in plants such as the hexaploid wheat *Triticum aestivum* (Travella et al. 2006) and *Arabidopsis Thaliana* (McGinnis et al. 2005). Boutros et al. demonstrated that a comprehensive library, containing approximately 500 base-pair dsRNAs, covering more than 90% of the *Drosophila* mRNA could be used for high-throughput screening to identify genes involved in cell viability. This study resulted in the identification of 438 dsRNAs that identified essential genes, amongst which 80% lacked mutant alleles.

There are three effective approaches for introducing dsRNA into the worm *C. elegans*, injection, soaking and feeding, to assign gene functions. Several experiments were carried out to feed *C. elegans* with *E. coli* expressing target gene dsRNA to determine the loss-of-function phenotype of genes of interest (Kamath et al. 2001; Timmons et al. 2001). Kamath et al. exploited this method by generating a library of some 17 000 siRNA constructs, representing 86% of the *C. elegans* transcriptome, to identify about 1700 mutant phenotypes, two thirds of which were novel.

It was later found that long dsRNAs are of limited utility in mammalian cells due to global suppression of gene expression by dsRNA-induced activation of the interferon response (Stark et al., 1998). By contrast, siRNAs (21–25 nt) do not generally stimulate the interferon response, and hence siRNAs can be used for library screening in mammalian systems. In the past few years, novel techniques have been developed to construct siRNA libraries for high-throughput screening coupled to microarray/microwell platforms, handled by robots and automatic data analysis (review Vanhecke and Janitz 2005). An example of using siRNAs for cDNA library screening was recently reported by Zhao and Ding (2007). In their study, a synthetic siRNA library, targeting 5000 human genes to identify natural repressors of osteogenic specification, was screened. The study led to the identification of 53 candidate suppressors, 12 of which were further confirmed for their crucial roles in suppressing osteogenic specification in human mesenchymal stem cells.

Although the molecular mechanism of RNAi in mammalian cells is still not fully understood, the phenomenon has already been exploited in experimental biology to study gene functions in cells in vitro or in model organisms in vivo.

10 cDNA Library Screening

cDNA library screening is a general and broad strategy for genomics and proteomics, by exploiting a range of methods (as described above) to identify novel genes. Here we focus on methods developed in our laboratory for cDNA library screening with the aim to identify novel signaling proteins and to map intracellular signal processing pathways. Our method of cDNA library screening identifies genes based on their function within a particular signaling network.

Mammalian gene expression cloning has been reported to be a powerful tool for examining the interactions between intracellular molecules. About twenty years ago, IL1R was first cloned (Sims et al. 1988) using a modified technique, a single cell autoradiography after radioligand binding, based on methods established by Aruffo and Seed (1987). To adapt this approach for mapping signaling pathways, we used co-transfection of a transcription reporter with a cDNA expression library. We named our method "Transcription Expression Cloning". It exploits the notion that overexpression of most signaling components mimics the effect of extracellular agonists (Deng and Karin 1994; Muzio et al. 1997; Tojima et al. 2000), and thus the downstream response can be detected by a specific and sensitive reporter system.

We applied this method to identify novel components in TIR receptor-induced signaling pathways where we used the *IL8* promoter, which contains NF-κB, c/EBP and AP-1 sites, to drive the reporter enhanced green fluorescent protein (EGFP) or luciferase (Kiss-Toth et al. 2000; Kiss-Toth et al. 2006). Although the proof-of-concept studies showed that EGFP as a reporter can be used for mapping pro-inflammatory signaling pathways by confocal analysis luciferase was used as a reporter (due to more straightforward automation). In this system, three plasmids were co-transfected into the HeLa cells, including (1) the reporter firefly luciferase driven by an *IL8* promoter (IL8-Luc), (2) an internal non-inducible control reporter Renilla luciferase reporter driven by the HSV Thymidine kinase promoter (pTK-rLuc), and (3) a pool of cDNA expression library clones (oligo-dT primed, non-directionally cloned, from a human peripheral blood mononuclear cells) driven by *CMV* promoter.

Our experimental strategy for high-throughput screening (HTS) screening is outlined in Fig. 1. Pools of 48 DNAs were screened and positive pools broken down to identify the bioactive cDNAs (Kiss-Toth et al. 2006). Using this approach, we and others identified a number of novel components in a range of signaling pathways in genome wide screens (Chanda et al. 2003; Iourgenko et al. 2003; Kiss-Toth et al. 2006). For example, our hits included transcription factors, redox/NF-κB regulators, and modulators of Mitogen

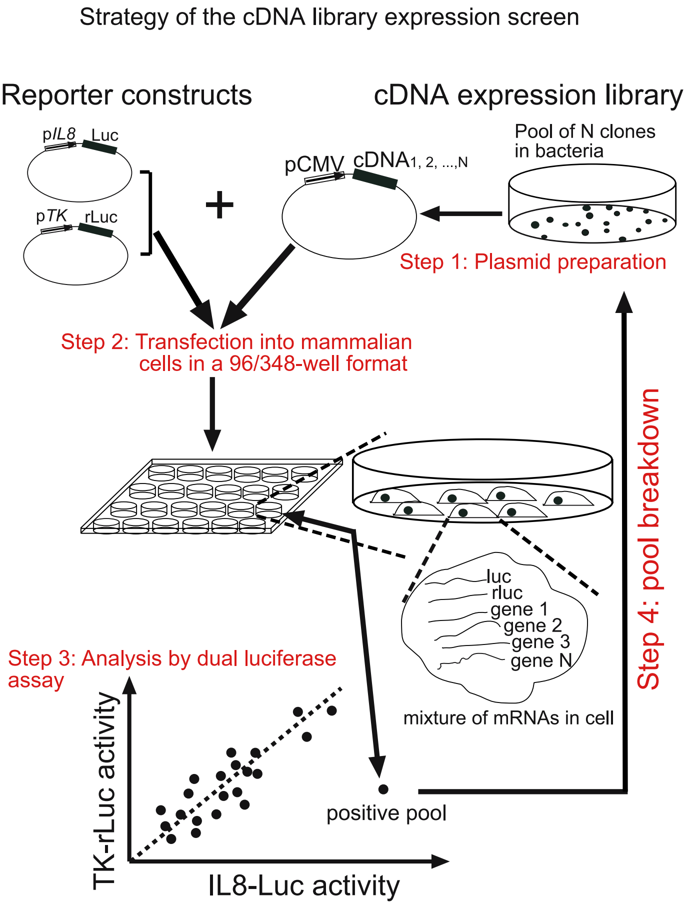


Fig. 1 Strategy of our previous cDNA library screen. A cDNA library (about 3 × 10^6 clones in size) was broken down to pools of *N*, where the pool size was determined by establishing the minimal amount of cDNA required for detection in *step 3*. The pools of plasmids were co-transfected with reporters (p*IL8*-Luc and p*TK*-rLuc) into mammalian cells (e.g. HeLa) on 96-well plates. After 24 h, the reporter activity was analyzed by dual luciferase assay (Promega) and positive pools were identified. The positive pools were then broken down into smaller subpools and eventually into single cDNA clones and *steps 1–3* repeated to identify the bioactive cDNA clone

Activated Protein Kinase networks. Therefore, the use of luciferase as a reporter, a high-throughput colony picker, a liquid handling robot, and a high-capacity luminometer allows screening of entire expression libraries (some 10^6 clones) in a timeframe feasible in regular research projects. These systems are able to catalogue genes based on their bioactivities in a short period of time and dramatically accelerate our speed of exploring protein interaction networks.

A limitation of our expression cloning method described above is that a strong viral *CMV* promoter was used to drive the components of a cDNA library, and an expression of the component far exceeding the natural concentration may result in artificial biological responses, generating false-positives. However, a fundamental feature of signaling systems is non-linear regulation, which is achieved by the generation of positive and negative feedback loops.

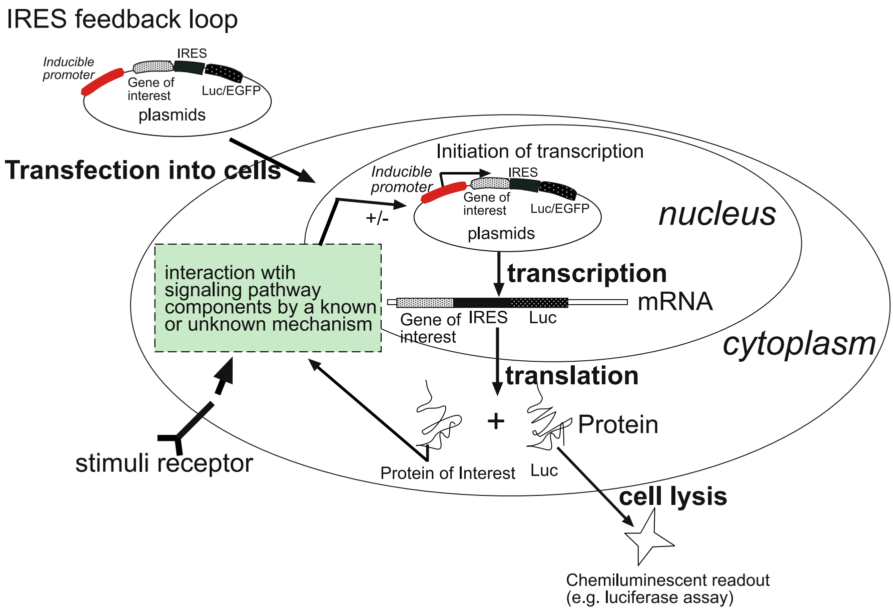


Fig. 2 IRES feedback loop. In this system, an IRES vector has been customized for screening of cDNA libraries. Expression of cDNA clones and reporter genes (luciferase or EGFP) were under the control of a promoter (*IL-8*), inducible by inflammatory signals. The gene of interest and the reporter were expressed on a single transcript in the cell, which ensures that both protein products are present in the same cell simultaneously. If the protein (product of the gene of interest in the IRES construct) is a regulator of a signaling network, whose end effector regulates the transcription initiation through the inducible promoter on the IRES constructs, a positive (or negative) feedback loop is formed, and thus the readout of the reporter (e.g. luciferase assay) reflects the non-linear perturbation of the signaling system of interest by introducing the IRES expression cassette into the cells

These play a key role in the modulation of physiological responses. Therefore, we have investigated an advanced version of the original method by incorporating a positive feedback loop in our system, to identify novel components in the TIR signaling pathway. We have used an IRES vector to clone the potential regulator (Y) to one of the expression cassettes and the reporter (EGFP/Luciferase) to the other, so that the potential regulator is transcribed in the same transcript as the reporter. The strategy is outlined in Fig. 2. We have verified this system by using well-characterized proteins such as RelA, TRAF6, MyD88 and IκBα, and shown that it can be used to select specific proteins in the TIR signaling pathway (Guan et al. 2006, 2007). The advantage of this improved method over previous systems developed by us and others is that both the potential regulator from a cDNA library and the reporter are transcribed on a single transcript and driven by an inducible promoter, instead of a strong viral promoter. Therefore, the expression of the test protein is self-regulated, nearer to the physiological concentration, thus generating less artificial responses.

11
Concluding Remarks

Understanding gene function at the molecular level requires characterization of protein–protein interactions. Formation of multiprotein complexes is often a dynamic process, where components of the complex need to be enzymatically modified for physiological interactions to take place or to be disrupted. In order to gain an in depth understanding of these events, it is often necessary to deploy a number of experimental approaches. In this review, we have summarized the most commonly used methods to assign gene function through detection of protein–protein interactions. Through the examples we listed above, it is evident that one of the main development paths for the development of technologies is towards automation and assay development in a format which enables global investigations of the proteome. Whilst these systems are undoubtedly promoting our understanding of molecular/cell biology to a higher level, many of the emerging technologies require significant investment in equipment, which makes research prohibitively expensive for many. Fortunately, this is often just a temporary limitation as instrument prices fall rapidly as a novel technique becomes more widely used. In addition, we believe that developing novel assays which do not necessarily require a major capital investment is of paramount importance. Examples of such approaches include modifications of the classical Y2H, the MAPPIT system or functional expression cloning. These enable access to cutting edge biology research for the wide scientific community, thus ensuring that knowledge is both shared and accumulated globally.

References

1. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294
2. Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. Nat Rev Drug Discov 3:301–317
3. Aronheim A, Engelberg D, Li N, al-Alawi N, Schlessinger J, Karin M (1994) Membrane targeting of the nucleotide exchange factor Sos is sufficient for activating the Ras signaling pathway. Cell 78:949–961
4. Aronheim A, Zandi E, Hennemann H, Elledge SJ, Karin M (1997) Isolation of an AP-1 repressor by a novel method for detecting protein–protein interactions. Mol Cell Biol 17:3094–3102
5. Aruffo A, Seed B (1987) Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. Proc Natl Acad Sci USA 84:8573–8577
6. Blackstock WP, Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol 17:121–127
7. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science 303:832–835
8. Chanda SK, White S, Orth AP, Reisdorph R, Miraglia L, Thomas RS, DeJesus P, Mason DE, Huang Q, Vega R et al (2003) Genome-scale functional profiling of the mammalian AP-1 signaling pathway. Proc Natl Acad Sci USA 100:12153–12158
9. Chen H, Puhl HL III, Koushik SV, Vogel SS, Ikeda SR (2006) Measurement of FRET efficiency and ratio of donor to acceptor concentration in living cells. Biophys J 91:L39–L41
10. De A, Gambhir SS (2005) Noninvasive imaging of protein–protein interactions from live cells and living subjects using bioluminescence resonance energy transfer. Faseb J 19:2017–2019
11. Deng T, Karin M (1994) c-Fos transcriptional activity stimulated by H-Ras-activated protein kinase distinct from JNK and ERK. Nature 371:171–175
12. Deshayes K, Schaffer ML, Skelton NJ, Nakamura GR, Kadkhodayan S, Sidhu SS (2002) Rapid identification of small binding motifs with high-throughput phage display: discovery of peptidic antagonists of IGF-1 function. Chem Biol 9:495–505
13. Doench JG, Petersen CP, Sharp PA (2003) siRNAs can function as miRNAs. Genes Dev 17:438–442
14. Dykxhoorn DM, Lieberman J (2005) The silent revolution: RNA interference as basic biology, research tool, and therapeutic. Annu Rev Med 56:401–423
15. Ecker JR, Davis RW (1986) Inhibition of gene expression in plant cells by expression of antisense RNA. Proc Natl Acad Sci USA 83:5372–5376
16. Fernandes PB (1998) Technological advances in high-throughput screening. Curr Opin Chem Biol 2:597–603
17. Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. Nature 340:245–246
18. Figeys D, McBroom LD, Moran MF (2001) Mass spectrometry for the study of protein–protein interactions. Methods 24:230–239
19. Finley RL Jr, Brent R (1994) Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. Proc Natl Acad Sci USA 91:12980–12984
20. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811

21. Fromont-Racine M, Rain JC, Legrain P (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet 16:277–282
22. Fujii I (2007) Directed evolution of antibody molecules in phage-displayed combinatorial libraries. Yakugaku Zasshi 127:91–99
23. Guan H, Holland K, Qwarnstrom E, Dower SK, Kiss-Toth E (2006) Feedback loops in intracellular signal processing and their potential for identifying novel signalling proteins. Cell Immunol 244:158–161
24. Guan H, Kiss-Toth E, Dower SK (2007) Analysis of Innate Immune Signal Transduction With Autocatalytic Expression Vectors. J Immunol Methods (in press), see also http://www.ncbi.nlm.nih.gov/pubmed/18155720?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum
25. Haab BB, Dunham MJ, Brown PO (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biol 2, RESEARCH0004
26. Hanson CL, Fucini P, Ilag LL, Nierhaus KH, Robinson CV (2003) Dissociation of intact Escherichia coli ribosomes in a mass spectrometer. Evidence for conformational change in a ribosome elongation factor G complex. J Biol Chem 278:1259–1267
27. Hu JC (2001) Model systems: Studying molecular recognition using bacterial *n*-hybrid systems. Trends Microbiol 9:219–222
28. Hudson JR Jr, Dawson EP, Rushing KL, Jackson CH, Lockshon D, Conover D, Lanciault C, Harris JR, Simmons SJ, Rothstein R, Fields S (1997) The complete set of predicted genes from Saccharomyces cerevisiae in a readily usable form. Genome Res 7:1169–1173
29. Ilag LL, Videler H, McKay AR, Sobott F, Fucini P, Nierhaus KH, Robinson CV (2005) Heptameric (L12)6/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria. Proc Natl Acad Sci USA 102:8192–8197
30. Iourgenko V, Zhang W, Mickanin C, Daly I, Jiang C, Hexham JM, Orth AP, Miraglia L, Meltzer J, Garza D et al (2003) Identification of a family of cAMP response element-binding protein coactivators by genome-scale functional analysis in mammalian cells. Proc Natl Acad Sci USA 100:12147–12152
31. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574
32. Kamath RS, Ahringer J (2003) Genome-wide RNAi screening in Caenorhabditis elegans. Methods 30:313–321
33. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M et al (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–237
34. Kamath RS, Martinez-Campos M, Zipperlen P, Fraser AG, Ahringer J (2001) Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in Caenorhabditis elegans. Genome Biol 2, RESEARCH0002
35. Kaur J, Sharma R (2006) Directed evolution: an approach to engineer enzymes. Crit Rev Biotechnol 26:165–199
36. Kiss-Toth E, Bagstaff SM, Sung HY, Jozsa V, Dempsey C, Caunt JC, Oxley KM, Wyllie DH, Polgar T, Harte M et al (2004a) Human tribbles, a protein family controlling mitogen-activated protein kinase cascades. J Biol Chem 279:42703–42708
37. Kiss-Toth E, Guesdon FM, Wyllie DH, Qwarnstrom EE, Dower SK (2000) A novel mammalian expression screen exploiting green fluorescent protein-based transcription detection in single cells. J Immunol Methods 239:125–135

38. Kiss-Toth E, Qwarnstrom EE, Dower SK (2004b) Hunting for genes by functional screens. Cytokine Growth Factor Rev 15:97–102
39. Kiss-Toth E, Wyllie DH, Holland K, Marsden L, Jozsa V, Oxley KM, Polgar T, Qwarnstrom EE, Dower SK (2006) Functional mapping and identification of novel regulators for the Toll/Interleukin-1 signalling network by transcription expression cloning. Cell Signal 18:202–214
40. Knudsen CR, Jadidi M, Friis I, Mansilla F (2002) Application of the yeast two-hybrid system in molecular gerontology. Biogerontology 3:243–256
41. Kritzer JA, Lear JD, Hodsdon ME, Schepartz A (2004) Helical beta-peptide inhibitors of the p53-hDM2 interaction. J Am Chem Soc 126:9468–9469
42. Lettre G, Kritikou EA, Jaeggi M, Calixto A, Fraser AG, Kamath RS, Ahringer J, Hengartner MO (2004) Genome-wide RNAi identifies p53-dependent and -independent regulators of germ cell apoptosis in C. elegans. Cell Death Differ 11:1198–1203
43. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP (2003) Vertebrate microRNA genes. Science 299:1540
44. Liu S, Liu S, Zhu X, Liang H, Cao A, Chang Z, Lai L (2007) Nonnatural protein–protein interaction-pair design by key residues grafting. Proc Natl Acad Sci USA 104:5330–5335
45. McGinnis K, Chandler V, Cone K, Kaeppler H, Kaeppler S, Kerschen A, Pikaard C, Richards E, Sidorenko L, Smith T et al (2005) Transgene-induced RNA interference as a tool for plant functional genomics. Methods Enzymol 392:1–24
46. Meng X, Wolfe SA (2006) Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nat Protoc 1:30–45
47. Mercurio F, Zhu H, Murray BW, Shevchenko A, Bennett BL, Li J, Young DB, Barbosa M, Mann M, Manning A, Rao A (1997) IKK-1 and IKK-2: cytokine-activated IkappaB kinases essential for NF-kappaB activation. Science 278:860–866
48. Michnick SW (2003) Protein fragment complementation strategies for biochemical network mapping. Curr Opin Biotechnol 14:610–617
49. Michnick SW (2004) Proteomics in living cells. Drug Discov Today 9:262–267
50. Muzio M, Ni J, Feng P, Dixit VM (1997) IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. Science 278:1612–1615
51. Napoli C, Lemieux C, Jorgensen R (1990) Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell 2:279–289
52. Nelson RW, Dogruel D, Williams P (1994) Mass determination of human immunoglobulin IgM using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 8:627–631
53. O'Farrell PZ, Goodman HM, O'Farrell PH (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. Cell 12:1133–1141
54. Osborne MA, Dalton S, Kochan JP (1995) The yeast tribrid system–genetic detection of trans-phosphorylated ITAM-SH2-interactions. Biotechnology (NY) 13:1474–1478
55. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics 6:308
56. Piston DW, Kremers GJ (2007) Fluorescent protein FRET: the good, the bad and the ugly. Trends Biochem Sci 32:407–414
57. Pollok BA, Heim R (1999) Using GFP in FRET-based applications. Trends Cell Biol 9:57–60
58. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V et al (2001) The protein–protein interaction map of Helicobacter pylori. Nature 409:211–215

59. Remy I, Michnick SW (2004a) A cDNA library functional screening strategy based on fluorescent protein complementation assays to identify novel components of signaling pathways. Methods 32:381–388
60. Remy I, Michnick SW (2004b) Mapping biochemical networks with protein-fragment complementation assays. Methods Mol Biol 261:411–426
61. Romano N, Macino G (1992) Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. Mol Microbiol 6:3343–3353
62. Rossi F, Charlton CA, Blau HM (1997) Monitoring protein–protein interactions in intact eukaryotic cells by beta-galactosidase complementation. Proc Natl Acad Sci USA 94:8405–8410
63. Rostom AA, Fucini P, Benjamin DR, Juenemann R, Nierhaus KH, Hartl FU, Dobson CM, Robinson CV (2000) Detection and selective dissociation of intact ribosomes in a mass spectrometer. Proc Natl Acad Sci USA 97:5185–5190
64. Sapsford KE, Liron Z, Shubin YS, Ligler FS (2001) Kinetics of antigen binding to arrays of antibodies in different sized spots. Anal Chem 73:5518–5524
65. Shih HM, Goldman PS, DeMaggio AJ, Hollenberg SM, Goodman RH, Hoekstra MF (1996) A positive genetic selection for disrupting protein–protein interactions: identification of CREB mutations that prevent association with the coactivator CBP. Proc Nat Acad Sci USA 93:13896–13901
66. Sims JE, March CJ, Cosman D, Widmer MB, MacDonald HR, McMahan CJ, Grubin CE, Wignall JM, Jackson JL, Call SM et al (1988) cDNA expression cloning of the IL-1 receptor, a member of the immunoglobulin superfamily. Science 241:585–589
67. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, Lawrence TS, Chinnaiyan AM (2001) Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. Cancer Res 61:7585–7593
68. Stark GR, Kerr IM, Williams BR, Silverman RH, Schreiber RD (1998) How cells respond to interferons. Annu Rev Biochem 67:227–264
69. Strauch EM, Georgiou G (2007) A bacterial two-hybrid system based on the twin-arginine transporter pathway of E. coli. Protein Sci 16:1001–1008
70. Sung HY, Guan H, Czibula A, King AR, Eder K, Heath E, Suvarna SK, Dower SK, Wilson AG, Francis SE et al (2007) Human tribbles-1 controls proliferation and chemotaxis of smooth muscle cells via MAPK signalling pathways. J Biol Chem 282(25):18379–18387
71. Tavernier J, Eyckerman I, Lemmers I, Van der Heyden J, Vandekerckhove J, Van Ostade X (2002) MAPPIT: a cytokine receptor-based two-hybrid method in mammalian cells. Clin Exp Allergy 32:1397–1404
72. Timmons L, Court DL, Fire A (2001) Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. Gene 263:103–112
73. Tirode F, Malaguti C, Romero F, Attar R, Camonis J, Egly JM (1997) A conditionally expressed third partner stabilizes or prevents the formation of a transcriptional activator in a three-hybrid system. J Biol Chem 272:22995–22999
74. Tojima Y, Fujimoto A, Delhase M, Chen Y, Hatakeyama S, Nakayama K, Kaneko Y, Nimura Y, Motoyama N, Ikeda K et al (2000) NAK is an IkappaB kinase-activating kinase. Nature 404:778–782
75. Tomashek JJ, Sonnenburg JL, Artimovich JM, Klionsky DJ (1996) Resolution of subunit interactions and cytoplasmic subcomplexes of the yeast vacuolar proton-translocating ATPase. J Biol Chem 271:10397–10404

76. Travella S, Klimm TE, Keller B (2006) RNA interference-based gene silencing as an efficient tool for functional genomics in hexaploid bread wheat. Plant Physiol 142:6–20
77. Tsien RY, Bacskai BJ, Adams SR (1993) FRET for studying intracellular signalling. Trends Cell Biol 7:242–245
78. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P et al (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 403:623–627
79. van Duijn E, Bakkes PJ, Heeren RM, van den Heuvel RH, van Heerikhuizen H, van der Vies SM, Heck AJ (2005) Monitoring macromolecular complexes involved in the chaperonin-assisted protein folding cycle by mass spectrometry. Nat Methods 2:371–376
80. Vanhecke D, Janitz M (2005) Functional genomics using high-throughput RNA interference. Drug Discov Today 10:205–212
81. Vidal M, Brachmann RK, Fattaey A, Harlow E, Boeke JD (1996) Reverse two-hybrid and one-hybrid systems to detect dissociation of protein–protein and DNA-protein interactions. Proc Natl Acad Sci USA 93:10315–10320
82. Walensky LD, Kung AL, Escher I, Malia TJ, Barbuto S, Wright RD, Wagner G, Verdine GL, Korsmeyer SJ (2004) Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. Science 305:1466–1470
83. Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I (1995) Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis 16:1090–1094
84. Woycechowsky KJ, Vamvaca K, Hilvert D (2007) Novel enzymes through design and evolution. Adv Enzymol Relat Areas Mol Biol 75:241–294, xiii
85. Zhao Y, Ding S (2007) A high-throughput siRNA library screen identifies osteogenic suppressors in human mesenchymal stem cells. Proc Natl Acad Sci USA 104:9673–9678
86. Zhu H, Bilgin M, Snyder M (2003) Proteomics. Annu Rev Biochem 72:783–812

Adv Biochem Engin/Biotechnol (2008) 110: 25–45
DOI 10.1007/10_2007_093
© Springer-Verlag Berlin Heidelberg
Published online: 25 April 2008

Cardiac Systems Biology and Parameter Sensitivity Analysis: Intracellular Ca^{2+} Regulatory Mechanisms in Mouse Ventricular Myocytes

Sung-Young Shin[1] · Sang-Mok Choo[2] · Sun-Hee Woo[3] ·
Kwang-Hyun Cho[1] (✉)

[1] Department of Bio and Brain Engineering and KI for the BioCentury,
Korea Advanced Institute of Science and Technology (KAIST), 305-701 Daejeon, Korea
ckh@kaist.ac.kr, http://sbie.kaist.ac.kr

[2] School of Electrical Engineering, University of Ulsan, 680-749 Ulsan, Korea

[3] College of Pharmacy, Chungnam National University, 305-764 Daejeon, Korea

Abstract Intracellular Ca^{2+} dynamics of cardiac myocytes are regulated by complex mechanisms of a variety of ion channels, transporters, and exchangers. Alterations of these Ca^{2+} regulatory components might lead to development of cardiac diseases. To investigate the regulatory mechanisms and hidden Ca^{2+} dynamics we use integrative systems analysis. Herein, we briefly summarize cardiac systems biology and, within the context of cardiac systems biology, identify the functional role of key Ca^{2+} regulatory proteins and their influence on intracellular Ca^{2+} dynamics (i.e., Ca^{2+} transient, SR Ca^{2+} content, CICR gain, half-decay time) using parameter sensitivity analysis based on an experimentally validated mathematical model of mouse ventricular myocytes. In addition, we analyze the influence of the pacing period (frequency) of a stimulus current since most of the Ca^{2+} regulatory proteins react with different timescales. Throughout the parameter sensitivity analysis, we found that alteration of SERCA or LTCC has a more significant effect on the Ca^{2+} dynamics than that of RyR or NCX. In particular, for the

70% down-regulation of LTCC, the Ca^{2+} influx through LTCC failed to initialize the SR Ca^{2+} release and thereby the intracellular Ca^{2+} dynamics was dramatically changed. We also found that the pacing period has a significant effect on the half-decay time of the Ca^{2+} transients. These findings provide us with new insights into the pathophysiology of cardiac failure as well as the development of new therapeutic strategies.

Keywords Ca^{2+} regulatory mechanism · Computer simulations · Functional analysis · Intracellular Ca^{2+} dynamics · Mathematical modeling · Mouse ventricular myocytes

Abbreviations

CaMKII	Calcium/calmodulin-dependent protein kinase II
CICR	Calcium-induced calcium release
LTCC	L-type Ca^{2+} channel
NCX	Na^{+}/Ca^{2+} exchanger
PKA	Protein kinase A
PLB	Phospholamban
PMCA	Plasma membrane calcium pump
RyR	Ryanodine receptor
SERCA	Sarco(endo)plasmic reticulum Ca^{2+}-ATPase
SR	Sarcoplasmic reticulum

1 Cardiac Systems Biology

1.1 Cardiac Excitation-Contraction Coupling

Cardiac excitation-contraction coupling is the process of pumping blood into a body, in which a heart contracts through electrical excitation of myocytes [1]. The ubiquitous second messenger Ca^{2+} plays a central role in cardiac electrical activity and direct activation of the myofilaments, which causes cardiac muscle contraction [2]. During depolarization of the cardiac action potential, Ca^{2+} enters a cell through voltage-gated Ca^{2+} channels. The Ca^{2+} then activates the ryanodine receptor (RyR) and triggers Ca^{2+} release from the sarcoplasmic reticulum (SR), which increases the Ca^{2+} concentration in the dyadic space between the L-type Ca^{2+} channel (LTCC) and RyR. The increased Ca^{2+} diffuses into the cytosol and raises the free intracellular (or cytosolic) Ca^{2+} concentration, allowing Ca^{2+} to bind to the myofilament protein troponin C, which then switches on the contractile machinery [2]. Although Ca^{2+} is the switch that activates myofilaments (the end effectors of excitation-contraction coupling), cardiac muscle contraction is graded and the contraction force depends on the cytosolic Ca^{2+} in a highly nonlinear way [1–3]. The strength of cardiac contractility is primarily regulated in two ways: the amplitude or duration of the Ca^{2+} transient and the sensitivity of the myofilaments to Ca^{2+} [2]. The myofilament Ca^{2+} sensitivity is enhanced dynamically by stretching the myofilaments (when the heart is filled with blood), resulting in a stronger contraction. However, this Ca^{2+} sensitivity is reduced by acidosis, elevated phosphate, and Mg^{2+} concentration, especially during ischemia [2]. β-adrenergic activation also reduces the myofilament Ca^{2+} sensitivity [4].

For relaxation of cardiac muscle, the cytosolic Ca^{2+} must decline, allowing Ca^{2+} to dissociate from troponin. There are four major Ca^{2+} transport systems: sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA), sarcolemmal Na^{+}/Ca^{2+} exchanger (NCX), sarcolemmal Ca^{2+}-ATPase or plasma membrane calcium pump (PMCA), and mitochondrial Ca^{2+} uniport [2]. In particular, SERCA and NCX play crucial roles in removing cytosolic Ca^{2+} although the quantitative importance varies between species [5]. In rabbit ventricular myocytes, SERCA removes 70% of the cytosolic Ca^{2+} and NCX removes 28% while PMCA and mitochondrial Ca^{2+} uniporter are responsible for only about 1%, respectively [6]. In the rat ventricle, the activity of SERCA is higher than that in the rabbit since the concentration of pump molecules is larger [7]. Quantitatively, 92% of the cytosolic Ca^{2+} is removed by SERCA, 7% by NCX, and 1% by PMCA and mitochondria [6]. However, during heart failure in many animal models including humans and rabbits, the functional expression and activity of these Ca^{2+} regulatory mechanisms is significantly altered [8].

1.2
Ca^{2+} Regulatory Mechanisms

The intracellular Ca^{2+} dynamics is regulated by a variety of Ca^{2+} regulatory mechanisms. In particular, LTCC, RyR, SERCA and NCX are known to play a central role [9, 10] and we call these the key Ca^{2+} regulatory mechanisms. In this subsection, we summarize the roles of these key Ca^{2+} regulatory proteins as well as their alterations in a failing heart.

Myocytes exhibit two types of voltage-dependent Ca^{2+} channels (L- and T-type) and the large electrochemical Ca^{2+} gradient drives extracellular Ca^{2+} into cytosol through these channels. As T-type Ca^{2+} channels are very few in most ventricular myocytes, we mainly consider L-type Ca^{2+} channels. LTCC (also known as dihydropyridine receptor: DHPR) are located primarily at sarcolemmal-SR junctions where the SR Ca^{2+} release channels exit. LTCC is activated by depolarization of the cell membrane, but Ca^{2+}-dependent inactivation at the cytosolic side limits the amount of Ca^{2+} entry during the action potential. This Ca^{2+}-dependent inactivation is a local effect and is mediated by calmodulin bound to the carboxyl terminus of the Ca^{2+} channel [11]. In addition, LTCC has been recently reported as interacting with calcium-binding protein-1 (CaBP1), calcium/calmodulin-dependent protein kinase II (CaMKII), A-kinase anchoring proteins (AKAPs), phosphatases, Caveolin-3, β-adrenergic receptor, PDZ domain proteins, sorcin, SNARE proteins, synaptotagmin, CSN5, the RGK family, and AHNAK1 [12]. The SR Ca^{2+} release during excitation-contraction coupling contributes to Ca^{2+}-dependent inactivation of LTCC. The total Ca^{2+} influx through LTCC is reduced by about 50% where the SR Ca^{2+} release occurs [5]. Thus, the SR Ca^{2+} release and LTCC create local negative feedbacks on Ca^{2+} influx. The higher Ca^{2+} influx and SR Ca^{2+} release turn off any further influx of Ca^{2+} through LTCC.

In a failing human heart with dilated and ischemic cardiomyopathy, the mRNA levels encoding the dihydropyridine receptor and dihyropyridine binding sites of LTCC were reported as being significantly decreased by 35–48% [13]. However, there are controversies concerning the expression levels of the dihydropyridine binding sites [14]. The Ca^{2+} currents in myocytes from a nonfailing heart were augmented by increasing stimulation frequencies whereas they were attenuated or lost in myocytes from the heart with reduced left ventricular function [15]. Several studies showed that alterations in the density of LTCC could deeply depend on animal age or diseases status [8].

Two different SR Ca^{2+} release channels are located on SR: RyR and inositol 1,4,5-triphosphate receptor (IP3R). However, in the cardiac SR, the density of RyR is significantly higher than that of IP3R, and the former is much more relevant for excitation-contraction coupling [16]. RyR located in the immediate vicinity of LTCC is activated by a local Ca^{2+} increase subsequent to Ca^{2+} influx through LTCC. RyR activated by small Ca^{2+} influx releases a large amount of Ca^{2+} from SR, which is termed the calcium-induced calcium release (CICR) process [17]. RyR forms a tetrameric structure composed of four monomers [18] that is stabilized by a channel-associated protein known as the FK506 binding protein (FKBP).

In the cardiomyopathic hamster, $[^3H]$ryanodine binding to cardiac membrane fractions was increased [19]. In contrast, $[^3H]$ryanodine binding and mRNA levels were reported as being decreased [20]. The decreased activity of RyR has been observed in two types of dog with a rapid ventricular pacing failure [21, 22]. In hypertensive and failing rats, the Ca^{2+} current density and the function of RyR were normal [23]; however, the relationship between the Ca^{2+} current density and the probability of evoking a spark was weakened.

SERCA pumps the cytosolic Ca^{2+} into SR, which transports two Ca^{2+} per one molecule of high-energy phosphate against a high ion gradient between the cytosolic (0.1–1 μM) and the SR Ca^{2+} (about 1 mM) [24]. SERCA is encoded by three genes and five different isoforms are expressed. Among these, SERCA2A is predominantly expressed in the cardiac and slow-twitch skeletal muscle cells [25]. SERCA is regulated by dephosphorylated PLB through direct protein–protein interaction. The binding of PLB to SERCA decreases the affinity of the calcium pump for Ca^{2+}. The phosphorylation of PLB by CaMKII and protein kinase A (PKA) results in the stimulation of SERCA through an increased affinity of SERCA for Ca^{2+} and an increased velocity (Vmax) of Ca^{2+} uptake [26, 27].

In Syrian hamsters with hereditary cardiomyopathy, gene expression levels of SERCA were decreased [28]. Moreover, SERCA mRNA levels and protein levels were decreased in a rat model of myocardial infarction induced by occluding the left coronary artery for 4–16 weeks [29]. The decrease of SERCA protein levels was also reported in failing guinea pig hearts following 8 weeks of banding of the descending thoracic aorta compared to an age-matched banded group without clinical signs of heart failure [30].

NCX is the dominant myocardial calcium efflux mechanism for muscle relaxation [31], extruding one Ca^{2+} for three Na^{+} through an electrochemical sodium gradient ("forward mode"). In this forward mode, a net movement of charge is produced and this results in a net inward current. NCX can also bring Ca^{2+} into a cell depending on the voltage ("reverse mode"). Some experimental results showed that NCX at high intracellular Na^{+} levels promote the calcium influx such that it induces excitation-contraction coupling [32, 33]. NCX is encoded by at least three different genes and a number of splice variants have been identified [34].

In rat cardiac hypertrophy, the decreased activity of NCX was reported [35]. In contrast, the enhanced activity of NCX was also observed in the cardiomyopathic Syrian hamster [36]. Recent studies indicate that NCX protein levels are significantly increased in both surface sarcolemma and T-tubular sarcolemma-enriched fractions in tachycardia-induced heart failure [37]. See [38] for further details on the regulation of NCX in normal and failing hearts.

1.3 Mathematical Models for Cardiac Systems Biology

It is being recognized that mathematical modeling of cardiac myocytes is a useful tool for investigating the complex biological process of electrical excitation and contraction [39, 40]. The mathematical description of cardiac myocytes can be traced back to the pioneering work of Hodgin and Huxly (called the "HH model") who first described the ionic currents of the squid giant axon quantitatively [41]. Ten years later, Noble developed a mathematical model by modifying the previous HH model to describe the long-lasting action and pace-maker potential of the Purkinje fibers of a heart [42]. In this model, he took account of three ion channels: Na^{+}, K^{+}, and Cl^{-}. In 1975, McAllister et al. published a cardiac action potential model of Purkinje fiber composed of nine ionic channels [43]. This model described for the first time the role of Ca^{2+} during the generation of action potential. The model consists of a rapid inward Na^{+} current (INa), a secondary inward current (ICa), a transient chloride current (ICl), a time-independent K^{+} current (IK1), a transient K^{+} current (IK2), and the fast (IX1) and slow (IX2) components of a new current. Di Francesco et al. published a mathematical model of the cardiac electrical activity by integrating ionic pumps, exchanger mechanisms, and concentration changes besides the ionic currents [44]. In particular, SR was represented for the first time by two compartments, including one for Ca^{2+} uptake and the other for a Ca^{2+} release store. In the 1990s, Rudy et al. published an expanded model of the guinea-pig ventricular myocyte by introducing the dependence of K^+ currents on K^+ concentration, the negative-slop characteristics of the time-independent K^+ current, novel potassium channels activated at plateau potentials and the modification of the fast Na^+ current [45, 46].

In the mid-1990s, the emphasis moved from general models of integrating voltage-clamp data from several species to more sophisticated ones based on data obtained from isolated cells of particular species since electrophysiological studies showed that action potential waveforms and ionic currents differ depending on species. For instance, the action potential of mouse and rat have no phase 2 plateau, but it exhibits rapid repolarization and a very short action potential duration compared with human, rabbit, guinea-pig and dog [47]. Mouse has been considered as a powerful tool to study the physiological effects of gene mutations, knockouts and transgenesis, and thereby the relevant mathematical model has also become increasingly important to understand the effects of these genetic manipulations, and to enable inferences about the effects expected in other species. The mathematical models of mouse and rat have been developed by Demir et al. [48] and Pandit et al. [49]. On the other hand, the mathematical models of the rabbit sinoatrial node were proposed by Zhang et al. [50] and Oehmen et al. [51]. A ventricular model of the rabbit was published by Puglish et al. [52]. Canine models were developed by Winslow et al. [53, 54], Ramirez et al. [55], and Cabo et al. [56].

The muscle contraction where force is developed by the attachment of a cross-bridge to the thin filament is a quite complex process which is controlled by the binding of free Ca^{2+} to troponin (TnC). There are several mathematical models (e.g., Kyoto model [57, 58]) that were developed to describe such muscle contraction and force generation processes. However, these models do not account for the entire force generation process in detail.

On the other hand, the β-adrenergic signaling pathway plays an important role in the regulation of cardiac myocytes and the development of heart failure [59]. This signaling pathway in response to sympathetic nerve activation or catecholamine activates the GTP-binding G-protein which subsequently activates adenylyl cyclase by converting ATP to cyclic AMP. The activated PKA by cAMP phosphorylates numerous target substrates including LTCC, PLB, RyR, SERCA, PMCA, several types of potassium channels, sodium channels, Tropoini I, etc., which alter ion currents and fluxes, and consequently intracellular Ca^{2+} dynamics and action potential. Saurcerman et al. developed a mathematical model by incorporating the signaling pathway into the excitation-contraction mechanism. Recently, more sophisticated mathematical models were developed by combining molecular-level processes and their regulations within the context of whole-cell functioning [60, 61]. These models have been used to investigate physiological and pathological phenomena including action potential adaptation to changes in heart rates and genetic mutations that are associated with a specific cardiac disease such as arrhythmias and heart failure [40].

1.4 Integrative Systems Analysis

In the foregoing sections, we have briefly summarized excitation-contraction coupling and its regulatory mechanism as well as mathematical models for cardiac systems biology. We find that the intracellular Ca^{2+} dynamics are regulated by complex mechanisms of a variety of ion channels, transporters, and exchangers. Hence, to investigate the regulatory mechanisms and hidden properties of intracellular Ca^{2+} dynamics of the cardiac myocytes, a "system-level integrative analysis" is required. For this purpose, experimentally validated mathematical models are now available as summarized in Sect. 1.3. We note that parameter sensitivity analysis of such mathematical models has been used as a powerful tool in exploring system dynamics and regulatory mechanisms (see Sect. 2.2 for a brief summary). Herein, we identify the functional role of key Ca^{2+} regulatory proteins and their influences on the intracellular Ca^{2+} dynamics (i.e., Ca^{2+} transient, SR Ca^{2+} content, CICR gain, and half-decay time) using parameter sensitivity analysis based on a mathematical model of mouse ventricular myocytes. In addition, we analyze the influence of the pacing period (frequency) of a stimulus current since most of the Ca^{2+} regulatory proteins react with different timescales. For instance, RyR and LTCC affect the Ca^{2+} dynamics in a relatively short time scale while SERCA and NCX have a more prolonged effect. Throughout the parameter sensitivity analysis, we found that SERCA and LTCC have more significant effects on the Ca^{2+} dynamics than either RyR or NCX. In particular, for the 70% down-regulation of LTCC, the Ca^{2+} influx through LTCC failed to initialize the SR Ca^{2+} release through RyR whereas the systolic Ca^{2+} level and the SR Ca^{2+} content were significantly changed. We also found that the pacing period has a significant effect on the half-decay time of cytosolic Ca^{2+} transient to parameter perturbations. All of the functional regulations were validated through multiple simulations with 10% random variation of parameters.

2 Quantification for Functional Analysis

2.1 Mathematical Modeling

We employ the mathematical model proposed by Bondarenko et al. for the functional analysis of the key Ca^{2+} regulatory proteins through the parameter sensitivity analysis [62]. This mathematical model consists of four compartments: junctional SR, network SR (NSR), dyadic (subspace) space, and cytosol. The action potential of the cell membrane and the intracellular Ca^{2+} including SR Ca^{2+} are assumed to be regulated by individual ion channels, and transporters that are usually found in mouse ventricular myocytes. In particular, the intracellular Ca^{2+} dynamics are assumed to be dominantly regulated by the key components: RyR, LTCC (regulating Ca^{2+} in dyadic space), SERCA, and NCX (regulating the cytosolic Ca^{2+}) (Fig. 1).

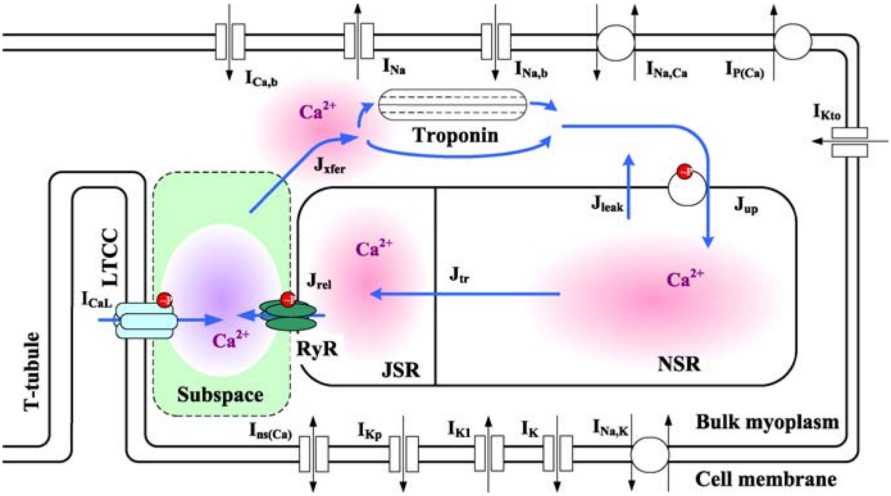


Fig. 1 Illustration of the intracellular Ca^{2+} regulatory mechanism. The regulatory system consists of four compartments: junctional SR, network SR, dyadic volume, and cytosol. The Ca^{2+} dynamics are regulated by the potassium channels, the sodium channels, the calcium channels, and transporters. The ion currents and the Ca^{2+} fluxes included in the mathematical model are as follows: I_{Ktof}, the transient outward K^+ current; I_{Cab}, the background Ca^{2+} current; I_{Na}, the fast Na^+ current; I_{Nab}, the background Na^+ current; I_{NaCa}, Na^+/Ca^{2+} the exchanger current; $I_{\mathrm{p(Ca)}}$, the Ca^{2+} pump current; I_{NaK}, the Na^+/K^+ pump current; I_{Ks}, the slowly delayed rectifier K^+ current; I_{K1}, the time independent K^+ current; I_{Kr}, the rapidly delayed rectifier K^+ current; $I_{\mathrm{na(Ca)}}$, the nonspecific Ca^{2+} current; I_{CaL}, the L-type Ca^{2+} current; J_{rel}, the Ca^{2+} flux released from SR; J_{tr}, the Ca^{2+} flux transferred from NSR to JSR; J_{leak}, the Ca^{2+} leakage from SR; J_{up}, the Ca^{2+} uptake through SERCA; J_{xfer}, the Ca^{2+} flux transferred from the dyadic space to the cytosol. Among these, we focus on the following key Ca^{2+} current and fluxes: I_{CaL}, J_{up}, J_{rel}, and J_{NaCa}

In the mathematical model, the Ca^{2+} flux released from SR through the RyR channel (J_{rel}) is described as a function of open probabilities of the gating variables (P_{O1} and P_{O2}) and the chemical gradient of Ca^{2+} between junctional SR and dyadic space Ca^{2+} as follows:

$$J_{\mathrm{rel}} = v_1(P_{O1} + P_{O2})\left(\left[\mathrm{Ca}^{2+}\right]_{\mathrm{JSR}} - \left[\mathrm{Ca}^{2+}\right]_{\mathrm{SS}}\right) P_{\mathrm{RyR}}\,,$$

where P_{RyR} denotes the RyR channel modulation factor, $[\mathrm{Ca}^{2+}]_{\mathrm{JSR}}$ and $[\mathrm{Ca}^{2+}]_{\mathrm{SS}}$ denote the Ca^{2+} concentration of junctional SR and dyadic space, respectively. The Ca^{2+} uptake (J_{up}) through SERCA is formulated based on a Hill-type equation as follows:

$$J_{\mathrm{up}} = \frac{v_3\left[\mathrm{Ca}^{2+}\right]_i^2}{K_{m,\mathrm{up}}^2 + \left[\mathrm{Ca}^{2+}\right]_i^2}\,,$$

where $[\mathrm{Ca}^{2+}]_i$ denotes the concentration of the cytosolic Ca^{2+}. The Ca^{2+} influx (I_{CaL}) through the L-type Ca^{2+} channel is described as a function of open probabilities of the gating variable (O), the cell membrane potential, and the reversal potential of the channel (E_{CaL}) as follows:

$$I_{\mathrm{CaL}} = G_{\mathrm{CaL}}O(V - E_{\mathrm{CaL}})\,,$$

where V denotes the membrane potential. NCX extrudes the cytosolic Ca^{2+} into the extra-cellular space at low negative voltages and allows for Ca^{2+} entry from the extra-cellular space to cytosol at relatively high voltages. So, the Ca^{2+} current (I_{NaCa}) through NCX is described as follows:

$$I_{\mathrm{NaCa}} = k_{\mathrm{NaCa}} \times \frac{\left[\exp(\eta VF/RT)\left[\mathrm{Na}^+\right]_i^3\left[\mathrm{Ca}^{2+}\right]_o - \exp((\eta-1)VF/RT)\left[\mathrm{Na}^+\right]_o^3\left[\mathrm{Ca}^{2+}\right]_i\right]}{(K_{m,\mathrm{Na}}^3 + \left[\mathrm{Na}^+\right]_i^3)(K_{m,\mathrm{Ca}} + \left[\mathrm{Ca}^{2+}\right]_o)(1 + k_{\mathrm{sat}}\exp((\eta-1)VF/RT))}\,,$$

where $[\mathrm{Na}^+]_i$ denotes the concentration of cytosolic Na^+. See [62] for further details on the mathematical model and parameters used in the simulation studies.

2.2 Sensitivity Analysis in Systems Biology

Parameter sensitivity analysis is the process of determining the sensitivity of responses to the change of parameter values [63]. It has been introduced as a powerful tool for systems biological approaches due to its practical applicability to model building and evaluation, understanding system dynamics, evaluating the confidence of a model under uncertainties, and experimental design [64–66]. For instance, Ihekwaba et al. applied parameter sensitivity analysis to the mathematical model of the NF-κB pathway and identified the parameters that most affect the oscillatory behavior of nuclear NF-κB [67]. Using this parameter sensitivity analysis, Mahdavi et al. revealed that overexpression of the receptor glycoprotein-130 results in reduced transcription-3 pathway activation and increased embryonic stem cell differentiation [68]. Hu et al. employed time-dependent parameter sensitivity analysis and found that phophatidylinositol 3′-kinase (PI3K) in the mitogene-activated protein kinase (MAPK) and PI3K-coupled pathway enhance the robustness of the MAPK pathway [69]. Cho et al. introduced a new strategy to parameter sensitivity analysis for experimental design and identification of key parameters in the TNFα-mediated NF-κB signaling pathway [70]. Multi-parameter sensitivity analysis was applied to investigate the key components and steps in the INF-γ-induced JAK-STAT signaling pathway [71]. By this approach, Zi et al. found that suppressor of cytokine signaling-1, nuclear phosphate, cytoplasmic STAT1, and the corresponding reaction steps are the most sensitive perturbation points in this pathway [71]. Herein, parameter sensitivity analysis is employed to investigate the functional roles and influences of the key Ca^{2+} regulatory proteins.

2.3 Quantification of Functional Regulatory Effects

To apply parameter sensitivity analysis, we varied each parameter from –70% to +70% of its nominal value as summarized in Table 1 and carried out computer simulations over the specified range of parameter values. The parameter perturbation range (–70% ~70%) was determined based on previous experimental results [8, 72–74] although the functional activity of some proteins might unusually vary out of this perturbation range [74]. We utilize percent change (PC) as an index to quantify the effect of functional regulation of a key Ca^{2+} regulatory protein (α) on the system response (x) as follows:

$$PC(p_\alpha) = \frac{x(p_\alpha \pm \Delta p_\alpha) - x(p_\alpha)}{x(p_\alpha)} ,$$

where p_α and Δp_α denote the nominal parameter values of α to be perturbed and the change of p_α, respectively. The system responses considered are the

Table 1 Parameter perturbations of the key Ca^{2+} regulatory proteins

Parameter	Description	Nominal value	Perturbation range
v_1	Maximum Ca^{2+} permeability	4.5 [ms^{-1}]	1.35 ~ 7.65
v_3	Maximum pump rate	0.45 [μM/ms]	0.135 ~ 0.765
G_{CaL}	Specific maximum conductivity	0.1729 [mS/μF]	0.05187 ~ 0.294
k_{NaCa}	Scaling factor	292.8 [pA/pF]	87.84 ~ 497.76

peak of the Ca^{2+} transient, the half-decay time, the mean of the SR Ca^{2+} content, and the CICR gain such that we can characterize the intracellular Ca^{2+} dynamics by these readouts. Note that the peak of the Ca^{2+} transient represents the maximum amplitude of the Ca^{2+} transient and the half-decay time represents a half of the time taken from the peak of the Ca^{2+} transient to its minimum amplitude. The mean of the SR Ca^{2+} content and the CICR gain are defined as follows:

$$\text{Mean of the SR Ca}^{2+}\text{ content} = \frac{1}{T}\int_{T} \left[\text{Ca}^{2+}\right]_{\text{TSR}}\,\mathrm{d}t$$

$$\text{CICR gain} = \frac{\max\limits_{0\le t\le T}\left(\dfrac{\mathrm{d}\left[\text{Ca}^{2+}\right]_i(t)}{\mathrm{d}t}\right)}{\max\limits_{0\le t\le T}\left(I_{\text{CaL}}(t)\right)}\,,$$

where $\left[\text{Ca}^{2+}\right]_{\text{TSR}} = \frac{\left[\text{Ca}^{2+}\right]_{\text{NSR}} V_{\text{NSR}} + \left[\text{Ca}^{2+}\right]_{\text{JSR}} V_{\text{JSR}}}{V_{\text{NSR}} + V_{\text{JSR}}}$, I_{CaL} denotes the L-type Ca^{2+} current, and T indicates the period of the electrical pulses. The mathematical model was coded in Matlab (V7.0, R14) and the full set of ordinary differential equations was solved by using the Runge–Kutta–Merson numerical integration algorithm on an HP workstation xw6000. To trigger the stimulated action potential, we employed a 0.5-ms 80 pA/pF depolarizing current with a pacing period of 1200 ms.

3 Parameter Sensitivity Analysis of Ca^{2+} Regulatory Mechanisms

3.1 Sarcoplasmic Reticulum Calcium Pump (SERCA)

The percent changes of the steady-state system response for parameter perturbations are illustrated in Fig. 2. The Ca^{2+} transient was almost linearly changed with respect to the parameter perturbation strength (Fig. 2A). The effects of the up- and down-regulation for the same perturbation strength were similar, but the effect on the CICR gain for the up-regulation was larger than that for the down-regulation (Fig. 2C) despite the similar effects on the Ca^{2+} transient. The half-decay time for the down-regulation of SERCA became nonlinearly prolonged with respect to the parameter perturbation strength (Fig. 2D). Although the primary reason for this is the decreased Ca^{2+} uptake rate, another important factor is the Ca^{2+} buffering. In other words, the decreasing Ca^{2+} transient amplitude increases the decay rate constant in a nonlinear way [10]. These findings are supported by several experimental

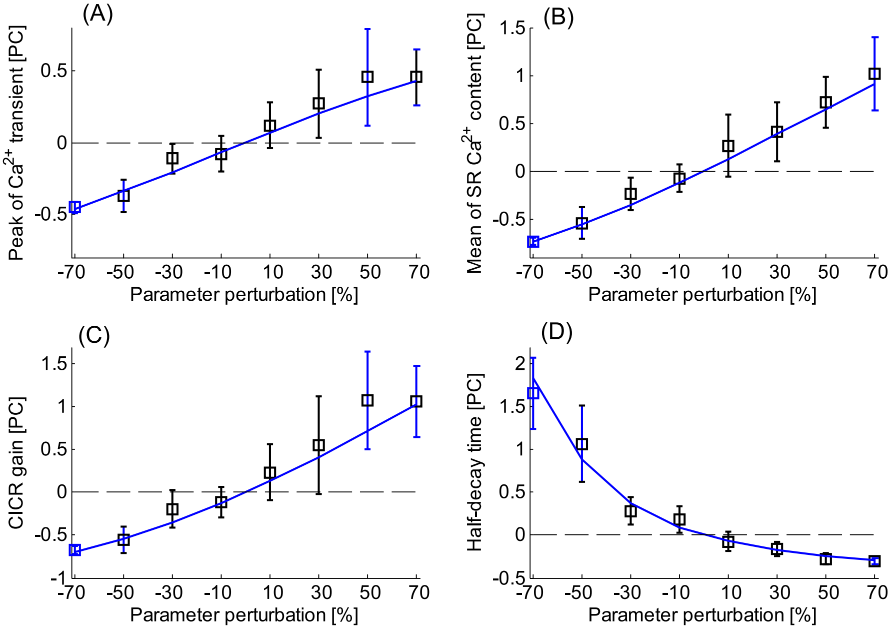


Fig. 2 The parameter perturbation effects of SERCA on system responses at steady-state beats. The *solid line* denotes the perturbation effect of SERCA when the other parameters are set to nominal values. The error bar indicates the perturbation effect of SERCA when the other parameters are subject to 10% random variations with respect to their nominal values. The *dashed line* denotes a reference level. **A** The percent change of the Ca^{2+} transient peak. **B** The percent change of the CICR gain. **C** The percent change of the half-decay time. **D** The percent change of the mean SR Ca^{2+} content. The system responses, except the half-decay time, almost linearly changed with respect to the parameter perturbation strength

results as follows: O'Neill et al. [75] showed that the SR Ca^{2+} content was significantly decreased (Fig. 2B); the half-decay time remarkably prolonged (Fig. 2D) during the application of SERCA inhibitor 2′,5′-di(tert-butyl)-1,4-benzohydroquinone (TBQ). In addition, the systolic Ca^{2+} transient was transiently (i.e., initially) increased during the same condition, which is in accord with the simulation results (data not shown) [75]. On the other hand, the increased activity of SERCA in PLB knockout mice increased the SR Ca^{2+} content and decreased the half-decay time [76].

The pacing period of the stimulus current did not significantly alter the percent change curves with respect to parameter perturbation except for the half-decay time (Fig. 3). As the pacing period shortened, the half-decay time became more significantly decreased especially for the down-regulation. At the same time, the minimum Ca^{2+} transient became considerably increased although the Ca^{2+} transient amplitude did not change (data not shown). We reason that the increased minimum Ca^{2+} transient for a short pacing period

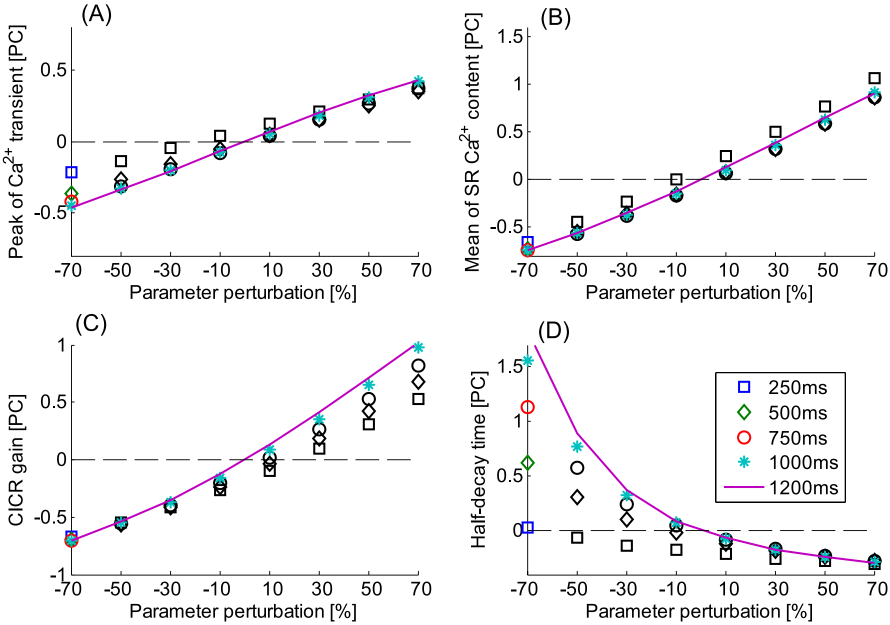


Fig. 3 The parameter perturbation effects of SERCA with respect to the variation of a pacing period. **A** The percent change of the Ca^{2+} transient peak. **B** The percent change of the CICR gain. **C** The percent change of the half-decay time. **D** The percent change of the mean SR Ca^{2+} content. The half-decay time decreased more significantly for a shorter pacing period

prolongs the half-decay time since the shortened diastolic duration diminishes the chance of Ca^{2+} removal by NCX and SERCA.

3.2 Sarcolemmal L-Type Calcium Channels (LTCC)

For the up-regulation of LTCC, the Ca^{2+} transient amplitude became almost linearly increased along with the parameter perturbation strength except for 70% down-regulation (Fig. 4A). The changes of the SR Ca^{2+} content and the CICR gain were much smaller compared with the effects of SERCA perturbations (Fig. 4B,C) but these changes were dramatic for 70% down-regulation of LTCC; the SR Ca^{2+} content was significantly increased and the CICR gain was decreased. The half-decay time was also significantly prolonged for 70% down-regulation, but, except for this perturbation, it became linearly decreased along with the parameter perturbation strength because of the enhanced activity of SERCA and NCX by the increased Ca^{2+} transient (Fig. 4D). Those dramatic nonlinear changes of the system responses for 70% down-regulation were presumably induced by the failure of the Ca^{2+} influx through

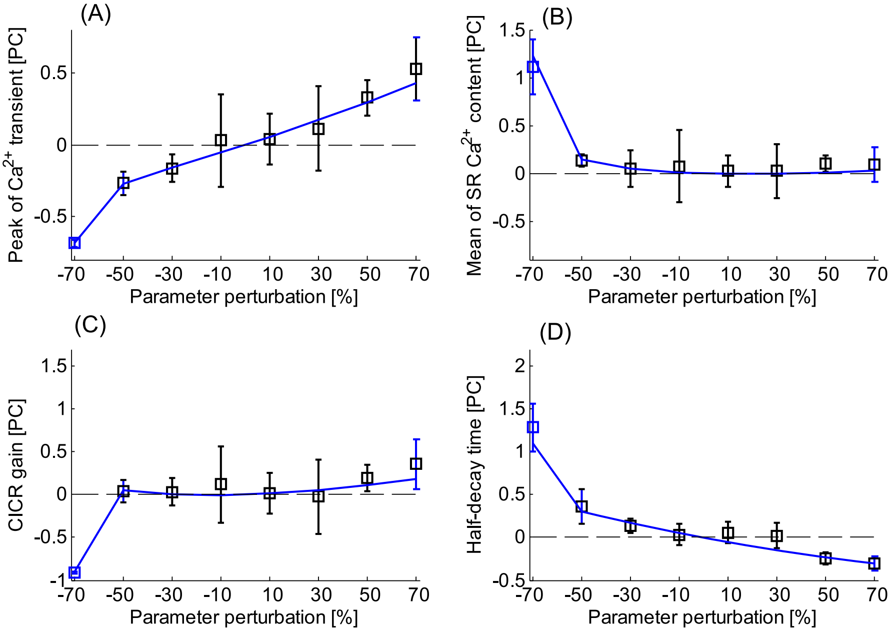


Fig. 4 The parameter perturbation effects of LTCC on system responses at steady-state beats. The functional regulations had more significant effects on the Ca^{2+} transient than the SR Ca^{2+} content

LTCC to initialize the SR Ca^{2+} release. Several previous experimental evidences support these simulation results on the functional regulation of LTCC. For instance, Trafford et al. [77, 78] showed that the increasing external Ca^{2+} concentration (which increases Ca^{2+} influx through LTCC) increased the systolic Ca^{2+} transient amplitude without any effect on the SR Ca^{2+} content, and the decreasing external Ca^{2+} concentration (which decreases Ca^{2+} influx through LTCC) decreased the systolic Ca^{2+} transient amplitude which produced a slight increase of the SR Ca^{2+} content. The reason why the Ca^{2+} transient is more sensitive to the parameter perturbations than that of the SR Ca^{2+} content can be explained by the following three facts (except for the case with 70% down-regulation at which the Ca^{2+} influx failed to trigger the SR Ca^{2+} release): First, the change of Ca^{2+} influx through the parameter perturbation has little effect on the SR Ca^{2+} content. Second, the total SR Ca^{2+} content (about 2.1×10^{-9} mol) is much larger on average than the cytosolic Ca^{2+} content (about 3.1×10^{-12} mol) [75]. Hence, the change of Ca^{2+} influx affects the cytosolic Ca^{2+} but not the SR Ca^{2+} content. Third, the larger Ca^{2+} transient due to the increased Ca^{2+} influx activates the larger Ca^{2+} efflux from the cytosol whereby the increased Ca^{2+} influx can be balanced [78]. From these simulation results, we showed for the fist time that 70% down-regulation of LTCC fails to initiate the SR Ca^{2+} release and this

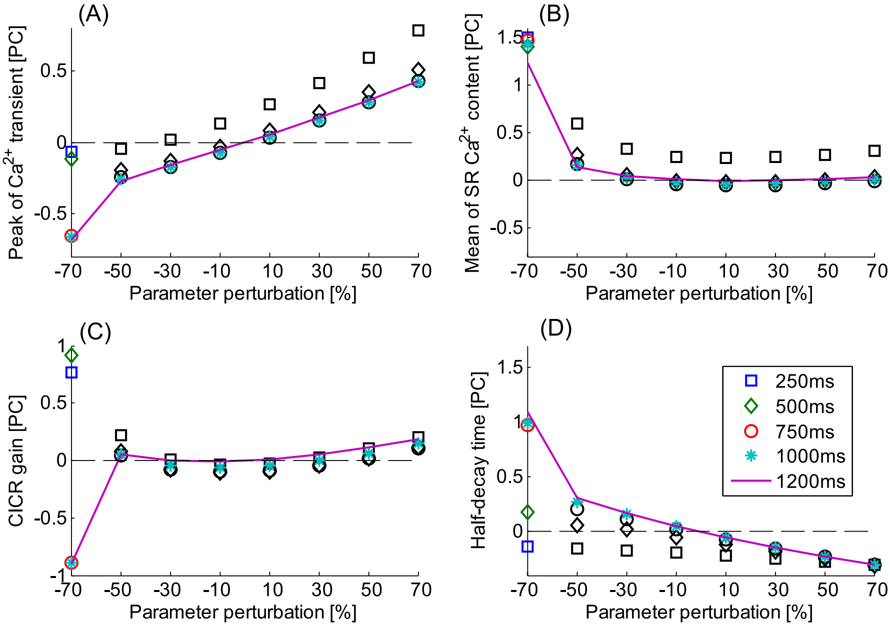


Fig. 5 The parameter perturbation effects of LTCC with respect to the variation of a pacing period. As the pacing period shortens, the Ca^{2+} transient and SR Ca^{2+} gain increase while the half-decay time decreases

leads to a dramatic change of intracellular Ca^{2+} dynamics. A specific perturbation limit inducing such an LTCC failure might depend on cell types, animal species and cell environmental conditions. So, further experimental studies are required to specify the perturbation limit value.

The shorter pacing period contributed to a greater extent to increasing the Ca^{2+} transient and the SR Ca^{2+} content (Fig. 5A,B). In particular, for 70% down-regulation of LTCC, the Ca^{2+} transient of the shortest pacing period was remarkably recovered (Fig. 5A) and thereby the CICR gain was significantly increased (Fig. 5C). The reason why the SR Ca^{2+} content increased to a greater extent for the shorter pacing period is presumably because the Ca^{2+} removal capacity of NCX was significantly restricted due to the shortened diastolic period.

3.3
Sarcoplasmic Reticulum Calcium Release Channels (RyR)

The Ca^{2+} transient became almost linearly increased along with the parameter perturbation strength (Fig. 6A) and the CICR gain showed a similar profile (Fig. 6C). However, the SR Ca^{2+} content change for the down-regulation of RyR was much larger than that of the up-regulation (Fig. 6B). These simu-

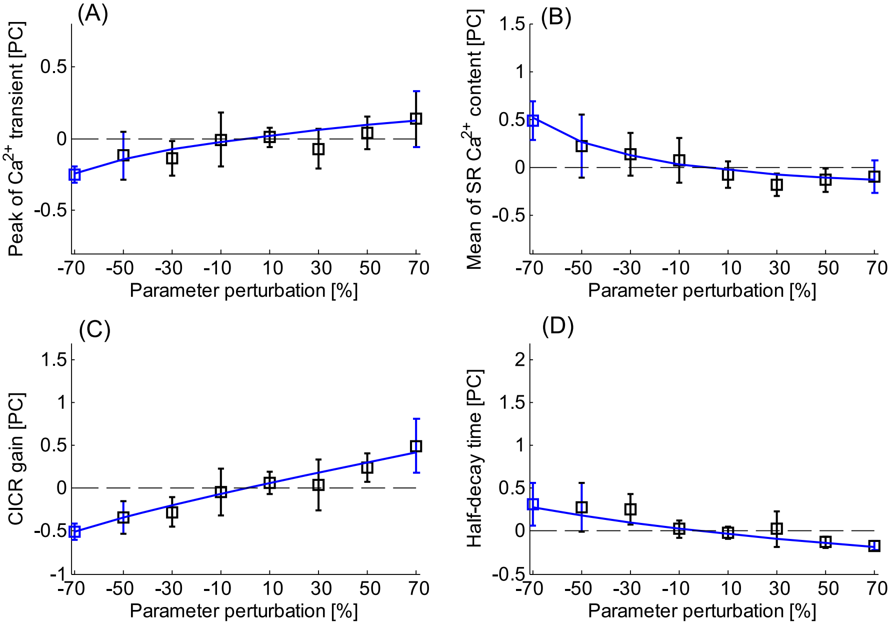


Fig. 6 The parameter perturbation effects of RyR on system responses at steady-state beats. The effects of down-regulations were relatively larger than those of up-regulations

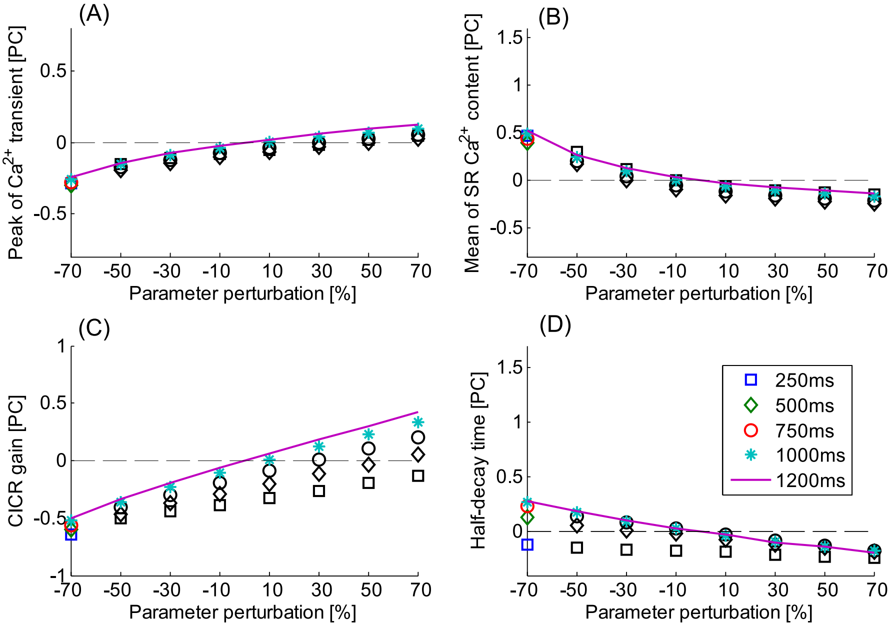


Fig. 7 The parameter perturbation effects of RyR with respect to the variation of a pacing period. The CICR gain and half-decay time decrease as the pacing period shortens

lation results are supported by previous experimental evidences. For instance, Eisner et al. [79] showed that the increase of the RyR activity by caffeine and butanedione monoxime (BDM) transiently increased the systolic Ca^{2+} transient, though it eventually decreased again to its initial level due to the decreased SR Ca^{2+} content. Diaz et al. [78] showed that local control by anesthetic tetracain which decreased the open probability of RyR could induce a large increase of SR Ca^{2+} content.

We found that the pacing period has little effect on the Ca^{2+} transient and SR Ca^{2+} content; however, the CICR gain and the half-decay time were considerably decreased for a shortened pacing period (Fig. 7).

3.4 Sarcolemmal Sodium–Calcium Exchanger (NCX)

The changes of system responses were linear with respect to the parameter perturbation strength (Fig. 8); however, the down-regulation effect was larger than that of up-regulation while the functional regulation of NCX had little effect on the half-decay time (Fig. 8D). The shorter pacing period increased the Ca^{2+} transient and SR Ca^{2+} content (Fig. 9A,B) because the Ca^{2+} removal capacity of NCX was constricted by the shortened pacing period. These results

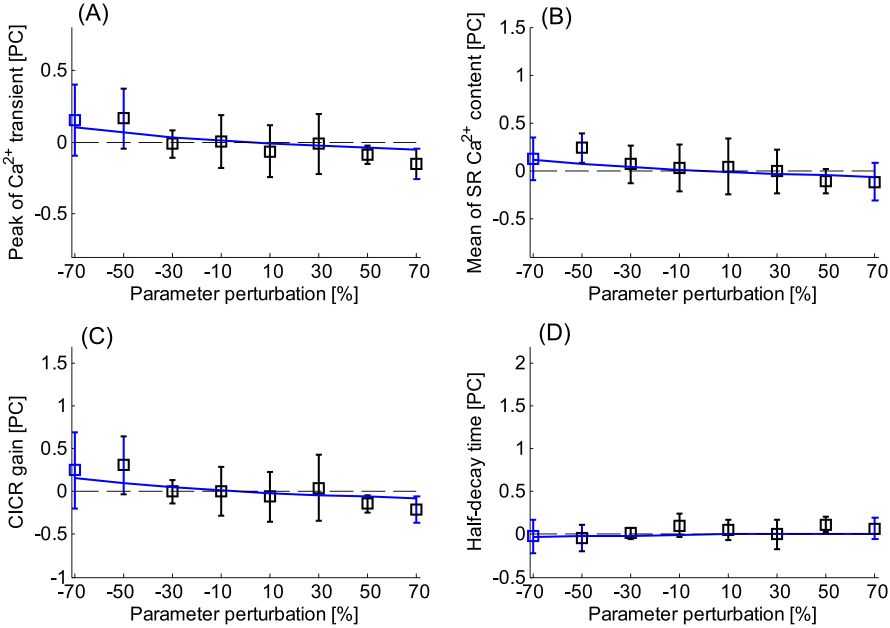


Fig. 8 The parameter perturbation effects of NCX on system responses at steady-state beats. The functional regulatory effect almost linearly changed along with the parameter perturbation strength

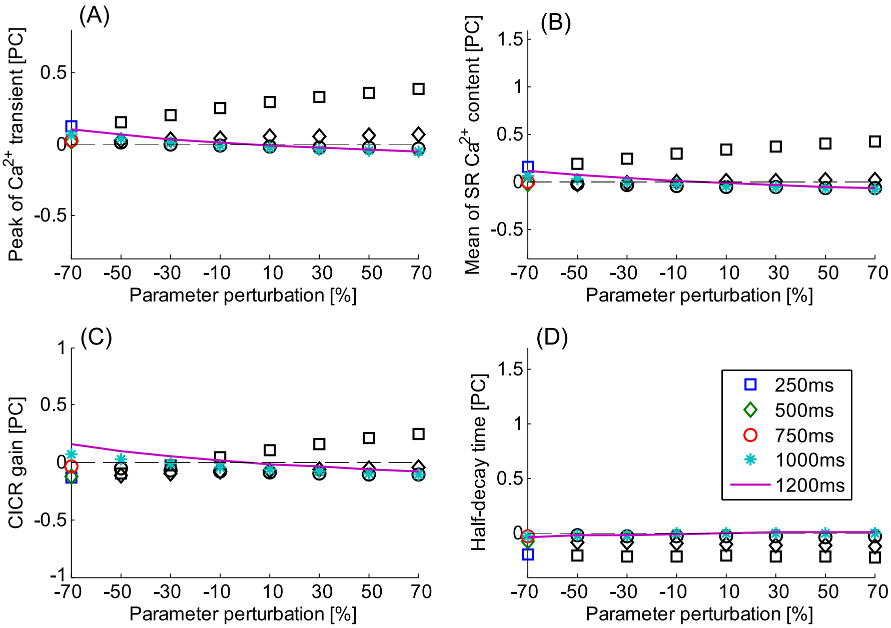


Fig. 9 The parameter perturbation effects of RyR with respect to the variation of a pacing period. The Ca^{2+} transient and SR Ca^{2+} content increase as the pacing period shortens

showed that the functional regulation of NCX had little effect on the system responses compared with effects of other Ca^{2+} regulatory proteins. This can be explained by the relatively low capacity of Ca^{2+} removal from the cytosol compared with SERCA. For instance, NCX is responsible for only 7% of the whole removal process in the rat [6]. This relatively low influence of NCX on the intracellular Ca^{2+} dynamics is also supported by other experimental evidences. For example, Goldhaber et al. [80] showed that a low level (or ablation) of NCX expression has a minimal effect on the Ca^{2+} transient and SR Ca^{2+} content in the genetically modified mice.

4 Conclusions

We have investigated the functional role and influences of key Ca^{2+} regulatory proteins in intracellular Ca^{2+} dynamics using parameter sensitivity analysis. From the simulation results, we found that SERCA and LTCC have the most significant effects on the system responses in various aspects, and RyR has a more significant effect than NCX. The decreased Ca^{2+} influx through the 70% down-regulation failed to initialize the SR Ca^{2+} release and thereby the intracellular Ca^{2+} dynamics were dramatically changed. From the pacing period analysis, we also found that the pacing period has a significant effect on the half-decay time of the Ca^{2+} transient depending on the strength of parameter perturbation. To take account of the robustness of parameters used for simulations, we repeated the simulations with 10% random variations of parameters and confirmed that the parameter perturbation effects on the Ca^{2+} dynamics are robust with respect to these variations. We also considered the effect of pacing periods since the Ca^{2+} regulatory components react over different timescales. By repeating the simulations over different pacing periods, we found that the pacing period was also an important factor affecting the system responses.

The parameter sensitivity analysis and the in silico simulation approach may be useful tools for probing time course cellular responses to interventions of Ca^{2+} regulatory proteins with different levels or durations with respect to a whole cell Ca^{2+} signaling. They would be particularly useful when there is no specific pharmacological modulator or when there are biological and technical limitations in experimental verification with intact cardiac myocytes. For example, to measure time-dependent changes in the SR Ca^{2+} content and CICR gain on a beat-to-beat basis in intact cells is experimentally very difficult. In addition, it is not always easy to estimate % change of a protein function at certain concentrations of modulating agents during an experiment.

Acknowledgements This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (M10503010001-07N030100112 and M10503010001-07N030100113) and also supported by the Korea Ministry of Science and Technology through the Nuclear Research Grant (M20708000001-07B0800-00110) and the 21C Frontier Microbial Genomics and Application Center Program (Grant MG05-0204-3-0).

References

1. Bers DM (2003) Excitation-Contraction Coupling and Cardiac Contractile Force. Kluwer Academic Publishers, Boston
2. Bers DM (2002) Nature 415:198
3. Solaro RJ, Rarick HM (1998) Circ Res 83:471
4. Nakae Y, Fujita S, Namiki A (2001) Anesth Analg 93:846
5. Puglisi JL, Yuan W, Bassani JW, Bers DM (1999) Circ Res 85:e7
6. Shannon TR, Bers DM (2004) Ann NY Acad Sci 1015:28
7. Hove-Madsen L, Bers DM (1993) Circ Res 73:820
8. Hasenfuss G (1998) Cardiovasc Res 37:279
9. Sjaastad I, Wasserstrom JA, Sejersted OM (2003) J Physiol 546:33
10. Eisner DA, Choi HS, Diaz ME, O'Neill SC, Trafford AW (2000) Circ Res 87:1087
11. Zuhlke RD, Pitt GS, Deisseroth K, Tsien RW, Reuter H (1999) Nature 399:159
12. Kobayashi T, Yamada Y, Fukao M, Tsutsuura M, Tohse N (2007) J Pharmacol Sci 103:347

13. Takahashi T, Allen PD, Lacro RV, Marks AR, Dennis AR, Schoen FJ, Grossman W, Marsh JD, Izumo S (1992) J Clin Invest 90:927
14. Rasmussen RP, Minobe W, Bristow MR (1990) Biochem Pharmacol 39:691
15. Piot C, Lemaire S, Albat B, Seguin J, Nargeot J, Richard S (1996) Circulation 93:120
16. Zalk R, Lehnart SE, Marks AR (2007) Annu Rev Biochem 76:367
17. Endo M (2007) Adv Exp Med Biol 592:275
18. Santonastasi M, Wehrens XH (2007) Acta Pharmacol Sin 28:937
19. Finkel MS, Shen L, Romeo RC, Oddis CV, Salama G (1992) J Cardiovasc Pharmacol 19:610
20. Lachnit WG, Phillips M, Gayman KJ, Pessah IN (1994) Am J Physiol 267:H1205
21. O'Brien PJ, Moe GW, Nowack LM, Grima EA, Armstrong PW (1994) Can J Physiol Pharmacol 72:999
22. Cory CR, McCutcheon LJ, O'Grady M, Pang AW, Geiger JD, O'Brien PJ (1993) Am J Physiol 264:H926
23. Gomez AM, Valdivia HH, Cheng H, Lederer MR, Santana LF, Cannell MB, McCune SA, Altschuld RA, Lederer WJ (1997) Science 276:800
24. Rossi AE, Dirksen RT (2006) Muscle Nerve 33:715
25. Periasamy M, Kalyanasundaram A (2007) Muscle Nerve 35:430
26. Kim M, Perrino BA (2007) Am J Physiol Gastrointest Liver Physiol 292:G1045
27. Rodriguez P, Mitton B, Nicolaou P, Chen G, Kranias EG (2007) Am J Physiol Heart Circ Physiol 293:H762
28. Kuo TH, Tsang W, Wang KK, Carlock L (1992) Biochim Biophys Acta 1138:343
29. Zarain-Herzberg A, Afzal N, Elimban V, Dhalla NS (1996) Mol Cell Biochem 163–164:285
30. Feldman AM, Weinberg EO, Ray PE, Lorell BH (1993) Circ Res 73:184
31. Bers DM, Ginsburg KS (2007) Ann NY Acad Sci 1099:326
32. Leblanc N, Hume JR (1990) Science 248:372
33. Levesque PC, Leblanc N, Hume JR (1994) Cardiovasc Res 28:370
34. Bers DM, Christensen DM, Nguyen TX (1988) J Mol Cell Cardiol 20:405
35. Hanf R, Drubaix I, Marotte F, Lelievre LG (1988) FEBS Lett 236:145
36. Hatem SN, Sham JS, Morad M (1994) Circ Res 74:253
37. Balijepalli RC, Lokuta AJ, Maertz NA, Buck JM, Haworth RA, Valdivia HH, Kamp TJ (2003) Cardiovasc Res 59:67
38. Reppel M, Fleischmann BK, Reuter H, Pillekamp F, Schunkert H, Hescheler J (2007) Ann NY Acad Sci 1099:361
39. Puglisi JL, Wang F, Bers DM (2004) Prog Biophys Mol Biol 85:163
40. Rudy Y, Silva JR (2006) Q Rev Biophys 39:57
41. Hodgkin AL, Huxley AF (1952) J Physiol 117:500
42. Noble D (1962) J Physiol 160:317
43. McAllister RE, Noble D, Tsien RW (1975) J Physiol 251:1
44. DiFrancesco D, Noble D (1985) Philos Trans R Soc Lond B Biol Sci 307:353
45. Luo CH, Rudy Y (1991) Circ Res 68:1501
46. Luo CH, Rudy Y (1994) Circ Res 74:1071
47. Yuan W, Ginsburg KS, Bers DM (1996) J Physiol 493(3):733
48. Demir SS, Clark JW, Murphey CR, Giles WR (1994) Am J Physiol 266:C832
49. Pandit SV, Clark RB, Giles WR, Demir SS (2001) Biophys J 81:3029
50. Zhang H, Holden AV, Kodama I, Honjo H, Lei M, Varghese T, Boyett MR (2000) Am J Physiol Heart Circ Physiol 279:H397
51. Oehmen CS, Giles WR, Demir SS (2002) J Cardiovasc Electrophysiol 13:1131
52. Puglisi JL, Bers DM (2001) Am J Physiol Cell Physiol 281:C2049

53. Winslow RL, Rice J, Jafri S, Marban E, O'Rourke B (1999) Circ Res 84:571
54. Greenstein JL, Wu R, Po S, Tomaselli GF, Winslow RL (2000) Circ Res 87:1026
55. Ramirez RJ, Nattel S, Courtemanche M (2000) Am J Physiol Heart Circ Physiol 279:H1767
56. Cabo C, Boyden PA (2003) Am J Physiol Heart Circ Physiol 284:H372
57. Matsuoka S, Sarai N, Kuratomi S, Ono K, Noma A (2003) Jpn J Physiol 53:105
58. Sarai N, Matsuoka S, Kuratomi S, Ono K, Noma A (2003) Jpn J Physiol 53:125
59. Sucharov CC (2007) Expert Rev Cardiovasc Ther 5:119
60. Cortassa S, Aon MA, O'Rourke B, Jacques R, Tseng HJ, Marban E, Winslow RL (2006) Biophys J 91:1564
61. Greenstein JL, Hinch R, Winslow RL (2006) Biophys J 90:77
62. Bondarenko VE, Szigeti GP, Bett GC, Kim SJ, Rasmusson RL (2004) Am J Physiol Heart Circ Physiol 287:H1378
63. Saltelli A, Ratto M, Tarantola S, Campolongo F (2005) Chem Rev 105:2811
64. Zhang Y, Rundell A (2006) IEE Proc Syst Biol 153:201
65. Schwacke JH, Voit EO (2005) J Theor Biol 236:21
66. Yue H, Brown M, Knowles J, Wang H, Broomhead DS, Kell DB (2006) Mol Biosyst 2:640
67. Ihekwaba AE, Broomhead DS, Grimley RL, Benson N, Kell DB (2004) Syst Biol (Stevenage) 1:93
68. Mahdavi A, Davey R, Bhola P, Yin T, Zandstra P (2007) PLoS Comput Biol 6:e130
69. Hu D, Yuan JM (2006) J Phys Chem A 110:5361
70. Cho K-H, Shin S-Y, Kolch W, Wolkenhaur O (2003) Simulation 79:726
71. Zi Z, Cho KH, Sung MH, Xia X, Zheng J, Sun Z (2005) FEBS Lett 579:1101
72. Balke CW, Shorofsky SR (1998) Cardiovasc Res 37:290
73. Siri FM, Krueger J, Nordin C, Ming Z, Aronson RS (1991) Am J Physiol 261:H514
74. Tomaselli GF, Marban E (1999) Cardiovasc Res 42:270
75. O'Neill SC, Miller L, Hinch R, Eisner DA (2004) J Physiol 559:121
76. Santana LF, Kranias EG, Lederer WJ (1997) J Physiol 503(1):21
77. Trafford AW, Diaz ME, Eisner DA (2001) Circ Res 88:195
78. Diaz ME, Graham HK, O'Neill SC, Trafford AW, Eisner DA (2005) Cell Calcium 38:391
79. Eisner DA, Sipido KR (2004) Circ Res 95:549
80. Goldhaber JI, Henderson SA, Reuter H, Pott C, Philipson KD (2005) Ann NY Acad Sci 1047:122

Adv Biochem Engin/Biotechnol (2008) 110: 47–66
DOI 10.1007/10_2008_102

Published online: 5 June 2008

Protein Interactions: Analysis Using Allele Libraries

Thomas G. Chappell[1] · Phillip N. Gray[2] (✉)

[1]Vista Biologicals, 2120 Las Palmas Drive, Carlsbad, CA 92011, USA

[2]Ambry Genetics, 100 Columbia #200, Aliso Viejo, CA 92656, USA
phillipngray@gmail.com

Abstract Interaction defective alleles (IDAs) are alleles that contain mutations affecting their ability to interact with their wild type binding partners. The locations of the mutations may lead to the identification of protein interaction domains and interaction interfaces. IDAs may also distinguish different binding interfaces of multidomain proteins that are part of large complexes, thus shedding light on large protein structures that have yet to be determined. IDAs may also be used in conjunction with RNAi to dissect protein interaction networks. Here, the wild type allele is knocked down and replaced with an IDA that has lost the ability to interact with a specific binding partner. As a result, interactions are disrupted rather than knocking out the entire gene. Thus, IDAs have the potential to be extremely valuable tools in protein interaction network analysis. IDAs can be isolated by reverse two-hybrid analysis, which was demonstrated over a decade ago, but high background levels caused by truncated IDAs have prevented its widespread adoption. We recently described a novel method for full-length allele library generation that eliminates this background and increases the efficiency of the reverse two-hybrid protocol (and IDA isolation) significantly. Here we discuss our strategy for allele library generation, the potential uses of IDAs as outlined above, and additional applications of allele libraries.

Keywords Allele library · Interaction defective allele · Protein interactions · Protein networks · Reverse two-hybrid

Abbreviations

AD	Transcriptional activation domain
3-AT	3-Amino-1,2,4-triazole

att	Attachment
bHLH	Basic helix–loop–helix
BxP	*att*B–*att*P recombination
CYH2	Ribosomal protein of the large (60S) ribosomal subunit that can mutate to cycloheximide resistance
DBD	DNA binding domain
5-FOA	5-Fluoro-orotic acid
HIS3	Histidine biosynthesis 3
HLH	Helix–loop–helix
IDAs	Interaction defective alleles
IPTG	Isopropyl β-ᴅ-thiogalactopyranoside
Kan^R	Neomycin phosphotransferase gene
LxR	*att*L–*att*R recombination
ORF	Open reading frame
PCR	Polymerase chain reaction
RNAi	RNA interference
URA3	Uracil biosynthesis 3

1 Introduction

As more researchers use a systems approach to study biological processes in higher eukaryotic organisms, there is a need for more robust tools that mimic many of the techniques that have been available for years in model organisms. The ability to overexpress genome-wide collections of open reading frames (ORFs), systematically alter gene expression using RNAi, and test protein–protein interactions using two-hybrid or coexpression has greatly expanded the toolsets. Yeast two-hybrid technology [1] has been used to create global protein–protein interaction maps in *Helicobacter pylori* [2], *Saccharomyces cerevisiae* [3, 4], *Caenorhabditis elegans* [5], and *Drosophila melanogaster* [5, 6]. More recently, the technology was used to produce preliminary human interactomes investigating the potential protein–protein interactions of millions of human protein pairs [7, 8]. These datasets are being mined by researchers to further characterize specific interactions and pathways – mapping functional domains and interaction surfaces by screening allele variants of wild type ORFs. In the case of interaction surfaces, isolation of interaction defective alleles (IDAs) from high coverage allele libraries allows the identification of residues and interfaces that mediate protein–protein or protein–nucleic acid interactions. IDAs may be isolated by reverse two-hybrid, which is a variation on yeast two-hybrid developed to identify *cis* and *trans* elements that disrupt protein interactions [9].

Most two-hybrid systems are based on split transcription factor technology. Briefly, transcription factors, such as Gal4p, are modular and consist of two domains: a DNA binding domain (DBD) and a transcriptional activation domain (AD). Each domain remains functional when physically separated, but transcription is only activated when the two domains are brought into close proximity. Thus, if two proteins, A and B, are fused to Gal4-DBD and Gal4-AD, respectively (i.e., DBD-A and AD-B), and an interaction occurs, the GAL4 transcription factor is reconstituted and capable of activating gene expression from promoters containing one or more Gal4p binding sites. Two-hybrid screens are conducted in auxotrophic strains of *S. cerevisiae* that possess reporter genes that are only expressed in response to a positive protein–protein interaction. This is accomplished by replacing the wild type promoters of genes involved in specific amino acid or nucleotide biosynthesis (e.g., HIS3 or URA3) with promoters containing specific transcription factor binding sites (such as GAL4 or LexA). In addition, exogenous genes such as β-galactosidase can be incorporated into yeast strains and protein interactions are further scored by enzymatic assay (i.e., color change of substrate) (Fig. 1).

Forward two-hybrid systems are utilized to identify protein–protein interactions, whereby a specific DBD fusion protein (“bait” protein) is cotransformed with a cDNA or ORF library cloned into the AD vector (“prey” library). Proteins expressed from the prey library that interact with the bait protein activate the reporter genes, allowing growth on medium lacking the appropriate supplements for the specific auxotrophic reporters. Each resulting yeast colony contains a single plasmid from the prey library that can then

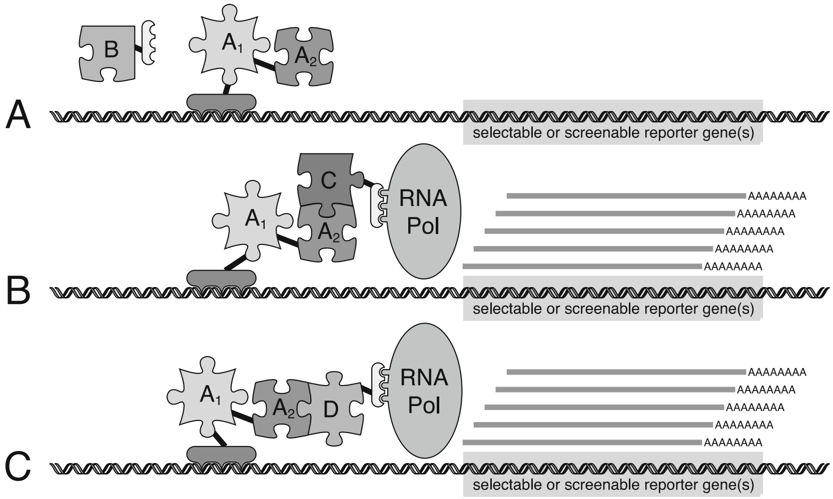

B
A1
A2
A
selectable or screenable reporter gene(s)
C
A1
A2
RNA
Pol
AAAAAAAA
B
selectable or screenable reporter gene(s)
A1
A2
D
RNA
Pol
AAAAAAAA
C
selectable or screenable reporter gene(s)

Fig. 1 Cartoon depiction of gene activation in a yeast two-hybrid system. “Bait” protein A contains two domains, and is expressed as a fusion protein with a DBD that localizes it upstream of one or more reporter genes. Introduction of “prey” proteins fused to ADs results in the localization of RNA polymerase to the reporter locus or loci, providing positive readout(s) for interacting proteins C and D, but not for noninteracting protein B

be isolated, reconfirmed by cotransformation with the original bait, and identified by sequence analysis.

In contrast, reverse two-hybrid systems are utilized to weaken or eliminate known protein–protein interactions. Starting with a protein–protein interaction that activates some or all of the reporter genes in a specific yeast strain, perturbations are made to the system in an attempt to attenuate the activity of one or more reporters. Perturbations can consist of the addition of potential small-molecule inhibitors [10], expression of possible inhibitory polypeptides, or the introduction of mutations into one or both of the binding partners. If the number of potential perturbations is relatively small, the status of the reporter genes can be determined simply by replica plating. For example, Amberg and coworkers used a defined set of 35 actin mutations to map actin–actin, actin–profilin, actin–Srv2p, and actin–SH3 domain interactions by simply screening cotransformants for loss of reporter activity [11].

For comprehensive coverage of an allele library of a binding partner using reverse two-hybrid, screening becomes virtually impossible. Therefore, negative selection or counter-selection reporters are incorporated into the two-hybrid yeast strain. Negative and counter-selection allows the isolation of rare IDAs from libraries consisting primarily of mutant alleles that behave as wild type. Negative selection can be performed with CYH2 as a reporter, where expression in the presence of cycloheximide is toxic to the cell [12]. Since there is no positive selection for CYH2 expression, the forward two-hybrid selection has to be done using another reporter, and interactions have to be screened for sensitivity to cycloheximide before a reverse two-hybrid selection can be initiated. This extra step can be eliminated by using a counter-selectable reporter gene, which allows for survival under one con-

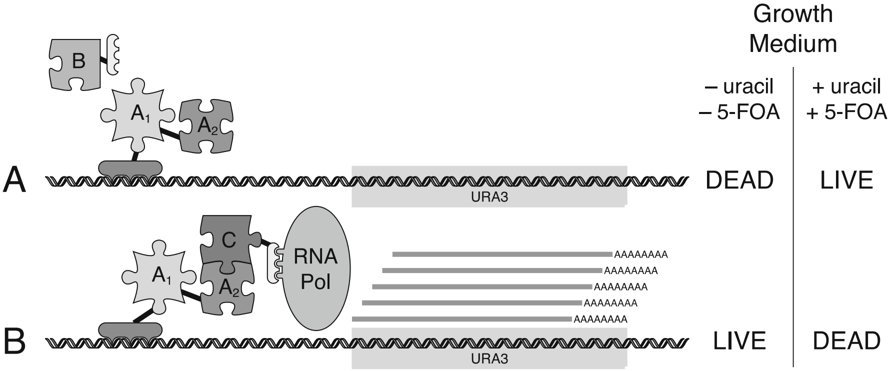


Fig. 2 Cartoon depiction of a reverse two-hybrid system. By using a counter-selectable reporter gene, media conditions can be altered to allow cell growth for either interacting or noninteracting protein pairs. In this example, the use of the *URA3* reporter results in growth on 5-FOA + uracil for prey protein B that does not interact with bait protein A

dition (during the forward selection) but is toxic under another (during the reverse selection). One such reporter is URA3; activation of URA3 allows survival of yeast in media lacking uracil, but activation in the presence of both uracil and the compound 5-fluoroorotic acid (5-FOA) is toxic to the cell (Fig. 2). Thus, when conducting reverse two-hybrid selection with URA3 or CYH2, only alleles containing mutations that either disrupt or attenuate the interaction will be resistant to 5-FOA (5-FOA^{R}) or cycloheximide (CYH2^{R}).

One major difficulty that has hampered the wider adoption of reverse two-hybrid technology is the fact that mutagenesis of an ORF leads to the generation of internal stop codons, resulting in truncated proteins within an allele library. Eighteen of the 61 nontermination codons are converted to ter-

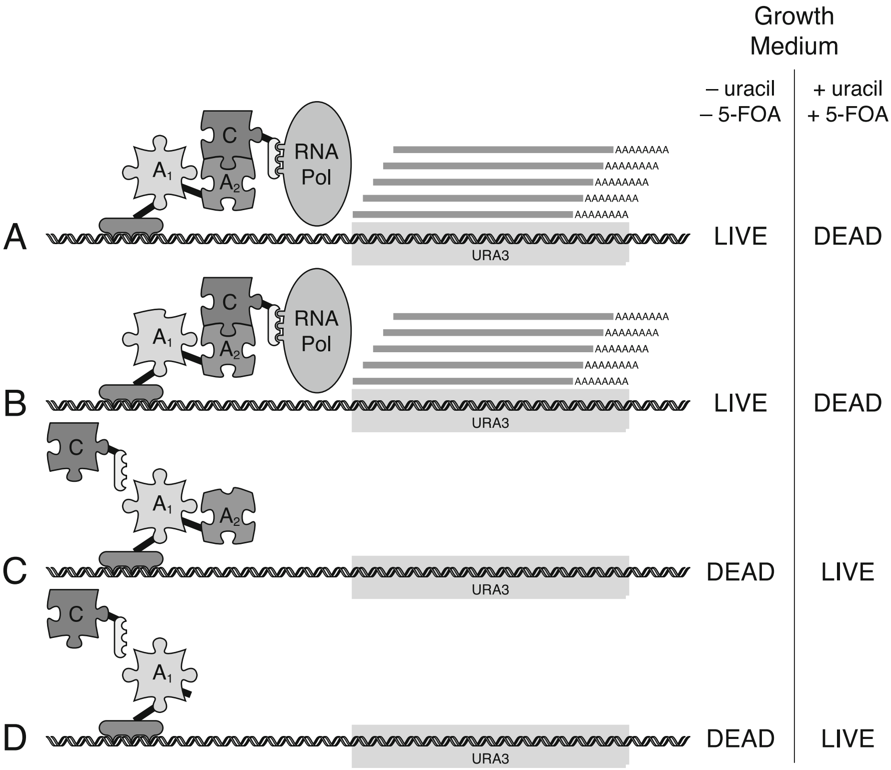


Fig. 3 Cartoon depiction of possible outcomes when an allele library of bait protein A is reverse screened against interacting prey protein C. *Panel A* shows the wild type interaction that does not grow on 5-FOA + uracil medium. *Panel B* shows an allele of protein A that does not disrupt the interaction with protein C. *Panel C* shows an interaction defective allele of *A*, in which an amino acid change in protein A disrupts the interaction with protein C. *Panel D* shows the most common type of allele isolated in traditional reverse two-hybrid screens, loss of a major portion of bait protein A through a frameshift or nonsense mutation within the ORF

mination codons with a single base change. Based on codon usage in the human genome, 30% of the single base changes within human ORFs will generate truncated proteins, with a significant proportion of those losing one or more protein domains. The usual result of a reverse two-hybrid screen with an allele library rich in truncated proteins is the isolation of a large number of IDAs, where virtually all isolates encode proteins containing internal termination codons (Fig. 3).

2 Allele Library Generation

The first step in characterizing protein interactions with reverse two-hybrid is allele library generation of one of the interacting partners. Typically, allele libraries are generated via the polymerase chain reaction (PCR) and must be cloned into the AD vector to express the alleles as an AD fusion. Initial protocols relied upon in vivo homologous recombination (i.e., gap repair) in the reporter yeast strain to clone the library into the AD expression vector, and simultaneously selected for IDAs and recombined plasmids. Gap repair in *S. cerevisiae* is an effective strategy to clone DNA fragments; however, when generating an allele library the number of clones recovered is not sufficient for a complex library (i.e., maximum number of clones containing single-codon changes evenly distributed throughout the ORF). Moreover, initial reverse two-hybrid screens using gap repair showed that greater than 97% of 5-FOA^R colonies contained alleles coding for truncated proteins [9, 13]. To screen out the truncated IDAs, a second, positive selection step following counter-selection on 5-FOA media was utilized. This required the addition of an easily detectable C-terminal fusion [13, 14] or epitope tag [15] to the expressed alleles. This method proved very labor intensive and made reverse two-hybrid protocols extremely inefficient. Thus, separating the small percentage of full-length IDAs from background resulting from truncated proteins proved to be a challenge and represented a technical obstacle that prevented widespread adoption of the technique.

We recently described an effective strategy that separates full-length selection from allele library generation and increases the efficiency of reverse two-hybrid screens 100-fold [16]. This strategy relies on Gateway™ technology, which is a recombinational cloning technology based on lambda phage recombination that facilitates the transfer of heterologous DNA sequences between vectors through site-specific attachment (*att*) sites [17–20]. We created a new Gateway™ vector, pDONR-Express, which facilitates the expression of ORFs as an N-terminal fusion to neomycin phosphotransferase and confers kanamycin resistance in *Escherichia coli*. By selecting against truncated proteins in *E. coli* prior to IDA selection in yeast, almost all background normally associated with reverse two-hybrid screens is eliminated. Moreover, when compared to gap repair mediated library assembly, combining Gateway™ recombination with the efficiency of *E. coli* transformation allows larger (10^6–10^7), more complex allele libraries to be evaluated.

This new method is outlined in Fig. 4. First, allele libraries are generated by PCR. The resulting PCR products are flanked by *att*B sites, which recombine with *att*P sites located in the pDONR-Express vector (B×P reaction). Ideally, libraries will consist of alleles containing single point mutations. This can be achieved by selecting the appropriate DNA polymerase and conditions based on the size of the ORF under study. The error rate of *Taq* polymerase under native conditions is sufficient when analyzing ORFs in the range of 300–800 bp. However, for small ORFs (∼ 100–200 bp), *Taq* polymerase combined with mutagenic PCR conditions is necessary for sufficient error generation. For genes up to ∼ 2 kb, the error rate of high fidelity polymerase is sufficient for introducing single point mutations [21]. In addition, allele libraries may be generated without PCR by utilizing mutagenic strains of *E. coli* [22]. However, the ORF of interest must be propagated in a vector containing *att*B sites (Gateway expression vector).

Next, the resulting entry clone allele library is transformed into *E. coli* and expression is induced using IPTG. Only entry clones expressing full-length ORFs will generate the neomycin phosphotransferase fusion necessary for

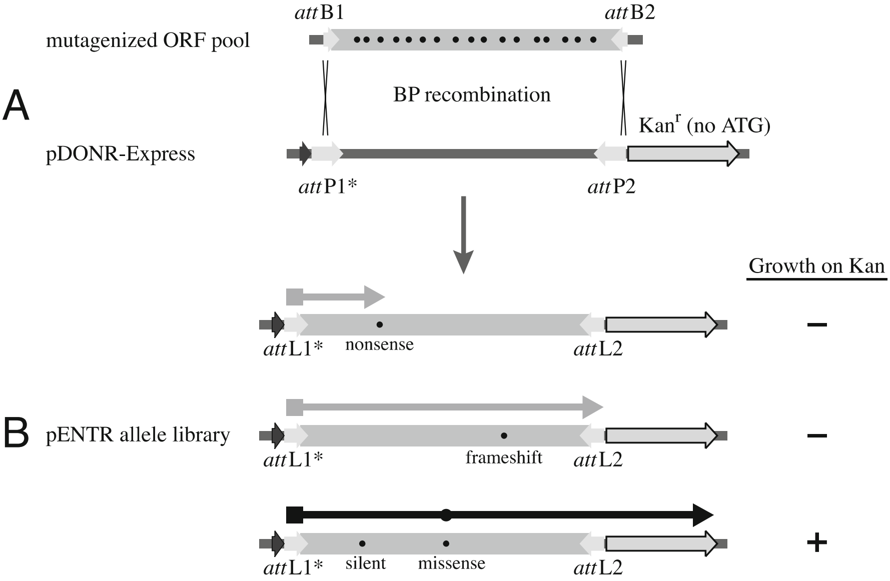


Fig. 4 Use of pDONR-Express to generate allele libraries containing full-length ORFs. By selecting in *E. coli* only those alleles that allow growth on kanamycin, the final library is greatly enriched in ORFs that contain only silent or missense mutations, eliminating most nonsense and frameshift mutations

kanamycin resistance. The full-length enriched entry clone library contains alleles flanked by *att*L sites, which are capable of recombination with any vector containing *att*R sites (L × R reaction). For reverse two-hybrid screens, allele libraries are LxR crossed into the AD vector, cotransformed into yeast with a DBD vector containing the protein binding partner, and plated on to media containing uracil and 5-FOA. Wild type interactions and alleles containing mutations outside the interaction domain transcribe the URA3 reporter and are 5-FOA sensitive. Alleles containing mutations in the interaction domain do not activate the URA3 reporter, require uracil for growth, and are 5-FOA^R.

3 Interaction Domain Identification

Reverse two-hybrid screens facilitate the isolation of IDAs, most of which will contain a unique mutation that disrupts, or attenuates, the interaction under study. IDAs from a specific screen may have mutations localized within a specific region of the protein, thus identifying the interaction domain. This is demonstrated in the analysis of the interaction between Id1 and MyoD1, an interaction mediated by the HLH region in Id1 and a bHLH domain in MyoD1 [23–26]. An allele library of the full-length MyoD1 ORF ($\sim$ 1 kb) was generated and the vast majority (99%) of alleles failed to grow in the presence of 5-FOA (i.e., most alleles contained a wild type sequence in the interaction domain). Several IDAs were isolated and a multiple sequence alignment of the translated alleles revealed mutation clustering to the bHLH domain, a defined region comprising only 20% of the protein molecule.

These mutations were mapped onto the crystal structure of a MyoD bHLH homodimer (PDB entry code 1MDY) and most localized to one side of either helix 1 or helix 2 at the interaction interface (Fig. 5). The effect of each mutation on the interaction was evaluated by 3-AT titration, which allows the relative strength of the interaction to be measured indirectly by correlating it with the level of HIS3 reporter gene transcription. The phenotypes observed under 3-AT selection were compared to the location of the point mutation in the crystal structure of each allele and good correlations were found. For the seven alleles containing mutations at the interaction interface, five were unable to survive in the presence of a minimum amount of 3-AT (10 mM), suggesting a completely disrupted interaction. In contrast, most alleles containing mutations outside the interaction interface were capable of growth on higher concentrations of 3-AT, suggesting these mutations only weakened the interaction.

The characterization of the interaction between *C. elegans* proteins GLA-3 and MPK-1 is another example of interaction domain identification using IDAs. A reverse two-hybrid screen with a MPK-1 allele library revealed 20 unique mutations leading to amino acid substitutions within 17 codons [27]. Nine of these 17 amino acid residues clustered to a small region on the surface

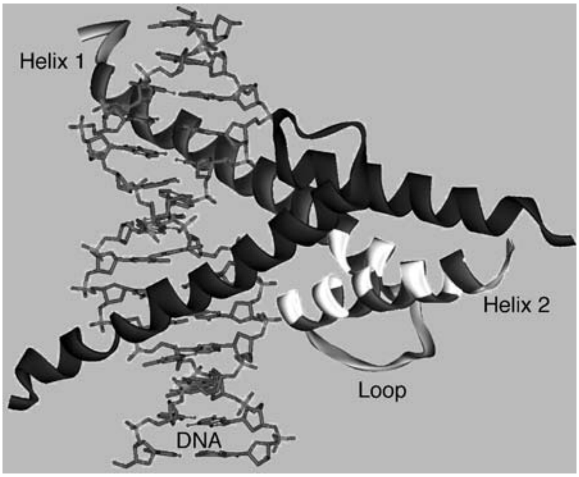


Fig. 5 Localization of interaction defective alleles on the structure of MyoD. Residues shown in *off-white* were identified as interaction defective allele locations. The screen identified almost all the amino acids at the interaction interface of the protein dimer

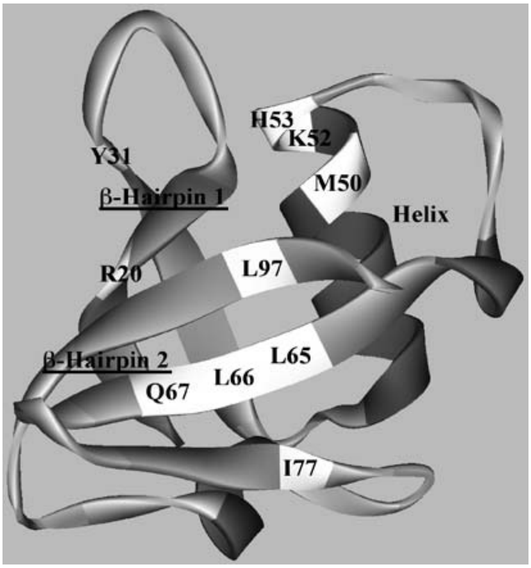


Fig. 6 Localization of interaction defective alleles on the structure of RalGDS. Residues shown in *off-white* were identified as interaction defective allele locations. In this case, there was little clustering of the alleles, which localized mostly to the hydrophobic core of the protein

of MPK-1, a region structurally conserved between MPK-1 and its mammalian homologue ERK2 that has been shown to function as a docking site for both regulators and targets of MAPK [28].

The analysis of Krev1 and the ras association (RA) domain of RalGDS is an example of a screen that yielded IDAs containing mutations that do not cluster in the primary amino acid sequence. An allele library of RalGDS-RA domain was generated and several IDAs containing single point mutations were recovered. Some clustering is seen, but a defined interaction interface was not apparent. When the mutations are mapped onto the RalGDS-RA crystal structure (PDB entry 1LFD [29]), most IDAs contained mutations in the hydrophobic core of the protein, suggesting the overall structure of the domain (rather than individual amino acids at the interaction interface) plays a critical role in facilitating interaction with Krev1 (Fig. 6).

4 Interaction Domain Identification in Other Binary Systems

Other systems available for identifying binary protein interactions include the split ubiquitin membrane yeast two-hybrid system, which allows the identification of membrane protein interaction partners [30]. This system may be used to screen allele libraries for IDAs, thus allowing interaction domains in membrane proteins to be readily identified. Mammalian systems available to identify binary interactions include the traditional split transcription factor system [31], a protein splicing system [32], and mammalian protein–protein interaction trap (MAPPIT) [33]. These systems allow proteins to be expressed in their native background so all posttranslational modifications will occur to the proteins of interest. These systems may be modified to screen allele libraries for IDA isolation.

Recent advances in gene delivery have made IDA isolation from mammalian cells a more viable approach. For instance, lentivirus-mediated gene delivery at low MOI allows for the transduction of one library member per cell [34], negating one of the reasons yeast was used in the first place for protein–protein interaction screens. If combined with high throughput microscopy, IDAs may be isolated without counter-selection if the system utilizes an easily detected reporter gene, such as GFP. Such high content screening protocols have been utilized for phenotype analysis using RNAi and transfection arrays [35, 36].

Reverse two-hybrid systems can be engineered to analyze DNA–, RNA–, and small-molecule–protein interactions. DNA–protein interactions are analyzed using a one-hybrid system [37]. In contrast to two-hybrid systems, one-hybrid systems lack a DBD, or bait, fusion protein. The system works by integrating the DNA sequence of interest upstream of reporter genes. In forward one-hybrid systems, cDNA libraries are screened as AD fusions to identify interacting partners. For reverse one-hybrid, allele libraries are made of the interacting protein and the interaction interface is defined. In addition, two-hybrid systems are available to identify protein–RNA interac-

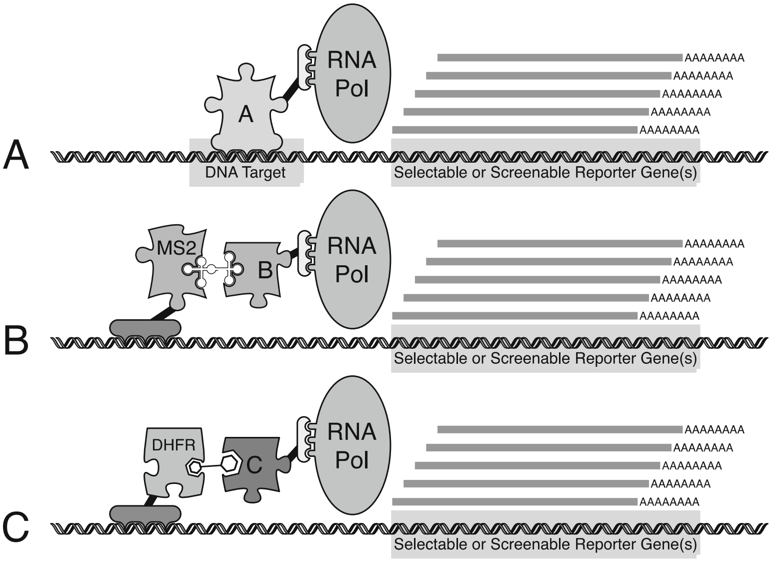

A
A
RNA Pol
AAAAAAAA
DNA Target
Selectable or Screenable Reporter Gene(s)
B
MS2
B
RNA Pol
AAAAAAAA
Selectable or Screenable Reporter Gene(s)
C
DHFR
C
RNA Pol
AAAAAAAA
Selectable or Screenable Reporter Gene(s)

Fig. 7 Cartoon depiction of variations of the yeast two-hybrid system. *Panel A* shows a one-hybrid system used to identify DNA binding proteins specific for DNA sequences introduced upstream of the reporter gene(s). *Panel B* shows a three-hybrid system to isolate RNA binding proteins. The RNA of interest is expressed as a fusion to an RNA structure that binds to the MS2 protein, creating an RNA–protein bait. *Panel C* shows a three-hybrid system to identify small-molecule binding proteins. In this case, the small molecule of interest is synthesized linked to methotrexate. The synthetic molecule can then bind to DHFR to create a small-molecule–protein bait. Each of these systems can be adapted to isolate IDAs by reverse screening

tions. This system consists of a fusion of the bacteriophage MS2 coat protein fused to LexA DNA binding protein. The RNA molecule of interest is transcribed as a fusion to an RNA sequence that binds MS2 coat protein, resulting in a LexA-MS2-RNA complex that can be used as a bait to identify interacting proteins from cDNA libraries [38–41]. Allele libraries may be screened against this complex to identify RNA–protein interaction domains. Finally, small-molecule–protein interaction interfaces may be identified using a three-hybrid system, whereby a small molecule is conjugated to methotrexate and subsequently bound to a LexA-DHFR-fusion [42, 43] (Fig. 7).

5 Multiprotein Complex Analysis

Complex structural genomics refers to the 3D structure determination of protein complexes and how these complexes interact with one another to form multiprotein complexes [44]. The majority (80%) of eukaryotic proteins are composed of multiple domains [45–47] and most interactions between multidomain proteins only involve one domain [48]. High throughput generation of interactome datasets has resulted in extensive network diagrams of protein–protein interactions. In many cases, individual protein nodes on these networks are linked to multiple partners, with some representing large protein complexes. Further mapping of the interaction pairs is then needed to determine whether specific interactions are independent, allosteric (either inhibitory or stimulatory), or mutually exclusive (see Fig. 8). We used reverse two-hybrid to study a binary interaction within the COPII complex. The COPII complex is involved in membrane trafficking and consists primarily of four proteins: Sec23p, Sec24p, Sec13p, and Sec31p [49–51]; but it is thought to

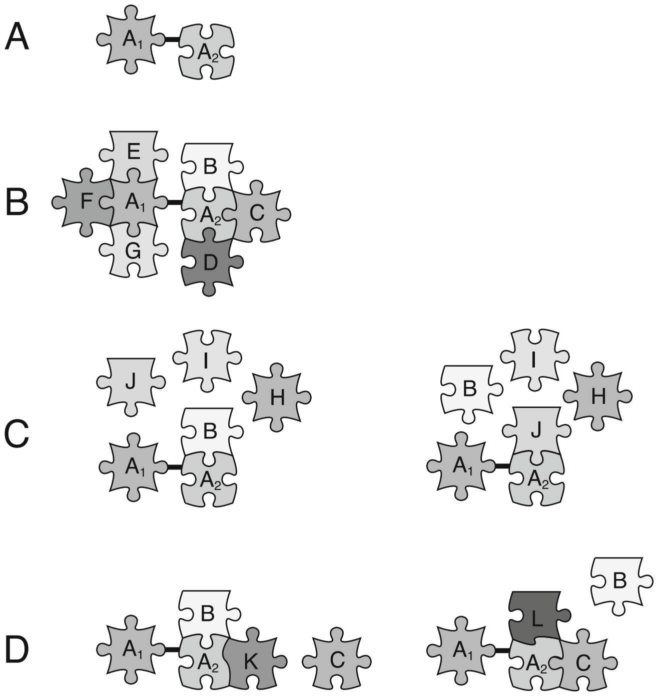


Fig. 8 Cartoon depiction of the types of interactions possible in multiprotein complexes. *Panel A* shows multidomain protein A without any interacting proteins. In *panel B*, proteins B–G interact with protein A independently; each protein can interact with A regardless of A's interactions with other proteins in the set. *Panel C* shows mutually exclusive interactions between protein A and proteins B, H, I, and J. In this case, only one of the four proteins can occupy the shared interaction site on protein A. In *panel D*, interactions are altered by allosteric effects. Binding of protein B results in conformation change in A that eliminates its interaction with protein C, but now allows its interaction with protein K. Similarly, binding of protein C eliminates the interaction site for protein B, allowing protein L to bind

exist in forms where Sfb2p or Sfb3p substitute for Sec24p, altering the cargo selection of the COPII coated transport vesicles [52]. Several models of the COPII complex have been proposed, but the structure of the entire complex has not been completely determined.

In the ProQuest two-hybrid system, the Sec23p–Sec24p is a weak interaction, weakly activating the *HIS3* reporter and showing no activation of the *URA3* reporter. The interaction between Sec23p and Sfb3p is stronger, activating both *HIS3* and *URA3*. An allele library of *SEC23* was generated and reverse screened against Sfb3p, resulting in the isolation of several dozen IDAs that contained mutations in multiple domains of the protein. When plotted on the Sec23p crystal structure, mutations appeared to localize to four distinct

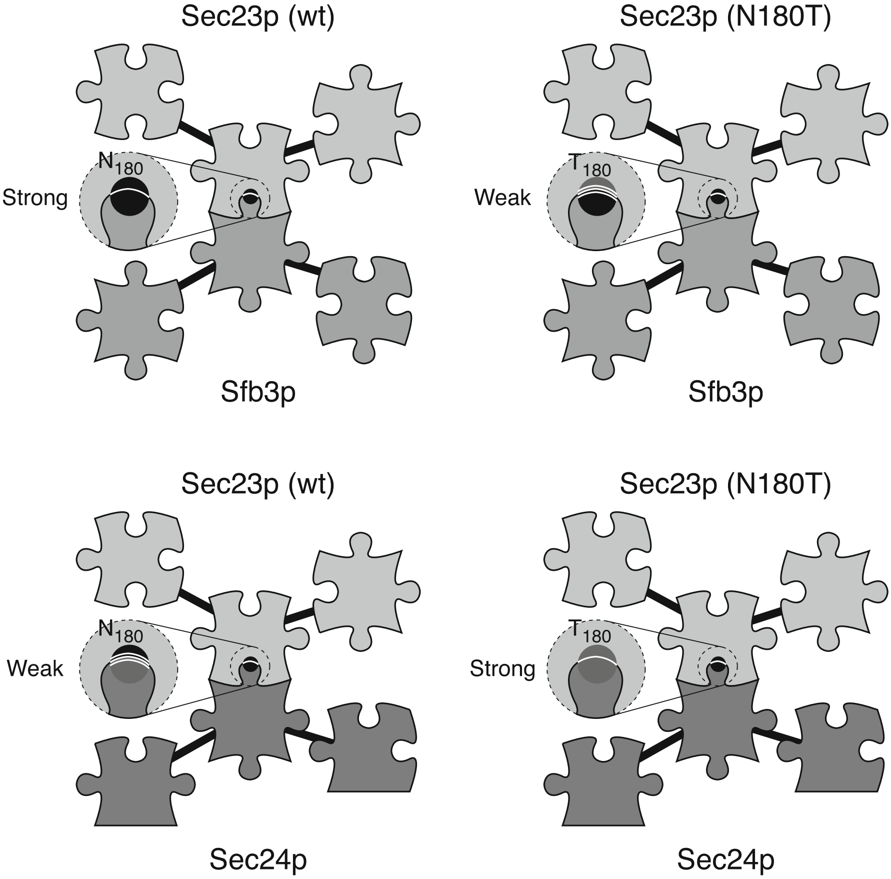


Fig. 9 Cartoon depiction of the N180T allele of *SEC23*. Through forward and reverse screens with Sfb3p and Sec24p, an allele of Sec23p was isolated that "flipped" the binding affinity of Sec23p for Sfb3p and Sec24p. The N180T allele of *SEC23* binds Sec24p more strongly than Sfb3p

domains: (1) zinc finger, (2) trunk domain, (3) β-sandwich, and (4) Gelsolin domain. The only mutation that strongly disrupted the interaction was a single mutation that localized to the trunk domain—the domain containing residues at the interaction interface between Sec23p and Sec24p. All other mutations mapped to either the zinc finger, β-sandwich, or gelsolin domains and only weakened the interaction—abolishing the activation of the *URA3* reporter but retaining reasonable activation of *HIS3*. This allele of *SEC23* (an N to T change at position 180) resulted in Sfb3p binding behavior similar to the binding behavior of Sec24p to wild type Sec23p. Most interestingly, as diagrammed in Fig. 9, this allele now showed strong binding to Sec24p, activating both the *HIS3* and *URA3* reporters. Thus this Sec23p allele swapped the binding strengths of the two partners, Sec24p and Sfb3p to Sec23p relative to the wild type form.

IDAs therefore may be used as tools to dissect interactions and relative affinities in large multiprotein structures. The entire COPII complex has yet to be solved, so combining these types of IDAs with affinity purification protocols may help to identify all protein interaction interfaces within the

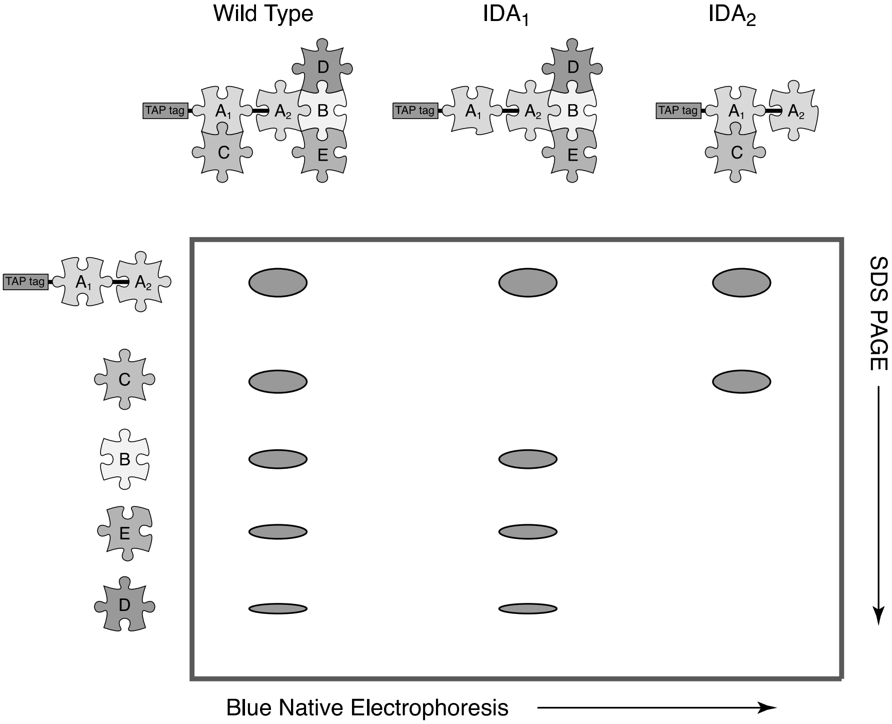


Fig. 10 Cartoon of blue native/SDS-PAGE for multiprotein complex analysis. Blue native PAGE retains complex structures, which can then be resolved into individual protein components using a second dimension SDS-PAGE step

complex. This is possible by performing affinity purification (such as tandem affinity purification, TAP) with several IDAs of the same protein, each with a mutation in a different potential interaction interface, and analyzing the purified protein complex with mass spectroscopy to monitor changes in complex composition. Interaction interfaces are identified when specific IDAs are paired with a particular protein dropping out of the complex.

Figure 10 illustrates the strategy for determining higher order protein structures using IDAs. Wild type and candidate IDAs are affinity purified to isolate interacting proteins. This can be accomplished by TAP (reviewed by Puig et al. [53]) or standard affinity purification protocols [54]. Next, the purified complexes are run out on blue native PAGE gels [55, 56]. Native gel purification is followed by standard SDS-PAGE to separate the individual proteins in each complex. In this example, purification of the wild type complex through an affinity tag on protein A yields the binary interactions A with B and A with C that were initially identified by two-hybrid analysis, but additionally shows proteins D and E to be members of the complex. Since the complex runs at a single molecular weight on a blue native PAGE gel, all five proteins appear to be present in a single complex. By analyzing the complexes purified when IDA_1 and IDA_2 are expressed with affinity tags, the interaction(s) that bind D and E to the complex can be shown to involve the interaction of A with B and not A with C.

Combining IDA isolation (R2H) with IDA-TAP experiments is readily available for complex structure determination of yeast protein complexes. For human protein complexes, the mammalian two-hybrid systems described above may be utilized.

6 Protein Interaction Network Analysis

Protein–protein interaction maps generated by large-scale yeast two-hybrid [2–8, 57] and affinity purification projects [54, 58] have features of scale-free networks. Scale-free networks are characterized by a power-law degree distribution, where the probability that a protein has k interactions follows $P(k) \sim k^{-\gamma}$ (γ is the degree exponent). As a result, the networks consist primarily of proteins with a small number of interactions and very few highly connected hub proteins. This feature of scale-free networks makes them resistant to system failure when random proteins (nodes) are knocked out or disabled. However, system failure will result if hub proteins are targeted for knockout (reviewed by Barabasi and Oltvai [59]). In support of this model, analysis of the *S. cerevisiae* protein interaction network revealed that highly connected hub proteins tended to be coded for by essential genes that have a lethal phenotype when knocked out [60].

Current approaches to studying protein interaction networks in cells are limited to knocking out an individual protein, or node. This can be accomplished with techniques such as RNAi expression or gene deletion studies.

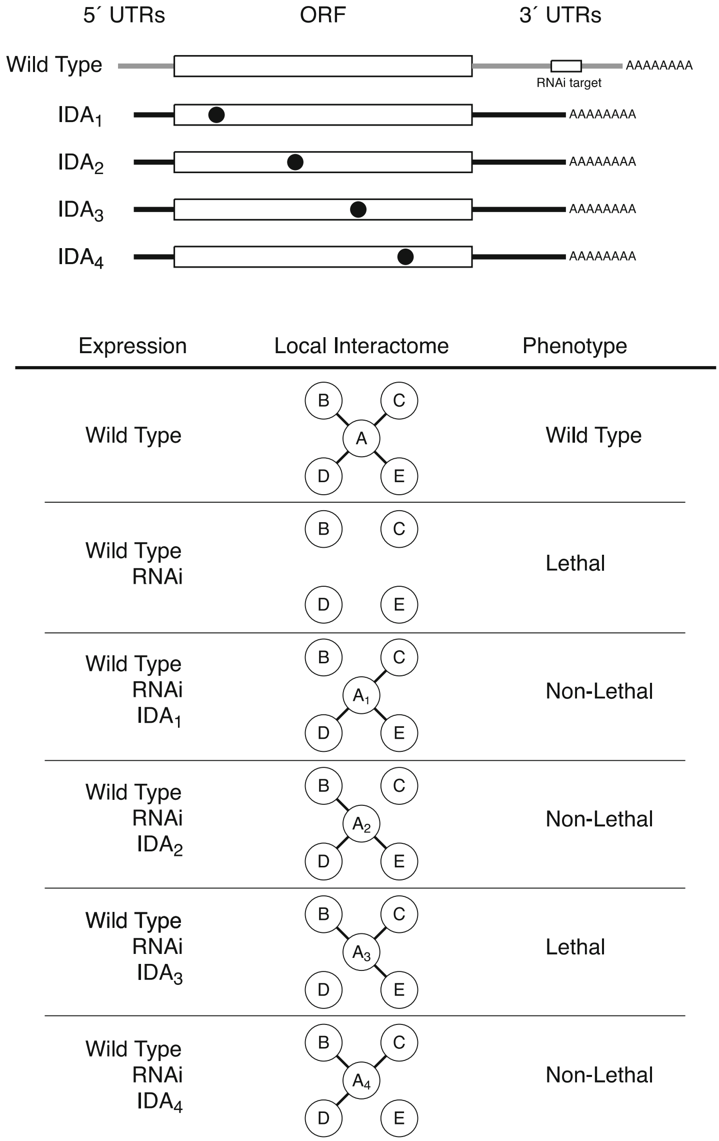


Fig. 11 Analysis of a local interactome using RNAi knockdown in conjunction with coexpression of interaction defective alleles. The lethal phenotype of the depletion of protein A can be isolated to its interaction with protein D through a combination of coexpression studies

However, if hub proteins are targeted, large regions of the network may fail, which may result in lethality. If specific edges (i.e., interactions) are targeted, the network as a whole may still function and more precise information gained. Disrupting a specific edge while maintaining all nodes is theoretically possible by isolating IDAs that have lost the ability to interact with one protein, but maintain interactions with all others. Such IDAs have been described in the literature [61–63]. With the recent publication of several metazoan protein interaction networks [5–8], the introduction of IDAs into these cells/organisms, and simultaneously knocking down the wild type gene with techniques such as RNAi, has the potential to analyze protein interaction networks as never before.

This strategy is outlined in Fig. 11. First, allele libraries are generated for hub protein A at the center of the local interactome. Reverse two-hybrid screens are conducted with each interacting partner (B–E) to isolate partner specific IDAs. The IDAs are then expressed in the native host cell while simultaneously knocking down the wild type ORF with RNAi and phenotypes are analyzed. In the illustrated case, the lethality of using RNAi to knock down protein A is linked to its interaction with protein D, rather than proteins B, C, or E. This approach can be extended to analyze nonlethal phenotypes that can be scored by any type of visualization or biochemical assay.

7 Summary

We have described a highly efficient method for full-length allele library generation. When combined with reverse two-hybrid technology, several IDAs may be isolated that all contain single point mutations within a specific region of the protein. Mutation clustering allows the identification of potential protein interaction domains. IDAs may also be used as tools in downstream applications for multiprotein complex determination and protein interaction network analysis.

References

1. Fields S, Song O (1989) Nature 340:245
2. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V, Chemama Y, Labigne A, Legrain P (2001) Nature 409:211
3. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) Nature 403:623
4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) Proc Natl Acad Sci USA 98:4569

5. Stanyon CA, Liu G, Mangiola BA, Patel N, Giot L, Kuang B, Zhang H, Zhong J, Finley RL Jr (2004) Genome Biol 5:R96
6. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) Science 302:1727
7. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Nature 437:1173
8. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) Cell 122:957
9. Vidal M, Brachmann RK, Fattaey A, Harlow E, Boeke JD (1996) Proc Natl Acad Sci USA 93:10315
10. Huang J, Schreiber SL (1997) Proc Natl Acad Sci USA 94:13396
11. Amberg DC, Basart E, Botstein D (1995) Nat Struct Biol 2:28
12. Leanna CA, Hannink M (1996) Nucleic Acids Res 24:3341
13. Shih HM, Goldman PS, DeMaggio AJ, Hollenberg SM, Goodman RH, Hoekstra MF (1996) Proc Natl Acad Sci USA 93:13896
14. Endoh H, Walhout AJ, Vidal M (2000) Methods Enzymol 328:74
15. Barr RK, Hopkins RM, Watt PM, Bogoyevitch MA (2004) J Biol Chem 279:43178
16. Gray PN, Busser KJ, Chappell TG (2007) Mol Cell Proteomics 6:514
17. Bushman W, Thompson JF, Vargas L, Landy A (1985) Science 230:906
18. Landy A (1989) Annu Rev Biochem 58:913
19. Ptashne M (1992) A genetic switch: phage lambda and higher organisms. Cell Press, Cambridge
20. Weisberg RA, Landy A (1983) In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA (eds) Lambda II. Cold Spring Harbor Press, Cold Spring Harbor, New York, p 211
21. Cline J, Braman JC, Hogrefe HH (1996) Nucleic Acids Res 24:3546
22. Camps M, Naukkarinen J, Johnson BP, Loeb LA (2003) Proc Natl Acad Sci USA 100:9727
23. Davis RL, Weintraub H, Lassar AB (1987) Cell 51:987
24. Weintraub H, Davis R, Tapscott S, Thayer M, Krause M, Benezra R, Blackwell TK, Turner D, Rupp R, Hollenberg S et al (1991) Science 251:761
25. Benezra R, Davis RL, Lassar A, Tapscott S, Thayer M, Lockshon D, Weintraub H (1990) Ann NY Acad Sci 599:1
26. Finkel T, Duc J, Fearon ER, Dang CV, Tomaselli GF (1993) J Biol Chem 268:5
27. Kritikou EA, Milstein S, Vidalain PO, Lettre G, Bogan E, Doukoumetzidis K, Gray P, Chappell TG, Vidal M, Hengartner MO (2006) Genes Dev 20:2279
28. Zhang J, Zhou B, Zheng CF, Zhang ZY (2003) J Biol Chem 278:29901
29. Huang L, Hofer F, Martin GS, Kim SH (1998) Nat Struct Biol 5:422
30. Stagljar I, Korostensky C, Johnsson N, te Heesen S (1998) Proc Natl Acad Sci USA 95:5187

31. Kuroda K, Kato M, Mima J, Ueda M (2006) Appl Microbiol Biotechnol 71:127
32. Kanno A, Ozawa T, Umezawa Y (2006) Anal Chem 78:556
33. Eyckerman S, Verhee A, der Heyden JV, Lemmens I, Ostade XV, Vandekerckhove J, Tavernier J (2001) Nat Cell Biol 3:1114
34. Naldini L, Blomer U, Gallay P, Ory D, Mulligan R, Gage FH, Verma IM, Trono D (1996) Science 272:263
35. Bailey SN, Ali SM, Carpenter AE, Higgins CO, Sabatini DM (2006) Nat Methods 3:117
36. Wheeler DB, Carpenter AE, Sabatini DM (2005) Nat Genet 37(Suppl):S25
37. Li JJ, Herskowitz I (1993) Science 262:1870
38. Bardwell VJ, Wickens M (1990) Nucleic Acids Res 18:6587
39. Lowary PT, Uhlenbeck OC (1987) Nucleic Acids Res 15:10483
40. SenGupta DJ, Zhang B, Kraemer B, Pochart P, Fields S, Wickens M (1996) Proc Natl Acad Sci USA 93:8496
41. Uhlenbeck OC, Carey J, Romaniuk PJ, Lowary PT, Beckett D (1983) J Biomol Struct Dyn 1:539
42. Baker K, Sengupta D, Salazar-Jimenez G, Cornish VW (2003) Anal Biochem 315:134
43. Bronson JE, Mazur WW, Cornish VW (2008) Mol Biosyst 4:56
44. Bravo J, Aloy P (2006) Curr Opin Struct Biol 16:385
45. Apic G, Gough J, Teichmann SA (2001) Bioinformatics 17(Suppl 1):S83
46. Apic G, Gough J, Teichmann SA (2001) J Mol Biol 310:311
47. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Science 300:1701
48. Aloy P, Russell RB (2004) Nat Biotechnol 22:1317
49. Antonny B, Gounon P, Schekman R, Orci L (2003) EMBO Rep 4:419
50. Lederkremer GZ, Cheng Y, Petre BM, Vogan E, Springer S, Schekman R, Walz T, Kirchhausen T (2001) Proc Natl Acad Sci USA 98:10704
51. Stagg SM, Gurkan C, Fowler DM, LaPointe P, Foss TR, Potter CS, Carragher B, Balch WE (2006) Nature 439:234
52. Karhinen L, Bastos RN, Jokitalo E, Makarow M (2005) Traffic 6:562
53. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) Methods 24:218
54. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Nature 440:631
55. Schägger H (2001) Methods Cell Biol 65:231
56. Schägger H, von Jagow G (1991) Anal Biochem 199:223
57. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) Science 303:540
58. Van Leene J, Stals H, Eeckhout D, Persiau G, Van De Slijke E, Van Isterdael G, De Clercq A, Bonnet E, Laukens K, Remmerie N, Henderickx K, De Vijlder T, Abdelkrim A, Pharazyn A, Van Onckelen H, Inze D, Witters E, De Jaeger G (2007) Mol Cell Proteomics 6:1226
59. Barabasi AL, Oltvai ZN (2004) Nat Rev Genet 5:101
60. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Trends Genet 20:227

61. Inouye C, Dhillon N, Durfee T, Zambryski PC, Thorner J (1997) Genetics 147:479
62. Jiang R, Carlson M (1996) Genes Dev 10:3105
63. Vidal M, Braun P, Chen E, Boeke JD, Harlow E (1996) Proc Natl Acad Sci USA 93:10321

Adv Biochem Engin/Biotechnol (2008) 110: 67–80
DOI 10.1007/10_2007_091

Published online: 29 January 2008

Identification of Protein–Protein Interactions by Mass Spectrometry Coupled Techniques

Mohamed Abu-Farha · Fred Elisma · Daniel Figeys (✉)

Ottawa Institute of Systems Biology (OISB), University of Ottawa, 451 Smyth Road.
Ottawa, Ontario, K1H 8M5, Canada
dfigeys@uottawa.ca

Abstract The use of mass spectrometry in protein identification has revolutionized the field of proteomics. Coupled to various affinity purification techniques, mass spectrometry is used to identify protein–protein interactions. This chapter looks at the use of these affinity purification techniques in the identification of protein interactions. Various tags are used to purify protein complexes including tandem affinity purification. The FLAG tag is another commonly used tag which is a small tag that tends not to interfere with the protein function. These different affinity purification methods are used to purify proteins that are further identified by either ESI-MS or MALDI-MS.

Keywords Affinity purification · Electrospray · Interactome · MALDI · Tandem Affinity purification

Abbreviations

MS	Mass spectrometry
Y2H	Yeast two hybrid
TAP	Tandem affinity purification
ESI	Electrospray ionization
MALDI	Matrix assisted laser desorption ionization
HA	Hemagglutinin
IP	ImmunoPurification

1 Introduction

The past decade has experienced a huge leap in the amount of data generated from different areas of life sciences. One of these advancements was the completion of the human genome project in 2003 [1]. The new challenge was to understand the function of all the genes identified by the human genome project shifting the focus from the DNA to proteins. Since proteins compose most of the functional units in the cell, the complete understanding of their role in the cell is very critical in understanding how the cell functions. Proteomics focuses on understanding the many aspects of proteins that can involve their structures, modifications, localization and their protein–protein interactions [2]. Proteomics can be further subdivided into expression and functional proteomics. Expression proteomics studies changes in protein expression under different conditions compared to normal cells. These studies can include changes in protein expression between cells exposed to different drugs, different types of stress and disease state and normal cells. This field of study has evolved from the traditional one-protein scale (Western blot analysis) to two-dimensional gels [3] and finally to large-scale analysis of changes in protein expression using techniques like isotope labeling [4] such as stable isotope labeling with amino acids in cell culture (SILAC) [5]. The second major area in proteomics is functional proteomics, which looks at understanding protein functions and elucidating their role in the cell. Identification of protein–protein interaction has emerged as one of the most important ways to understanding the functions of different proteins. This stems from the fact that most proteins are not "island" [6] and function by forming different complexes under different conditions [7]. Hence, understanding protein–protein interactions in the cell offers an invaluable tool to understanding the functions of many unknown proteins [7].

A key advancement that revolutionized the field of proteomics is the use of mass spectrometry (MS) in protein identification. Use of MS in biomolecule analysis had been extremely inefficient due to the fact that biomolecules are large and polar ions making their transfer to the gas phase very hard. Use of MS in biomolecule analysis had to wait the development of ionization methods that can solve these problems. It was only possible to readily analyze these large molecules in the past two decades as a result of the development of electrospray Ionization by John Bennett Fenn [8] and matrix-assisted laser desorption/ionization (MALDI) by Koichi Tanaka [9]. The second key development was the efforts in sequencing and annotating genomes, which lead to a wealth of sequences database that became the foundation of high-throughput bioinformatics for protein and peptide analysis by mass spectrometry. Also, the development of suites of separation techniques and data analysis reinforced the utility of these approaches. Together these advancements were very important in the increasing interest in MS and its development to become the key tool in proteomic research [10]. This chapter will focus on the use of MS coupled techniques to identify protein–protein interactions. We will focus on the process of identifying protein interactors by affinity purification and their identification by MS.

2 Mapping a Protein–Protein Interaction

The less-than-anticipated number of genes identified by the human genome has further enforced the idea that proteins can have multifunction in the cell, although one should not discount the importance of no-coding RNA as regulatory elements and functional elements [11, 12]. The functions of a single protein can vary according to its interaction partners and its localization. Interesting examples, such as moonlighting of proteins illustrate the multifunctional aspects [13]. Many techniques can be used to look at protein–protein interaction in the cell. The most common high-throughput techniques are the yeast two hybrids (Y2H) and affinity purification coupled to MS [14]. Y2H assays are based on the fact that transcription factors have a DNA binding domain and an activation domain. In this assay, the two components are separated and fused to two potentially interacting proteins [15]. Upon their interaction, these proteins will activate a reporter gene that is easily detected. This technique is widely used to test for protein–protein interaction [15]. The main criticism that is used against Y2H is that it can have a false-positive rate as high as 50% [16]. This high false-positive rate could be due to many things including the fact that the assay investigates the interaction between over-expressed fusion proteins in the yeast nucleus. An alternate method that has been greatly used is affinity purification coupled to MS. In this method, a protein of interest is tagged with a specific tag and then the tag is used in the purification of protein complex using an anti-tag system immobilized on a solid support. Protein sample is then separated using one of multiple separation methods and then digested into small peptides that are identified by MS. This strategy offers the advantage of using anti-tag systems that are highly specific and commercially available in different formats. This technique has also been made easier with the advancement in molecular biology techniques that make the tagging process very robust and simple. The use of different tags in the affinity purification of protein complexes and their identification by MS will be discussed in the rest of this chapter.

2.1 Affinity Purification of Protein Complexes

In recent years, a number of affinity-based protein purification methods have been used to identify protein–protein interactions. These methods typically

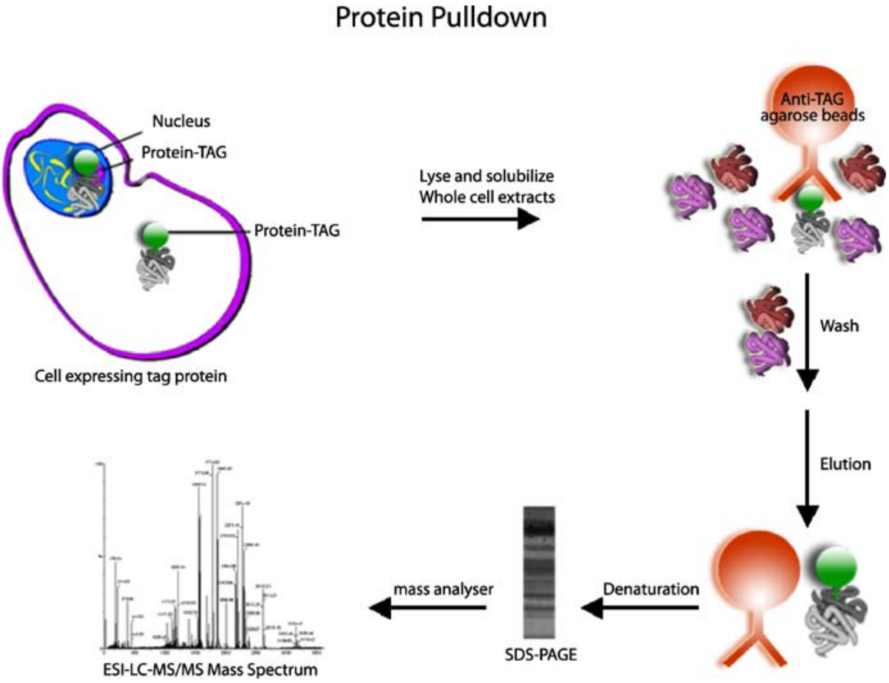


Fig. 1 Diagram showing the different steps in the anti-tag protein purification system

depend on the expression of a protein of interest with an affinity tag. These tags are generally made of short hydrophilic peptides such as the FLAG, hemagglutinin (HA), or poly-His tags. Other tags are small proteins like GST, thioredoxin or GFP tag. Figure 1 shows the outline of the affinity purification process. Many of these tags can be used in combination with MS to identify protein–protein interactions. In this section we will focus on large scale immunopurification (IP) coupled to MS strategies for the identification of protein interactions.

2.1.1 Tandem Affinity Purification

Tandem affinity purification (TAP) was developed as a method to purify protein complexes expressed at physiological levels under normal conditions [17]. This method relies on the use of two tags, as the name implies. In their original paper, Rigaut et al. (1999) tested a number of tags including FLAG tag, two IgG-binding units of protein A of *Staphylococcus aureus* (ProtA), the Strep tag, the His-tag, the calmodulin-binding peptide (CBP) and the chitin-binding domain (CBD) [17]. Although none of the tags interfered with the protein function, ProtA and the CBP gave the highest recovery efficiency [17]. The two tags are spaced by a tobacco etch virus (TEV) protease recognition site [17]. Gavin et al. (2002) used the TAP method coupled to MS to identify the interaction partners of 589 proteins. This study resulted in the identification of 232 multi-protein complexes [6]. The quest to identify the rest of the yeast interactome using the TAP purification techniques was achieved by two other studies that looked at all the yeast 6466 ORFs [18, 19]. In these studies, proteins of interest were fused to TAP tag by homologous recombination. This process allows the expression of these proteins under the control of their endogenous promoters offering physiological levels of tagged protein expression. Cellular lysates containing the tagged protein were applied to IgG-sepharose where the tagged protein binds to the IgG-sepharose through its ProtA tag. The tagged protein along with its binding partners were washed to reduce the level of contamination. To further reduce the level of contamination the immobilized protein complex was incubated with TEV protease to release the protein of interest as well as its interactors. Then a second purification step is performed using the calmodulin-sepharose, which binds to the CBP tag on the protein of interest in the presence of calcium. After washing, the protein complex was eluted with EGTA [17]. Eluted protein complexes can then be resolved using different methods. In the case of Gavin et al. (2002) it was a 1-D SDS-PAGE gel. The gel was then stained and bands of interest were proteolyticly digested and analyzed by MALDI-MS [6]. Ionization and generation of gas-phase molecules in MALDI is facilitated by a matrix that is mixed at a high ratio with the sample being analyzed (analyte) [9]. The matrix is usually made of small organic molecules. To generate protonated gas-phase molecules the matrix is mixed with the analyte at a high ratio and spotted onto a metal substrate. The dried crystals are irradiated by a laser beam that will ionize the matrix [10]. Ionization of the analyte is believed to occur through the matrix as it transfers part of its charge to the analyte. At the same time the matrix also offers protection to the analyte from the disruptive energy of the laser beam. In general, most ions generated by MALDI ionization are singly charged ions, but multiply charged ions can also be observed [10]. Figure 2 gives an overview of protein identification by affinity purification coupled to MALDI ionization.

The use of yeast as a model system offers the advantages of using homologues recombination to tag the protein of interest and have it expressed under the control of its own promoter [20]. This process has the advantage of eliminating untagged proteins from the cell as well as expressing the protein at its normal level. Different methods have been used in mammalian systems to overcome the problem of not having homologues recombination. These include the use of transient transfection, stable cell lines and the use of inducible promoters [20]. Due to these and other problems such as sample size that face researchers working on mammalian cells, identification of protein–protein interaction has been more limited and done on smaller scales. One of the early large-scale studies looks at positive TNF-alpha/NF-kappa B sig-

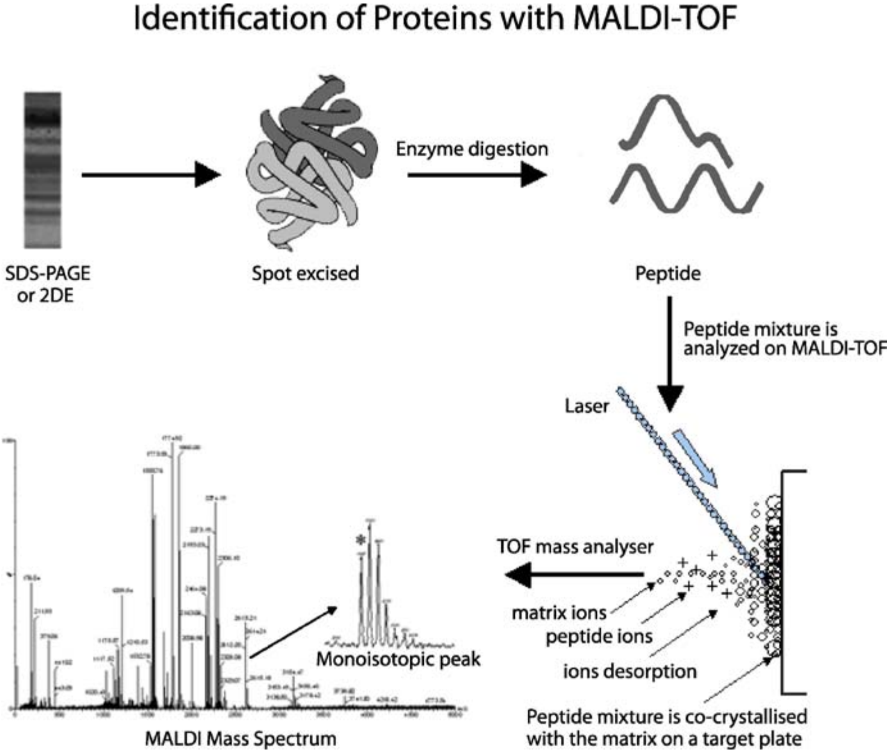

Identification of Proteins with MALDI-TOF
SDS-PAGE or 2DE
Spot excised
Enzyme digestion
Peptide
Peptide mixture is analyzed on MALDI-TOF
Laser
TOF mass analyser
matrix ions
peptide ions
ions desorption
Peptide mixture is co-crystallised with the matrix on a target plate
Monoisotopic peak
MALDI Mass Spectrum

Fig. 2 Protein identification by MALDI ionization coupled to MS

nal transduction pathway. This study uses the TAP tag approach to look at the interaction of 32 known and candidate NF-alpha/ NF-kappa B pathway components [21]. TNF-alpha-responsive HEK293 cells were stably transfected with the different tagged proteins. Protein complexes were purified from the non-induced and TNF-alpha-induced cells [21]. This study had the advantage of identifying protein interactions under different conditions.

2.1.2 One Tag Immunopurification, a Special Look at FLAG Tag

The large-scale analysis of yeast protein interaction in 2002 by Gavin et al. was also paralleled by another study performed by Ho et al. (2002) [22]. This study looked at the interaction of 725 proteins detecting 3,617 interactions between 1,578 unique proteins covering about 25% of the yeast genome. IPs were performed using anti-FLAG antibodies. Briefly, cellular lysates from cells transfected with the FLAG tagged protein of interest were immunopurified using anti-FLAG-antibodies conjugated to sepharose beads. Protein complexes were then eluted and analyzed by 1-D SDS-PAGE and stained with colloidal Coomassie stain. Bands were then excised from polyacrylamide gels, reduced and S-alkylated, and then subjected to trypsin hydrolysis. Digested peptides were then analyzed by electrospray ionization (ESI) coupled to LC-MS/MS. ESI is another soft ionization method that is used to generate gas-phase protonated molecules [8, 23]. In this process, analyte is dissolved at low concentration in a volatile solvent. The solvent containing the analyte is pumped through a hypodermic needle at a low flow rate and a high voltage to electrostatically disperse, or electrospray, small, micrometer-sized droplets. These droplets rapidly evaporate imparting their charge onto the analyte molecules. Electrospray ionization occurs under atmospheric pressure preserving the structure of the sample being analyzed [23]. One method to stabilize the spray is to use nebulizer gas. Molecules are then transferred into MS with high efficiency for analysis [10]. An overview of protein identification by affinity purification coupled to electrospray ionization is shown in Fig. 3.

FLAG immunopurification is a more simple but robust technique for identifying protein interactors. It also has the advantage of being a small hy-

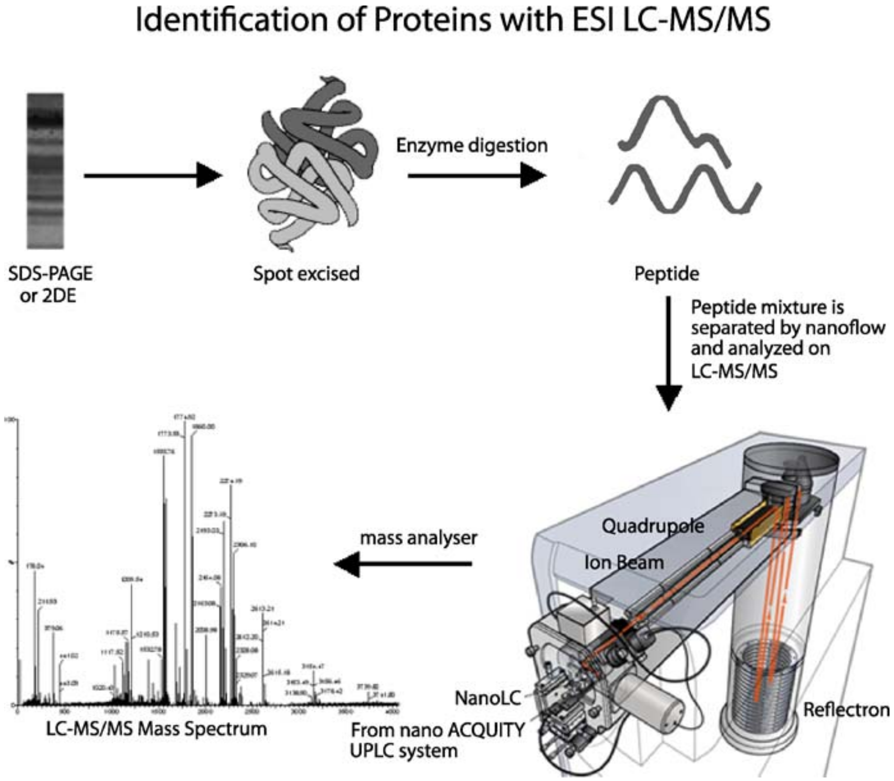

Identification of Proteins with ESI LC-MS/MS
SDS-PAGE or 2DE
Spot excised
Enzyme digestion
Peptide
Peptide mixture is separated by nanoflow and analyzed on LC-MS/MS
Quadrupole
Ion Beam
Reflectron
NanoLC
From nano ACQUITY UPLC system
mass analyser
LC-MS/MS Mass Spectrum

Fig. 3 Protein identification by ESI ionization coupled to MS

drophilic peptide (~1 kDa) compared to the large TAP tag (original TAP is ~20 kDa). It was reported that 18% of C-terminus TAP tagged essential yeast proteins gave rise to non-viable strains [6]. This high percentage of non-viable strains shows the great advantage FLAG has over the TAP tag and the need to use smaller size TAP tags. On the other hand, FLAG-IP suffers from a higher false-positive protein-interaction identification rate compared to the TAP tag [24]. The high false-positive rate in protein–protein interaction studies creates the need for cross-validation of reported interactions. A new large-scale study of 338 human protein–protein interactions was performed by Ewing et al. (2007). This study used the FLAG-IP system coupled to ESI-LC-MS/MS to look at protein interaction in HEK293 cells (see Fig. 3 for an outline of this process). Analysis of these protein interactions resulted in the identification of 24 540 potential protein interactions that was further validated to generate 6,463 interactions between 2,235 unique proteins. Data set generated using this method was validated using different methods generating high confidence rate in the quality of this data [25].

In other attempts to study protein–protein interactions in lower organisms, Arifuzzaman et al. (2006) performed a large-scale pull-down study using 4,339 His-tagged Escherichia coli ORFs. Unlike FLAG-IP, which utilizes anti-FLAG antibodies conjugated to sepharose beads, the His-tag is purified on a nickel column [26]. Purified proteins were then identified using MALDI-TOF MS [26].

2.1.3 Phosphopeptides Purification

The dynamic nature of proteomes makes their analysis not a trivial issue. One of the methods for controlling proteins function is through post-translational modifications (PTMs). Addition of different PTMs can change the conformational structure of a protein leading to change in its interaction partners. One of these important PTMs that regulate many functions in the cell is phosphorylation [27–29]. Shulze et al. (2005) have performed a study using synthetic peptides of all the cytosolic ErbB-receptor family to identify proteins interacting with these peptides as a result of their phosphorylation [30]. In this study they analyzed 94 pairs of singly phospho-, non-phospho- and the doubly phosphopeptides of the 89 tyrosine residues of the cytosolic ErbB-receptor family. Peptides containing tyrosine residues were synthesized with a desthiobiotin tag and immobilized on streptavidin-coated magnetic beads. After incubation with cell lysates protein complexes were eluted by biotin and analyzed by MS. Identified interactors included many of the known interactions of the ErbB-receptor family. Their data also showed that the EGF receptor and ERBB4 played a more prominent role in signaling than ERBB2 and ERBB3 [31]. This study showed how peptide–protein interaction screens can be used on a large scale to gain a global understanding of whole protein family interactions [31].

3
Gel Free Mass Spectrometry

Proteomic analysis has been historically linked to two-dimensional electrophoresis (2DE). 2DE gels were used to resolve complex protein mixture and to visualize these proteins by different types of staining [32]. Protein identification of different spots appearing in the 2DE gels was the main challenge until the development of MS instruments that were capable of identifying the different spots [32]. Due to 2DE limitations and the need to create high-throughput techniques gel-free methods were developed. In this part we will focus on a technique called multidimensional protein identification technology (MudPIT) that has been used in identifying protein–protein interactions [33]. MudPIT is a gel-free peptide separation technique that employees two chromatographic separation steps prior to sample ionization and identification by MS. Normally the first chromatographic separation dimension is a strong cation exchange (SCX) column. After that, the sample is separated in a second dimension by reverse chromatography (RP). RP offers the advantage of being compatible with electrospray ionization and being efficient at desalting the sample mixture [33].

Sample preparation involves the denaturation and then reduction and alkylation. After that the sample is digested with the appropriate digestion enzyme. Finally, the samples are acidified before being loaded onto the SCX column. MudPIT has the advantage of separating the peptides according to their charge state in the first dimension and then according to their hydrophobicity in the second dimension [33]. One example of the use of MudPIT in identifying protein–protein interactions is a study performed by Graumann et al. (2004). In the study the authors studied the interactions of 21 proteins involved in transcription and progression through mitosis. Proteins of interest were TAP-tagged and then purified and analyzed by MudPIT-MS. Using this method the authors were able to identify 102 previously known and 279 potential physical interactions [34].

The use of gel-free techniques requires an in-solution digestion of the samples being analyzed. Different methods are used to digest proteins in solution such as immobilized trypsin in monolythic columns followed by peptide fractionation or protein separation coupled to immobilized trypsin [35]. A new in-solution digestion method developed by our group has been recently published. This method uses a single microfluidic device called the proteomic reactor to pre-concentrate, clean up, derivatize, and digest proteins [35]. Protein digestion using the proteomic reactor involves loading cell lysates at a low pH into SCX column. Under these conditions, most peptides will have a positive charge favoring their binding to the reactor material. Due to low pH, trypsin remains inactive. Trypsin is activated by increasing the pH leading to protein digestion. Peptides are then eluted using buffers compatible with MS analysis [35].

The development of different gel-free proteomic techniques gives researchers new tools to add to the proteomic toolkit to answer various biological questions.

4 Quality of Protein Interaction Data

Generation of large protein interaction data sets is a great resource to understand the functions of many previously uncharacterized proteins. Nonetheless, validation of these data sets is a very daunting task. As the availability of large-scale data sets increases, it gives researchers the chance to compare data generated by different methods. For example, comparison of data generated by Gavin et al. (2002), Ho et al. (2002) and the two other Y2H studies [36, 37] shows a very high false-positive rate as high as 80%. This high false-positive rate can be due to many reasons including the techniques. For example, as we mentioned earlier Y2H suffers from major issues that increase its false rate. The use of a tag that is added to the protein of interest will affect the structure of the protein. This affect is clearer when a larger tag is added like the TAP tag. This issue can be reduced using smaller tags and alternating them between the N- and C-termini of protein achieving the least interference with the protein structure.

One of the methods used to validate protein–protein interaction is co-immunopurification. The high number of interactions in the large-scale studies makes the validation process near impossible. Alternatively, bioinformatic tools can be used to give higher confidence levels to large-scale data sets. For example, in the large-scale experiment performed by Ewing et al. (2007) the authors used various methods to ensure the quality of their data. Data generated from each IP was scored according to six different parameters [25]. This method generated a confidence score for each protein–protein interaction enabling the authors to judge its validity and setting a cut-off point to accept generated data. Using these criteria, the number of interactions was reduced from 6463 protein–protein interactions of which 2251 had high confidence scores [25]. Another example is the use of an unsupervised probabilistic scoring scheme developed by Hart et al. (2007) [38]. This approach consists of giving a confidence score to each interaction that was generated by the matrix method interpretation (technique used for the creation of the interactions data sets that will include all the prey–prey interactions from given bait pull-down). This method not only increases recall and/or precision over other methods like the standard spoke model interpretation (only bait–prey interactions) but can be used to integrate data sets from other sources. The authors used this scoring scheme to combine the data generated by Gavin et al. (2002), Krogan et al. (2006) and Ho et al. (2002). The results show that the scoring metric is more accurate than the filtering schemes used by the other groups.

One main area in which the study of protein–protein interaction can be improved is the creation of standard experimental guidelines. This should include explicit information about the origin of samples and how they were analyzed. One initiative toward setting these standards is the Proteomics Standards Initiative (PSI) which is aimed at developing guidelines for various proteomics workflows that will help facilitate data comparison, exchange and verification [39, 40]. These documents are collectively known as the "minimum information about a proteomics experiment" (MIAPE) guidelines [41]. One of these MIAPE modules is the MIMIx ("minimum information about a molecular interaction experiment") [41, 42]. Theses guidelines are aimed at giving the user the ability to assess the quality of the presented data and to find interactions of their protein(s) of interest and then to access the original publications for the complete experimental design [42]. Although, not fully able to meet the needs for standardizing protein interaction data, these guidelines offer a road map for future more complete guidelines.

5
Protein Interaction Databases

Generation of large protein interaction data sets created the need to make a public database of all the interactions. A number of databases collect data from various types of protein–protein interaction experiments were launched. Examples of these databases include BIND [43], DIP [44], IntAct [45], MINT [46], MPact (MIPS) [47], BioGRID [48] and HPRD [49]. Initially these databases were functioning in isolation with no common extraction, curation and storage protocols and not all of them explored the same scientific papers, therefore making data sharing a difficult task. In fact, we observed that only small section of those databases were overlapping and therefore in many cases the information was more complementary and can be unified to increase and improve our knowledge about interactome networks. In recent years, a group of protein interaction databases launched a special project calling for a community-standard for the representation of protein interaction data [50]. This model was developed by members of the Molecular Interaction (MI), a subgroup of the PSI. Major interaction databases have already adopted this work model. Using these standards as a road map, a group of databases have jointly formed the International Molecular Interaction Exchange consortium (IMEx) [42]. IMEx has begun sharing the curation load and aims to interchange data curated [42].

6 Conclusions

In conclusion, the use of MS in protein identification has tremendously advanced the field of proteomics. The development of new types of MS machines and new affinity purification approaches will further advance this field. More rigorous experimental design and data analysis will decrease the rate of false positive and negative rates. Finally, the stage is now set to move forward for identification of the human interactome. Ultimately, researchers will be able to compare the interactome of different disease states or cells treated with different cues and look into changes in protein interactions under different conditions at a large scale.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) Nature 422:835
2. de Hoog CL, Mann M (2004) Annu Rev Genomics Hum Genet 5:267
3. Lopez JL (2007) J Chromatogr B Analyt Technol Biomed Life Sci 849:190
4. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Nat Biotechnol 17:994
5. Ong SE, Foster LJ, Mann M (2003) Methods 29:124
6. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, 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 MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Nature 415:141
7. Figeys D (2002) Curr Opin Mol Ther 4:210
8. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Science 246:64
9. Koichi T, Hiroaki W, Yutaka I, Satoshi A, Yoshikazu Y, Tamio Y, Matsuo T (1988) Rapid Commun Mass Spectrometry 2:151
10. Mann M, Hendrickson RC, Pandey A (2001) Annu Rev Biochem 70:437
11. Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E (2006) Nature 440:556
12. Chen PY, Meister G (2005) Biol Chem 386:1205
13. Jeffery CJ (1999) Trends Biochem Sci 24:8
14. Fields S, Song O (1989) Nature 340:245
15. Gietz RD, Triggs-Raine B, Robbins A, Graham KC, Woods RA (1997) Mol Cell Biochem 172:67
16. Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Mol Cell Proteomics 1:349
17. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) Nat Biotechnol 17:1030
18. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Nature 440:631

19. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Nature 440:637
20. Gingras AC, Aebersold R, Raught B (2005) J Physiol 563:11
21. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, Hopf C, Huhse B, Mangano R, Michon AM, Schirle M, Schlegl J, Schwab M, Stein MA, Bauer A, Casari G, Drewes G, Gavin AC, Jackson DB, Joberty G, Neubauer G, Rick J, Kuster B, Superti-Furga G (2004) Nat Cell Biol 6:97
22. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Nature 415:180
23. Wilm M, Shevchenko A, Houthaeve T, Breit S, Schweigerer L, Fotsis T, Mann M (1996) Nature 379:466
24. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Nature 417:399
25. Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, Taylor R, Dharsee M, Ho Y, Heilbut A, Moore L, Zhang S, Ornatsky O, Bukhman YV, Ethier M, Sheng Y, Vasilescu J, Abu-Farha M, Lambert JP, Duewel HS, Stewart II, Kuehl B, Hogue K, Colwill K, Gladwish K, Muskat B, Kinach R, Adams SL, Moran MF, Morin GB, Topaloglou T, Figeys D (2007) Mol Syst Biol 3:89
26. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A, Tsuzuki K, Nakamura S, Altaf-Ul-Amin M, Oshima T, Baba T, Yamamoto N, Kawamura T, Ioka-Nakamichi T, Kitagawa M, Tomita M, Kanaya S, Wada C, Mori H (2006) Genome Res 16:686
27. Karin M (1994) Curr Opin Cell Biol 6:415
28. Morley SJ (1994) Mol Biol Rep 19:221
29. van der Geer P, Hunter T, Lindberg RA (1994) Annu Rev Cell Biol 10:251
30. Schulze WX, Deng L, Mann M (2005) Mol Syst Biol 1:1
31. Mann M (2006) Nat Rev Mol Cell Biol 7:952
32. Roe MR, Griffin TJ (2006) Proteomics 6:4678
33. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd (1999) Nat Biotechnol 17:676
34. Graumann J, Dunipace LA, Seol JH, McDonald WH, Yates JR 3rd, Wold BJ, Deshaies RJ (2004) Mol Cell Proteomics 3:226
35. Ethier M, Hou W, Duewel HS, Figeys D (2006) J Proteome Res 5:2754
36. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) Nature 403:623

37. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) Proc Natl Acad Sci USA 98:4569
38. Hart GT, Lee I, Marcotte ER (2007) BMC Bioinformatics 8:236
39. Taylor CF, Hermjakob H, Julian RK Jr, Garavelli JS, Aebersold R, Apweiler R (2006) Omics 10:145
40. Hermjakob H (2006) Proteomics 6(2):34
41. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR 3rd, Hermjakob H (2007) Nat Biotechnol 25:887
42. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-Aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J, Bader J, Uetz P, Tyers M, Legrain P, Fields S, Mulder N, Gilson M, Niepmann M, Burgoon L, Rivas Jde L, Prieto C, Perreau VM, Hogue C, Mewes HW, Apweiler R, Xenarios I, Eisenberg D, Cesareni G, Hermjakob H (2007) Nat Biotechnol 25:894
43. Bader GD, Betel D, Hogue CW (2003) Nucleic Acids Res 31:248
44. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) Nucleic Acids Res 32:D449
45. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) Nucleic Acids Res 32:D452
46. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) FEBS Lett 513:135
47. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) Nucleic Acids Res 34:D436
48. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) Nucleic Acids Res 34:D535
49. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A (2003) Genome Res 13:2363
50. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R (2004) Nat Biotechnol 22:177

Adv Biochem Engin/Biotechnol (2008) 110: 81–114
DOI 10.1007/10_2008_098

Published online: 2 July 2008

Lab-on-a-chip in Vitro Compartmentalization Technologies for Protein Studies

Yonggang Zhu[1] (✉) · Barbara E. Power[2]

[1]Division of Materials Science and Engineering, CSIRO Australia, P.O. Box 56,
VIC 3190 Highett, Australia
yonggang.zhu@csiro.au

[2]AntibOZ Pty Ltd, VIC 3130 Blackburn, Australia

Abstract In vitro compartmentalization (IVC) is a powerful tool for studying protein–protein reactions, due to its high capacity and the versatility of droplet technologies. IVC bridges the gap between chemistry and biology as it enables the incorporation of unnatural amino acids with modifications into biological systems, through protein transcription and translation reactions, in a cell-like microdrop environment. The quest for the ultimate chip for protein studies using IVC is the drive for the development of various microfluidic droplet technologies to enable these unusual biochemical reactions to occur. These techniques have been shown to generate precise microdrops with a controlled size.

Various chemical and physical phenomena have been utilized for on-chip manipulation to allow the droplets to be generated, fused, and split. Coupled with detection techniques, droplets can be sorted and selected. These capabilities allow directed protein evolution to be carried out on a microchip. With further technological development of the detection module, factors such as addressable storage, transport and interfacing technologies, could be integrated and thus provide platforms for protein studies with high efficiency and accuracy that conventional laboratories cannot achieve.

Keywords In vitro compartmentalization · Lab-on-a-chip · Microdrop · Microfluidics · Protein evolution · Protein studies

1 Introduction

In vitro compartmentalization (IVC) refers to cell-like compartments generated artificially as reaction chambers in which protein transcription and translation reactions can occur. In biology, cell walls confine networks of chemical reactions so that they can proceed in isolation from the rest of the environment, but toxic proteins are unable to be produced in living systems. Biological degradation systems have the advantage that they remove misfolded proteins from the environment. In IVC, chemically modified amino acids can be incorporated into proteins, expanding the number of variations now available [1] so that different or toxic proteins or enzymes can now be produced that were previously not possible in biological protein expression systems. Tawfik and Griffiths [2] described a system using micron-size aqueous droplets dispersed in an oil medium in which individual gene sequences were proximal to the enzyme variant they encoded. The emulsion droplets in IVC systems may range in size from a few to a few tens of micrometers. The volume of the droplets is (10^4 to $\sim 10^{10}$ times) smaller than that used in a typical transcription/translation reaction (20 μL). Such small droplets ($\sim$ 500 million 10 μm drops per milliliter of sample) provide the opportunity to physically contain one copy of DNA, in an environment containing the components for transcription and translation into protein, so that the synthesized protein is in the same droplet as the DNA that it encodes. In conjunction with versatile controls, these droplets are an ideal means of compartmentalizing biochemical and genetic assays. These advantages have fueled the increasing effort in the development of droplet-based IVC systems in the last few years [3–11].

The advances in microfluidic technologies provide unique opportunities for IVC development. Instead of the conventional methods of preparing emulsions such as homogenizers, stirrers or extruding devices, which produce only polydisperse droplets, microfluidic chips can be used for microdrop formation and control. It has been demonstrated that monodispersed droplets/slugs can be formed and manipulated reliably on-chip for various protein and cell analysis applications. Although the development of microfluidic IVC technology is recent, it has attracted a great deal of attention. A number of review articles have been published in droplet-based microfluidic devices, e.g., [9, 12–17].

This chapter reviews the current development of lab-on-a-chip technologies that may be applicable for IVC. Initially we give a brief introduction to protein production and directed evolution, then summarize the applications of IVC in protein synthesis. Then we review the current lab-on-a-chip technologies, which include microdrop formation in a microfluidic chip and its manipulation, sorting, and detection technologies. Finally, future directions in microfluidic technologies will be discussed.

2
IVC for Protein Synthesis

Protein synthesis is accomplished within the droplet, which contains diluted DNA so that each droplet contains approximately one gene copy of DNA. The aqueous phase of the droplet contains all of the components for transcription. These could also be polymerase chain reaction (PCR) reagents enabling the amplification of the initial single copy of DNA. The heating, cooling, and extension steps can also be done within the droplet.

After the DNA has been amplified within the droplet, another droplet can be merged to the first drop. The second droplet may contain all of the translation reagents required to synthesize protein, including unnatural amino acids conferring new properties to the newly synthesized protein.

The new protein droplet can now be split into two drops. One drop can be used for protein detection of a certain property, which may require the addition of another reagent to the droplet; the droplet is discarded after detection. The second droplet is sorted according to the detection result of the paired drop. If the droplet is retained due to desirable properties, then the DNA molecules can be isolated from the droplet and used for further rounds of modification or protein evolution.

3
IVC for Protein Evolution

3.1
Directed Evolution

Directed evolution is a method used in protein engineering to evolve proteins or RNA with desirable properties for use in agricultural, medical, and industrial applications [18–25]. The protein evolution experiments typically involve the following main steps [18, 21, 26]: a library of mutant genes is generated first through random or targeted mutagenesis and/or gene recombination of target genes using techniques such as error-prone PCR and DNA shuffling. The DNA is translated into protein. The presence of mutants (variants) in the library is screened, then selected for the desired property. The variants identified are further analyzed by DNA sequencing in order to understand what mutations have occurred. The process is repeated using these functionally improved proteins as the templates until the goal is achieved or no further improvement is detectable.

Directed evolution can be performed with or without living cells (in vivo or in vitro evolution). In vivo evolution is preferred when the evolved protein or RNA is to be used in living organisms, whereas in vitro evolution has the advantages of generating larger libraries and utilization of more versatile selection techniques.

3.2 Directed Evolution in Emulsions

While the expression of genes is usually carried out in a host cell, cell-free or in vitro expression (such as IVC systems using water-in-oil emulsions) has been gaining popularity recently. The main advantage of using IVC for protein evolution is that a combination of methods can be used to introduce variation into the selection pool. The gene library can have engineered changes in the DNA and then another layer of variation can be introduced at the protein level, where the amino acids used for protein synthesis can also contain unusual properties. A few protocols using emulsion IVC systems have been published recently in *Nature Methods* [10, 27, 28]. An example of the IVC cycle for enzyme evolution is shown in Fig. 1 [11]. The enzyme-encoding genes and the corresponding in vitro transcription and translation reagents are emulsified as water droplets in an oil phase. The gene concentration in the droplets is kept very low such that each droplet contains statistically one copy of the mutant gene. The active enzyme encoded by the gene modifies its own DNA by substrate turnover. This prevents the DNA from being digested after breaking the emulsion. Enriched genes are recovered by PCR for molecular evolution and further characterization.

Currently, the IVC system is used for the selection of proteins, DNA and RNA enzymes, enzyme inhibitors and so on. For example, such a system has been used for the selection of peptide ligands [29–31], selection of restriction endonucleases using *Fok*I as a model enzyme [32], protein synthesis through solubilizate exchange between droplets [33], and selection of ribozymes that catalyze multiple-turnover Diels–Alder cycloadditions [34]. It has also been applied for the directed evolution of DNA methyltransferases [35], bacterial phosphotriesterase [36], Taq polymerase [37], protein inhibitors of DNA-nucleases [38], and for molecular evolution of catalytic

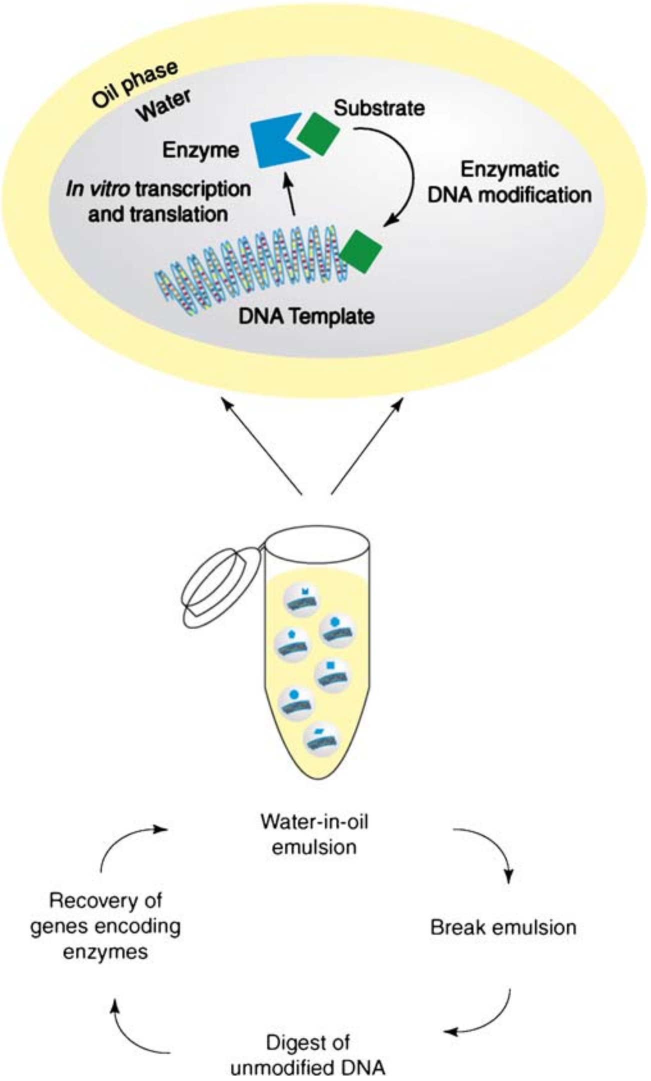


Fig. 1 Schematic evolution cycle for a DNA-modifying enzyme using emulsion IVC. Reprinted from [11] with permission from Elsevier, copyright 2006

proteins from large libraries [39]. Traditionally, protein translation has been achieved with systems such as eukaryotic wheat germ extract [31, 40, 41] and rabbit reticulocyte systems [23] to provide post-translational modification of the synthesized protein, but these systems have proved difficult in the IVC system. In a cell-free environment, IVC systems have used bacterial S30 extracts and purified components for protein production (PURE system) [42].

4 Lab-on-a-Chip Technologies

All of the above examples used emulsion droplets prepared using conventional techniques. The size distribution of the droplets was usually large, sometimes with several orders of magnitude difference! With such a large size variation, reaction, detection, and sorting of the droplets could not be well controlled, thus limiting the applications. With the advance of microfabrication and microfluidic technologies, researchers have developed lab-on-a-chip devices to improve droplet formation and control functionalities. These technologies could allow protein synthesis and directed evolution to be carried out on-chip, providing a high efficiency and accuracy that the current laboratory protocols cannot achieve.

4.1 What is Required for the Ultimate IVC Chip?

The main functionalities of a droplet chip depend on the desired tasks to be performed on the chip. For protein synthesis and enzyme evolution, a typical chip would at least consist of these main components, i.e., (1) droplet formation, (2) drop fusion, (3) detection, and (4) sorting. Figure 2 shows a flow chart of a typical IVC chip and its possible functionalities. The droplet formation compartmentalizes reagents into single droplets, which may contain genes and in vitro transcription and translation solutions. Drop fusion is needed for bringing multi-reagents into single droplets for reaction or for adding substrate solutions. After fusion, an incubation step may be needed to provide an environment for the biochemical reactions to occur. When the reaction is completed or at the required stage, the droplets need to be examined to see if proteins of interest were indeed synthesized. This can be achieved by, for example, laser-induced fluorescence techniques. The droplets with the desired synthesized protein react with the substrate to emit light. Such fluo-

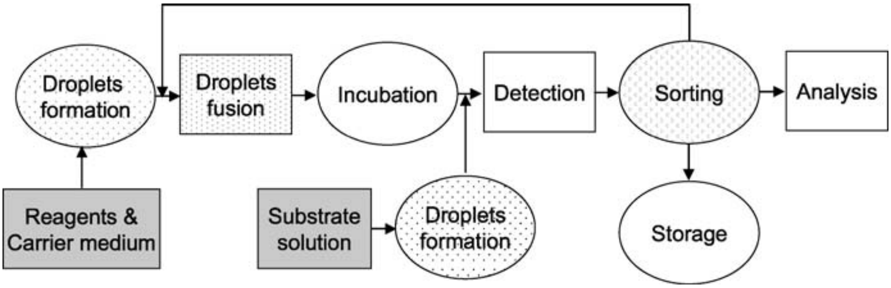


Fig. 2 Typical flow chart of an IVC chip for enzyme evolution

rescence signals can be used to sort out droplets of interest for analysis and for the next round of evolution, if required.

Other steps may be necessary, such as on-chip PCR, droplet storage, or interfacing with analytical instruments. If the protein concentration in the droplets is low due to the limited number of genes present in each droplet (preferably one gene per droplet), the detection techniques need to be very sensitive, ideally at single molecule level.

Another approach to circumvent this problem is to amplify the DNA molecules in the droplets. Initially each droplet contains preferably one copy of DNA, which can be amplified (by PCR) to increase the number of copies of DNA within the droplet. The advantage of an increased number of (template) DNA copies is that higher amounts of protein can be produced in a short time frame. If the concentration of the protein produced is high enough, then less sensitive detectors are required. The PCR steps can also be achieved on-chip. After the sorting step, droplets may be stored for in-situ monitoring and future analysis or directed to off-chip analytical instruments such as mass spectrometers. On-chip storage chambers are needed, which may enable each droplet to be addressable. For off-chip analysis the interfacing function needs to be realized, such as mass spectrospray techniques. To form an evolution cycle, the proteins in the droplets may have to be purified and re-injected into the droplet stream for the subsequent round of evolution.

The following sections review the current developments in various aspects of the droplet-based microfluidic devices that have been used or have the potential to be used for IVC applications.

4.2 Microdrop Formation

4.2.1 Basic Principle of Droplet Generation

Microdrop generation has been investigated extensively for industrial applications such as ink-jet printing. The fundamental issue is how to effectively break out the jet of one liquid into another immiscible liquid. In the monograph by Lee [43], various techniques for droplet formation were reviewed, including thermal, acoustic, and electrical methods. These techniques are capable of generating picoliter volume droplets in air at mega-Herz speed. For IVC applications, the frequency required is much lower due to the time required for reaction and processing. Therefore, fluid dynamic forces have mainly been utilized for drop formation, with assistance in some cases by other forces such as electricity or centrifugal forces. The most commonly used methods utilize a network of microchannels of either a T-junction or cross-junction geometry (Fig. 3) to achieve the jet break-up.

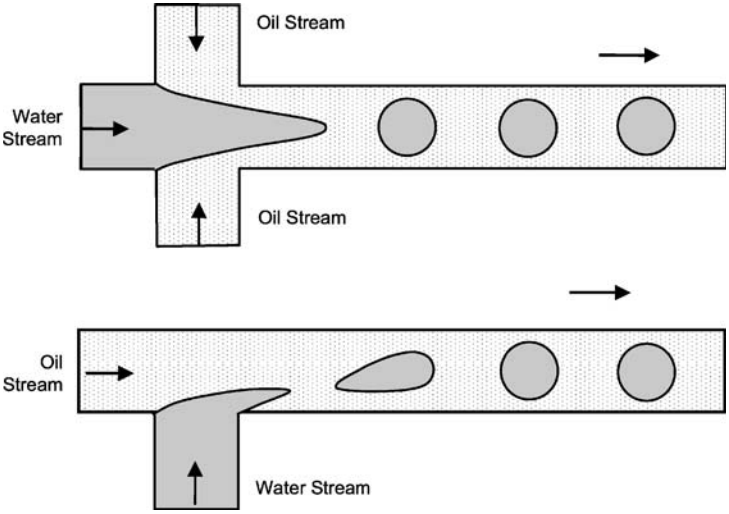


Fig. 3 Geometries of microchannel networks for droplet formation. *Top* cross-channel geometry, *bottom* T-junction geometry

In the T-junction geometry, two immiscible liquids are introduced from the two inlets, respectively, to create dispersed and continuous phase flows. The first phase (e.g. water) is forced through the microchannel junction into another channel where the second phase (e.g. oil) breaks the first phase to form microdroplets. The droplets are formed due mainly to the shear force from the mainline and other forces like surface tension, buoyancy, momentum, and inertia of both phases.

In the cross-junction geometry, the first phase is introduced from the left microchannel while the second immiscible liquid flows from two side channels. The two liquids are then forced to flow through a nozzle into the main channel. The outer fluid pushes the inner fluid into a narrow thread, which then breaks into droplets at the nozzle or downstream of the nozzle.

4.2.2 Development of On-Chip Droplet Generation Techniques

4.2.2.1 T-Junction Geometry

The T-junction geometry has been used previously in numerous studies to produce monodisperse droplets (see for example, [44–54]). The first on-chip application of drop generation using such geometry was carried out by Thorsen et al. [55]. Figure 4 shows the channel geometry and a picture of droplet formation from the T-junction. The droplet size could be controlled in

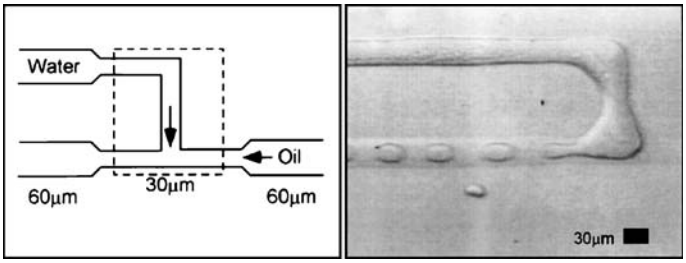


Fig. 4 Droplet formation by a T-junction geometry. Schematic view of the chip design (*left*) and an image of droplet formations (*right*). Reprinted from [55] with permission, copyright 2001, American Physical Society

the range of about 10 to 40 µm with the change of oil and water flow rates. The T-junction geometry can also be used to form monodisperse double emulsions in a microfluidic chip (Fig. 5) [47]. Such geometry could be used in IVC systems that allow interfacing with existing fluorescence-activated cell sorting (FACS) machines.

The T-junction geometry has been used as on-chip reaction platforms. For example, microfluidic chips have been developed for the formation of arrayed droplets for protein crystallization analysis [47, 56, 57]. An example of the chip is shown in Fig. 6. Just before the drop formation, multiple streams of aqueous solutions were brought together so that the formed droplets could contain all reactants. A similar geometry was also used in the study for the formation of alginate gel drops [58].

Recently, the T-junction format has been used for generating droplets for cell analysis [59, 60]. For example, Fig. 7 shows a polydimethylsiloxane (PDMS) device for producing uniformly sized and spaced aqueous droplets, which contain cells of various loading conditions. Cells with expressed fluorescent proteins are distinguished by a vertical spike in the fluorescence signal above the weaker background signal of the aqueous droplets.

4.2.2.2
Cross-Junction Geometry

This type of drop formation is a classical flow mechanics problem that has been studied extensively over the last few decades, mainly in the form of a nozzle in a quiescent or slow moving flow, e.g. [61–71], and has attracted more attention recently for microfluidic applications, e.g. [48, 55, 59, 72–88]. The simplest design of cross-junction geometry is just two straight channels, with constant cross-section areas, crossing each other [89].

Figure 8 shows the flow-focusing geometry implemented in a microfluidic device made of PDMS [72]. A range of droplet sizes were produced from the chip, varying from much bigger to much smaller than the orifice radius. Al-

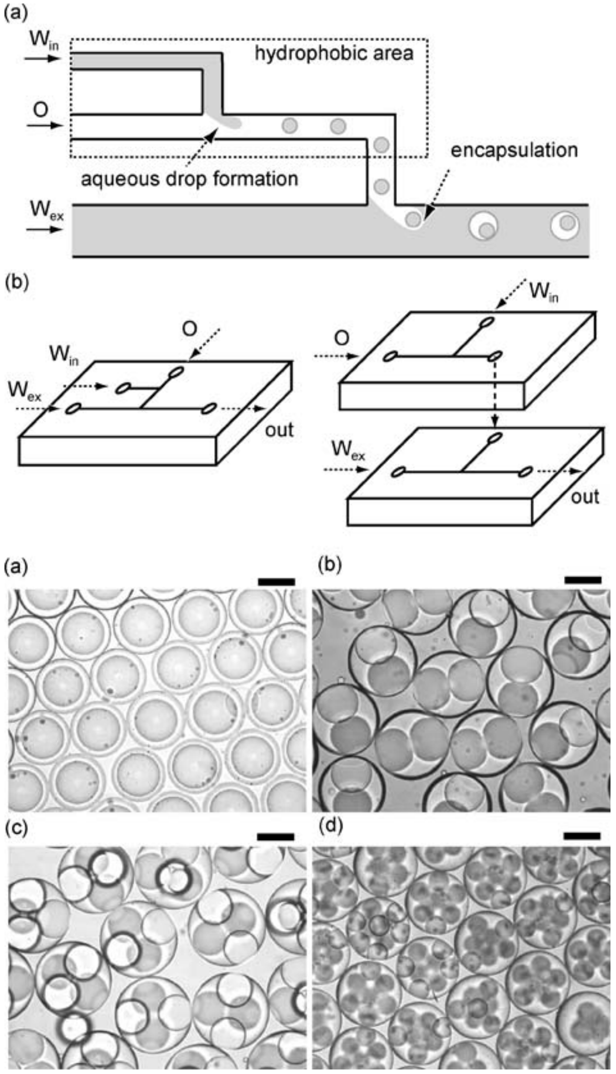


Fig. 5 Double emulsion formation using the T-junction geometries. *Above*: schematic views of the chip designs. *Below*: **a**–**d** images of emulsions produced. From [47], reproduced by permission of the Royal Society of Chemistry

though droplet formation was flow-dependent, it was found that there was a range of flow conditions where drops with diameters comparable to the orifice width were formed independent of the flow rates (see Figs. 8a–d).

Figure 9 shows another design [90] where, at the cross-junction, two water phases were sheared by the silicon oil stream and monodispersed droplets were generated alternatively and synchronously. Generated droplets were

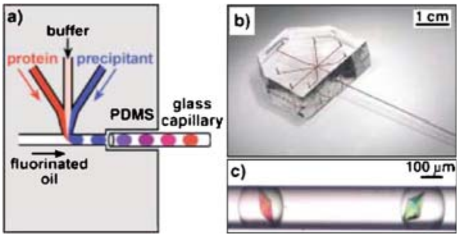


Fig. 6 Protein crystallization chip using the T-junction drop formation geometry (**a**). Pictures of crystals formed inside the drops (**b**),(**c**). Reproduced from [168] with permission, copyright 2004, Wiley-VCH

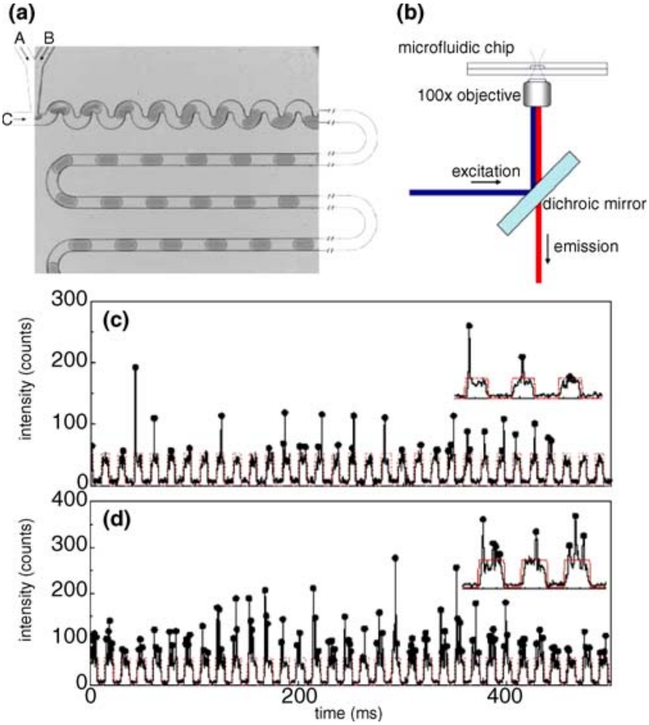


Fig. 7 PDMS device for producing uniformly sized and spaced aqueous droplets for single cell analysis, and the fluorescence signals detected from the droplets. **a** Optical image of the PDMS chip device with 50 μm square channel. **b** Schematic of the laser-induced fluorescence optical setup. **c**,**d** Optical readout of the fluorescence signals under two different cell loading conditions. From [59], reproduced by permission of the Royal Society of Chemistry

aligned in the tapered chamber before fusion and exiting to the outlet channel. The system was designed to create an alternation in droplet generation

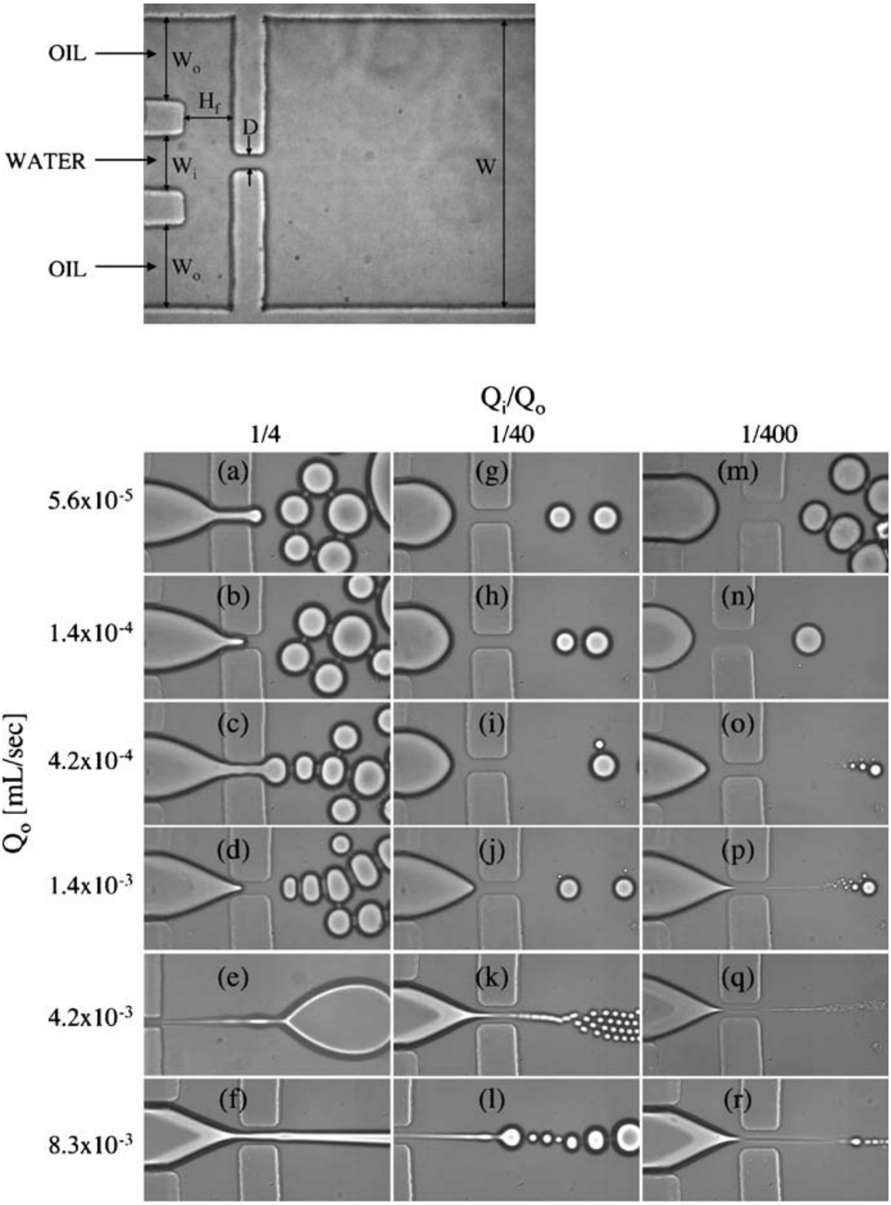


Fig. 8 *Above*: microfluidic device for drop formation. *Below*: **a**–**r** images of drops formed under different flow conditions. Reprinted with permission from [72]. copyright 2003, American Institute of Physics

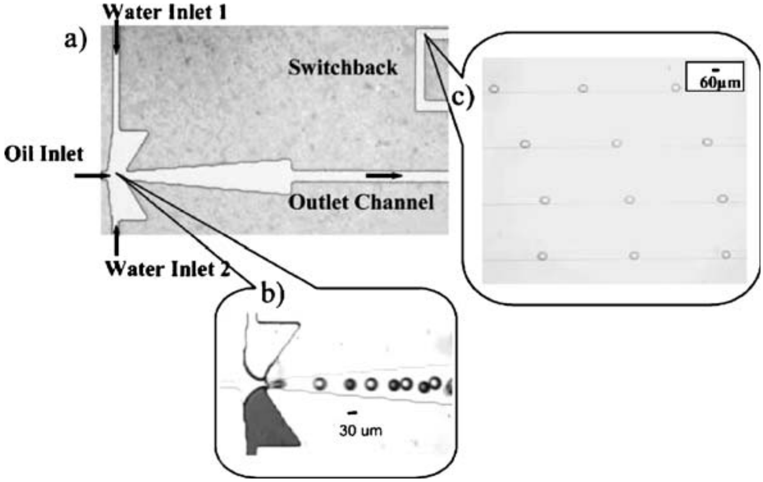


Fig. 9 PDMS chip with modified cross-section for producing oil droplets in water. **a** PDMS chip layout, **b** drop formation at the nozzle, and **c** droplet movement along the channel. From [90], reproduced by permission of the Royal Society of Chemistry

as well as controlled droplet fusion. CdS nanoparticle synthesis was demonstrated by this device.

The smallest droplets formed on-chip without active control were demonstrated in [91]. The water stream was hydrodynamically focused into microthreads before break-up to produce droplets with sizes ranging from 9 to

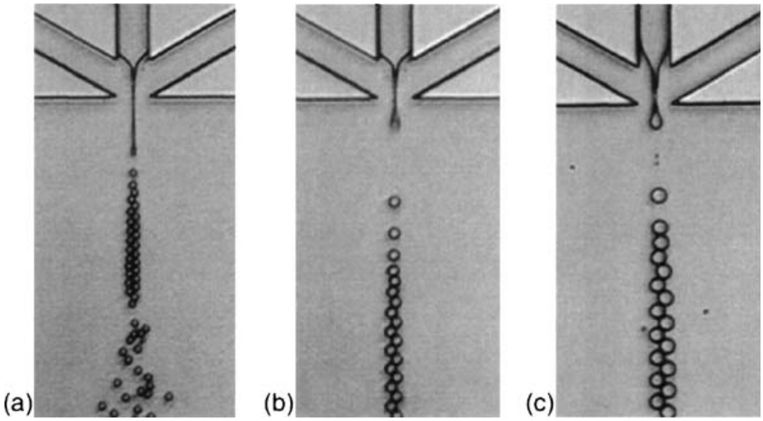


Fig. 10 Focused microthread device for precise drop formation control. The nominal droplet size, focusing flow rate from side channels, and water flow rate from the center channel were: **a** 9 µm, 0.45 mL h^{-1}, 0.002 mL h^{-1}; **b** 12 µm, 0.49 mL h^{-1}, 0.002 mL h^{-1}; and **c** 16 µm 0.15 mL h^{-1}, 0.001 mL h^{-1}, respectively. The channel size was 5 µm for all three cases. Reprinted with permission from [91], copyright 2004, American Institute of Physics

16 µm (Fig. 10). However, to achieve this, the channel depth has to be very small, ~ 5 µm. Some researchers also tried to develop 3D structures for drop formation [81, 92].

4.2.3 Formation Mechanism Studies

While droplet formation has been successfully realized in microchannels and various devices have been developed, little information is available for predicting droplet size and generation rate for a given flow condition and channel geometry. A vast volume of studies have been carried out so far on drop and bubble formation under quiescent conditions in macrochannel systems. Only a few studies have been carried out so far for microchannel systems, e.g., [14, 44, 48, 51, 55, 93–97]. In spite of the length-scale difference, the droplet generation mechanism at microscales is still based on the fundamental fluid dynamics principles observed at macroscales, i.e., the instability of a jet of liquid issued into another, due to the effects of surface tension and viscous forces, causes the jet to break into droplets (see the pioneering work of Lord Rayleigh [98] and the subsequent work by, e.g., Taylor [67], Tomotika [68]), albeit that the effect of liquid–boundary interaction is more pronounced at microscales. Quantitative models have been proposed for predicting droplet sizes for both Newtonian and non-Newtonian flows [14, 55, 94]. For example, it has been postulated that the shear force dominates the formation [14, 55, 67, 68] and therefore the droplet size d follows:

$$d \propto \frac{\sigma}{\mu\dot{\varepsilon}} \propto \frac{\sigma D^3}{\mu Q}\,, \tag{1}$$

where $\dot{\varepsilon}$ is the shear rate, D is the nozzle width and Q is the carrier phase flow rate. While all previous data show a trend that is consistent with the above equation (i.e., drop size decreasing with increasing flow rate of the continuous phase), e.g., [14, 78], the rate of change was much smaller than that implied by Eq. 1 as revealed by Zhu et al. [99] (Fig. 11). This is consistent with that proposed by [95], $d \propto Q^{0.25}$, for flow rate-controlled droplet generation.

The flow rate dependence of droplet size was also observed for T-junction geometry. Garstecki et al. (2006) [44] proposed a model derived from the postulated mechanism of break-up for two different viscosities of the continuous fluid and showed that the drop size was a linear function of the flow rate ratio between the dispersed and continuous phase, i.e., $d = 1 + \alpha Q_\mathrm{W}/Q_\mathrm{O}$, where Q_W and Q_O are water and oil phase flow rates, respectively (Fig. 12). However, such a relationship was not observed in the experiments by [100], where the capillary number $c_a \equiv \mu V/\sigma$ (where μ is the viscosity of fluid, σ is the interfacial tension and V is the characteristic flow velocity) was much larger than those in [44].

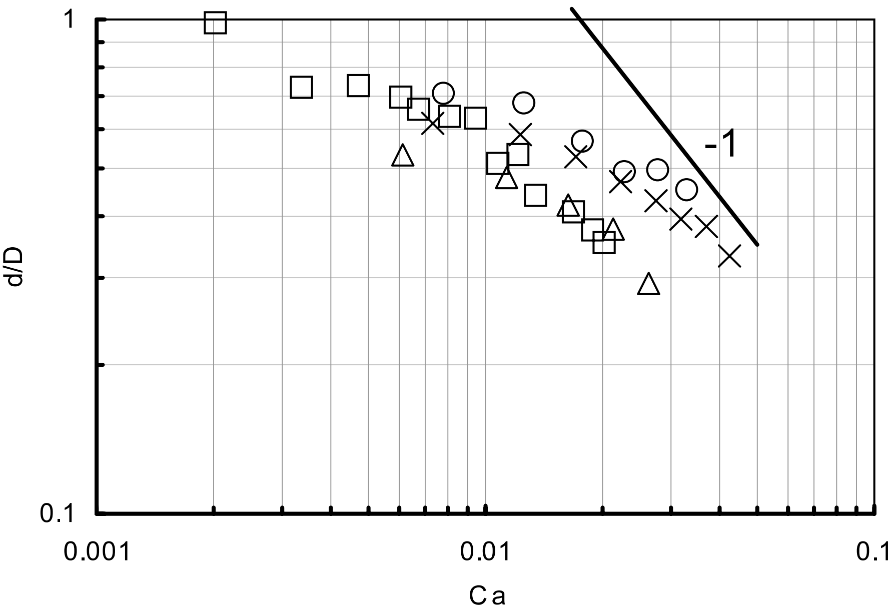


Fig. 11 Variation of droplet size as a function of capillary number. The capillary number is defined as $c_a \equiv \mu V/\sigma$, where μ is the viscosity of fluid, σ is the interfacial tension and V is the characteristic flow velocity. The *lines* indicate the slope implied from Eq. 1 and the magnitudes were arbitrarily adjusted for clarity of comparison. Measurements by Zhu et al. [99]: □, D = 100 μm. Measurement by Tan et al. [78]: Q_w = 0.3 (△), 0.5 (×), and 0.6 (○) μL/min, respectively; D = 48 μm

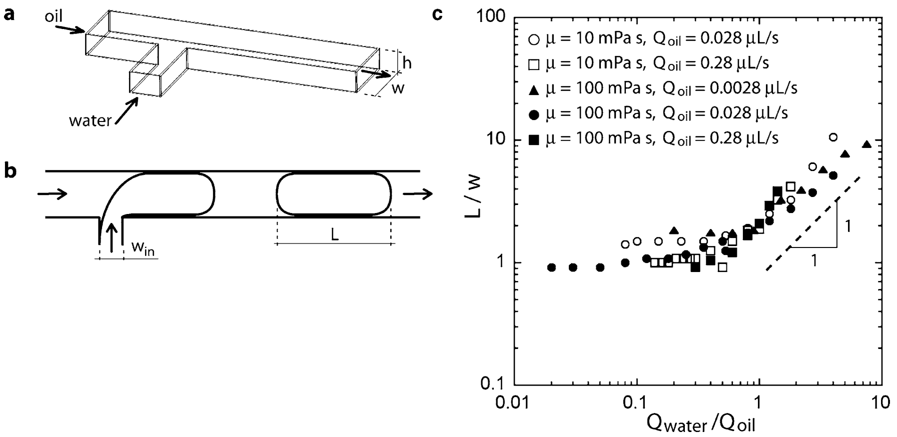


Fig. 12 Drop length as a function of flow rate and viscosity. **a** Schematic of microfluidic T-junction composed of rectangular channels. **b** Top view of the same schematic in a two-dimensional representation. **c** Dimensionless length of the droplets (L/w) plotted as a function of the ratio of the rates of the discontinuous (Q_{water}) and continuous (Q_{oil}) phases. From [44]: reproduced by permission of the Royal Society of Chemistry

4.3 Droplet Fusion

When a chemical or biological reaction is required, two or more reagents can be combined into one drop by bringing together streams of materials at the point where droplets are made. This strategy has been implemented in several droplet-based microfluidic systems [50, 56, 57, 59, 77, 101, 102]. However, if multiple steps are required and new reagents are added during each step, fusion of droplets is required. The two droplets streams may be produced independently from each nozzle. Droplet fusion can be achieved by simply bringing two streams of droplets together and allowing droplets to collide and coalesce. Figure 13 shows an example of a microfluidic chip for generating two streams of droplets [99]. The fusion occurred when the flow conditions were identical and material conditions were appropriate. However, when there are subtle changes in conditions such as flow conditions, surface tension, symmetry of chip design, mode of generation (i.e. if synchronous), drop size, chemical composition of the droplets and draining forces, the success rate of fusion may be reduced significantly. For example, when stabilizing surfactants are used, the fusion of droplets may not occur at all. Further, for droplets much smaller than the cross-section of the microchannel, they might

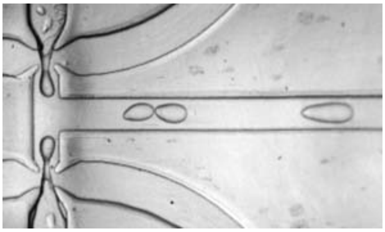


Fig. 13 Droplet chip with two streams of droplets generated and fused in the main channel

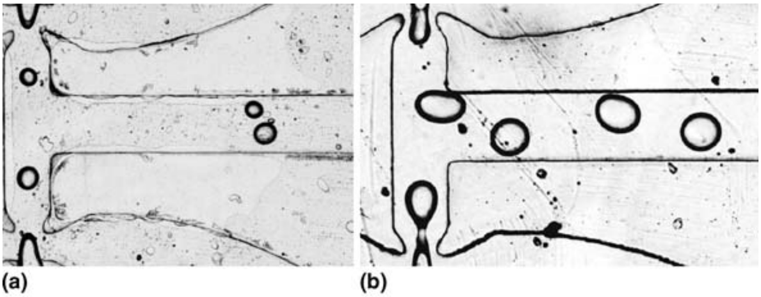


Fig. 14 Images of droplets generated from the two nozzles: **a** asynchronous generation, **b** synchronous generation. No fusion observed

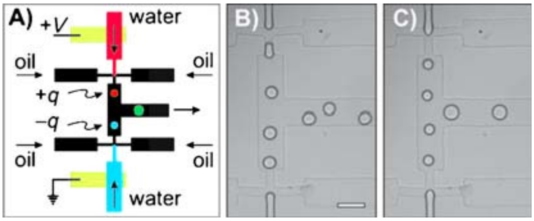


Fig. 15 Electrocoalescence of water droplets in oil medium: **A** schematic view of electrode arrangements, **B** no coalescence without electrical field, and **C** droplet coalescence with the application of electrical field. Reprinted from [106]. copyright Wiley-VCH, reproduced with permission

flow side by side without contact as shown in Fig. 14a. If the flow conditions were slightly different between the two nozzles, the two streams of droplets were produced asynchronously (Fig. 14b), possibly with different size and rate. In this situation, the two droplets had little chance of coalescence near the T-junction, except for cases where the two droplets happened to be very close to each other.

To achieve reliable droplet fusion, a number of techniques have been investigated, which include electrocoalescence (both AC and DC fields), electrocapillarity, thermocapillarity, thermal control, magnetic beads, and surface-directed and channel geometry-directed methods.

The electrocoalescence technique uses electrical force to attract two or more droplets and induce coalescence. It is one of the most commonly used techniques for separating oil from water in the petrochemical industry [103]. Such a technique has also been used for microfluidics applications [104–109]. Figure 15 shows an example of one such application, where two streams of aqueous droplets were brought together and coalesced to form one stream of larger drops [106].

The surface of the channel walls can also be modified actively or passively to control the movement of droplets. For example, electrocapillarity [110–116] such as electrowetting-based actuation has been used for the manipulation of droplets in the microfluidic systems (Fig. 16). The surface energy changes by the applied electric field and such a change can be utilized for changing the wettability of the dielectric surface and the contact angle of the droplet. Thus, the droplet under control moves into the directed position and merges with the other droplet. Fidalgo et al. [117] presented a method of drop fusion by locally achieving a hydrophilic surface such that droplets of different components are trapped and fused with the ones behind.

Droplet fusion can also be achieved passively by special channel designs to force the contact of droplets [79, 90, 101]. These include (1) channel restriction to force the contact of drops [118], (2) channel geometry variation to slow down drop movement in the front to enhance the contact (i.e., tapered

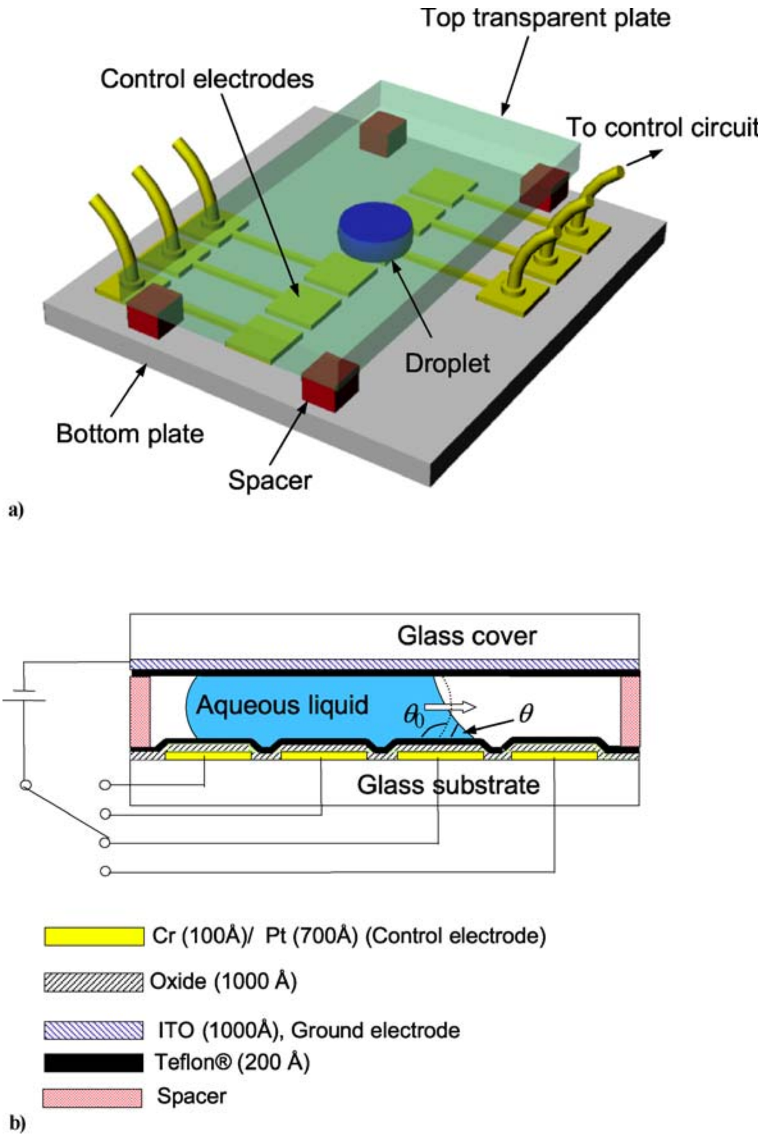


Fig. 16 EWOD-based device for droplet control and coalescence. **a** Artist's view; *top plate* is transparent. **b** Cross-section view; *dotted line* indicates the shape of meniscus before actuation. Reprinted with permission from [111], copyright 2003, IEEE

chamber as shown in Fig. 17 [90], sudden expansion [119–121]), (3) channel jointing to enable spontaneous merging [101], and (4) trapping mechanisms [119].

Kohler et al. [119] proposed the idea of slowing down one droplet by altering local temperature to enhance the contact with the faster-moving

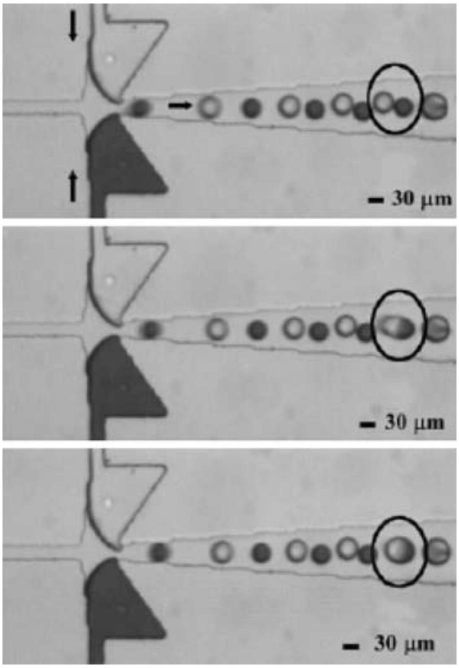


Fig. 17 Dynamic fusion of alternating droplets. From [90], reproduced by permission of the Royal Society of Chemistry

droplet behind. Thermocapillary force was used successfully [73, 122] for fusing droplets. An example is shown in Fig. 18, in which a dye-filled droplet (187 µm diameter, lower left) was moved by the focus of a laser (black ar-

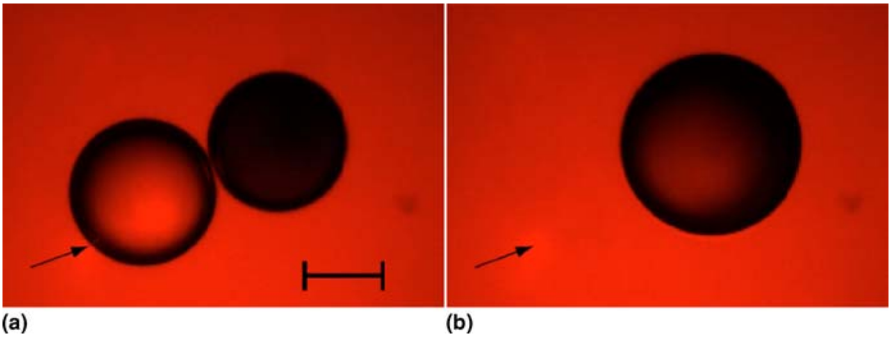


Fig. 18 Fusion of small droplets based on the thermal Marangoni effect using laser heating: **a** before fusion, **b** after fusion. *Black arrow* indicates the laser. Reprinted with permission from [122], copyright 2004, American Institute of Physics

row) to a droplet containing black India ink (182 μm diameter) and fused into one. The droplet fusion and mixing took place on a time scale of less than 33 ms. A review of the thermocapillary force for various microfluidic control was presented in [123]. Demonstration of droplet fusion using thermal control and magnetic beads techniques were given in [119] and [124], respectively.

4.4 Droplet Splitting

The splitting of droplets is mainly achieved by channel branching or obstacles. At the junction or obstacle point, the splitting of carrier flow will break the droplet into two and each flows into a separate branch channel. This technique works when the droplet size is of the same order as the channel width or larger. By arranging a difference in the branching channel size (thus different flow resistance), the droplets can break into different size droplets accordingly. Such a technique is relatively easy to implement on-chip and has been used in a number of studies [46, 49, 83, 119, 125–128]. Figure 19 shows two examples of droplet splitting using obstacles and channel branching [125, 126]. For droplets much smaller than the channel size, the splitting is much more difficult since they tend to flow in the middle of the channels and are not affected by the boundaries.

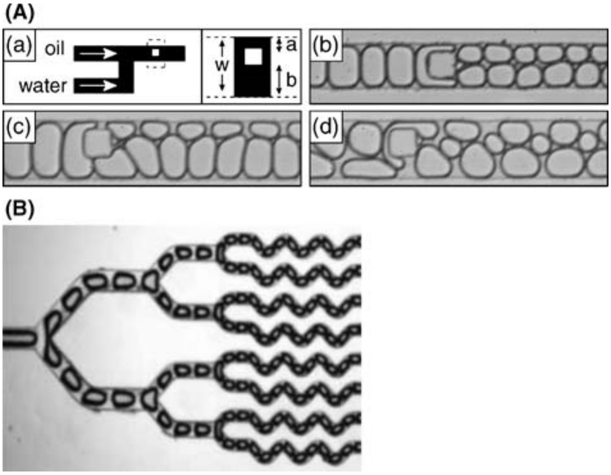


Fig. 19 **A** Passive droplet breakup using obstacles: **a** shows obstacle position; **b**, **c**, and **d** show how different sized droplets can be formed. **B** Passive droplet breakup using channel branching. **A** Reprinted from [126] with permission, copyright 2004, American Physical Society. **B** Reprinted from [125] with permission

4.5 Droplet Sorting

A droplet sorter is an essential component of a microfluidic droplet-reaction system. Sorting techniques have been extensively used in flow cytometry for cell and particle analysis. There is a vast volume of publications on techniques for sorting cells using microfluidic technologies [129–144]. For microfluidic IVC applications, the task of sorting is to identify the droplets of interest and select them. The identification may be based on the fluorescence intensity, which indicates if the protein of interest has been synthesized in the droplet. The selection process is to direct the droplets of interest into a separate fluidic channel for further analysis or storage. These droplets may also be further purified and fed back into the process for a second iteration of synthesis.

For droplet sorting, many of the forces or principles that have been used for droplets fusion can also be utilized for sorting purposes [123]. For example, electrical force has been commonly used in lab-on-a-chip devices for sorting droplets [106, 109, 131, 145–147]. The use of electrostatic charging of droplets provides a means of control with high precision and speed, and without moving parts. Figure 20 shows an example of sorting using dielectrophoretic (DEP) force [145]. Droplets are diverted into the selection channel by the DEP force that is present when potential is applied to the electrodes. Without the potential, droplets will simply flow into the waste port.

A different arrangement of electrical technique is presented in [106], in which an electrical field was applied near a bifurcating channel to direct charged droplets. Electrowetting-based techniques have been used success-

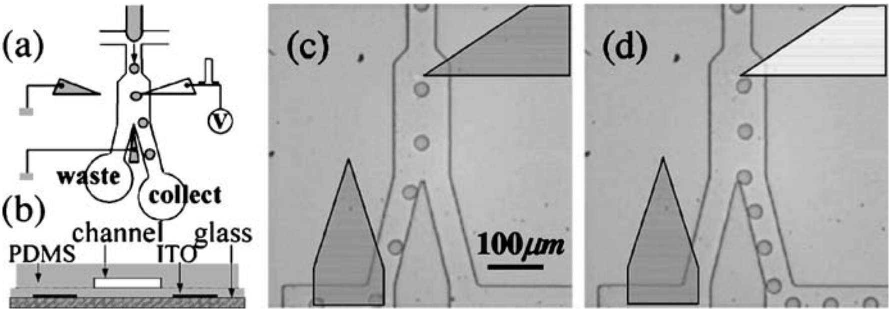


Fig. 20 Droplet sorting using dielectrophoretic force. **a** Top view of the droplet formation and sorting device. **b** Cross-section of the device. The molded PDMS microfluidic channel is aligned to the 30-micron PDMS layer, which is spin-coated onto the patterned ITO electrodes. **c** In the absence of an electric field water drops flow into waste channel. **d** Applying an electric field, the drops are attracted toward the energized electrode and flow into the collect channel. Transparent ITO electrodes have been drawn in *gray* for grounded electrodes and *white* for energized electrodes. Reprinted from [145] with permission, copyright 2004, American Institute of Physics

fully for the control of droplet movement. They can also be used for sorting droplets. Perhaps the first example of a microfluidic sorter using the technique was presented in [116]. Although the original device was developed for sorting droplets in air, the principle should be applicable for water-in-oil droplets. Other examples of the techniques for droplet control include [108, 111, 113–115].

Thermocapillary or Marangoni force was used for directing droplet flows in microchannels to achieve selection [73, 122, 148, 149]. The force could be induced by laser power [122] or microheaters near the microchannel [150].

A number of studies used hydrodynamic flow control either actively [132, 142, 144, 151, 152] or passively [153] to sort cells or droplets. The cell sorting techniques are also applicable for droplet sorting. An example of the sorter is shown in Fig. 21 [144]. A side channel was used to push the cell streams from one channel to another when required. Magnetic force has

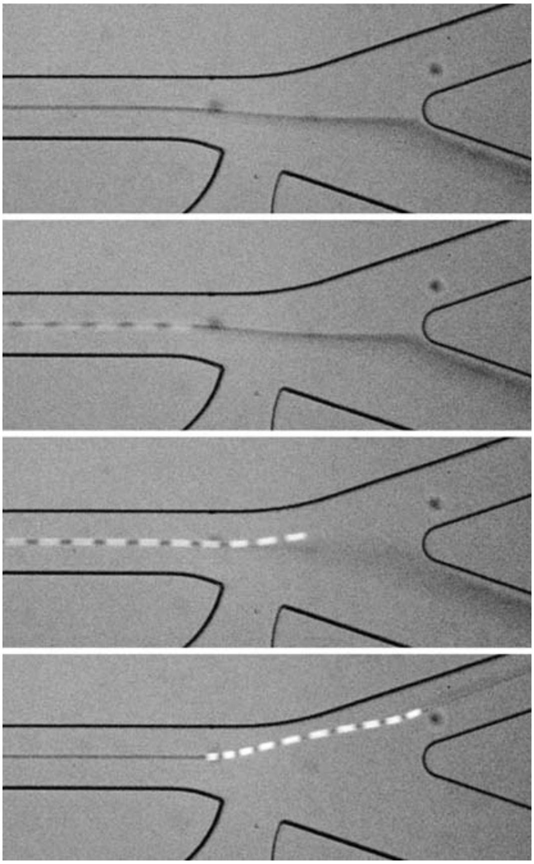


Fig. 21 Cell sorting using hydrodynamic flow manipulation. Reprinted from [144] with permission, copyright 2006, Springer

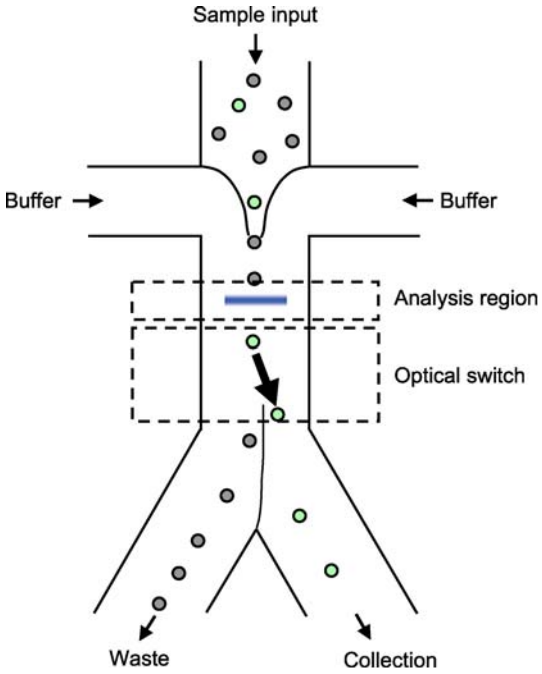


Fig. 22 Layout of the microfluidic sorting junction and the optical switch. Reprinted from [141] with permission, copyright 2005, Macmillan

been utilized in micro-devices for flow manipulation. For example, a permanent magnet was used for separating apoptotic cells loaded with magnetic beads [154]. Such a technique has also been used for the manipulation of droplets [124]. A magnetohydrodynamic switch was developed using two AC MHD micropumps that control the pressure of the two channels. Fluid flow was switched from one channel to another when a certain electrical current phase difference between the two pumps was reached [155]. Such a device was tested using 5 μm beads.

Optical techniques have been used for cell, particle, and droplet manipulation [129, 135, 136, 138, 141, 156]. A laser beam is normally focused to a location of a microchannel and traps a cell or particle to the position. Subsequent switching allows the cell or particle to change flow path to achieve sorting. Figure 22 shows an example of such an optical switch [141]. After being aligned to the center of the channel by flow focusing, cells are analyzed and then switched according to their detected fluorescence. Target cells are directed by the laser to the collection output while all other cells flow to the waste output.

4.6 Other Technologies

In addition to the technologies discussed above, a complete droplet-based IVC chip may also need components such as micromixer, droplet storage, PCR, sample delivery and so on.

4.6.1 Addressable Storage

In the case of long incubation time or when selected droplets need to be stored for future analysis, a storage chamber is required to keep all selected droplets. The crucial microfluidic task is to make sure all droplets in storage are addressable. Little work has been done in this area. The issue has only drawn attention very recently. The work of Kovac and Voldman [157] reported an opto-fluidic cell sorting device in which cells flowing in carrier fluids settle into wells. An imaging technique was used to visualize cells by fluorescence intensity. An optical scattering force was used to push cells of interest into the main flow for downstream collection. Tan and Takeuchi [158, 159] introduced a device that hydrodynamically traps beads in an array and employs microbubbles to release arrayed beads from traps (Fig. 23).

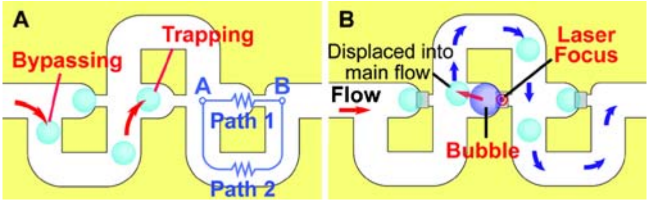


Fig. 23 Diagram of the microfluidic trap (**A**) and release device (**B**). Reprinted from [159] with permission

4.6.2 Microdrop-Based Polymerase Chain Reaction

In IVC applications, the linkage of genotype to a phenotype ideally requires a single gene molecule to appear in one droplet. Only a limited number of copies of the protein molecules could be produced in such a case. This makes the detection of the proteins very challenging, especially when the detection is achieved at a single time-point. Many of the single molecule detection techniques such as fluorescence correlation spectroscopy (FCS), atomic force microscopy (AFM), surface enhanced Raman spectroscopy (SERS) etc. cannot be easily applied here. One way of circumventing such a problem is to apply a gene amplification technique to the droplets so that many copies of the

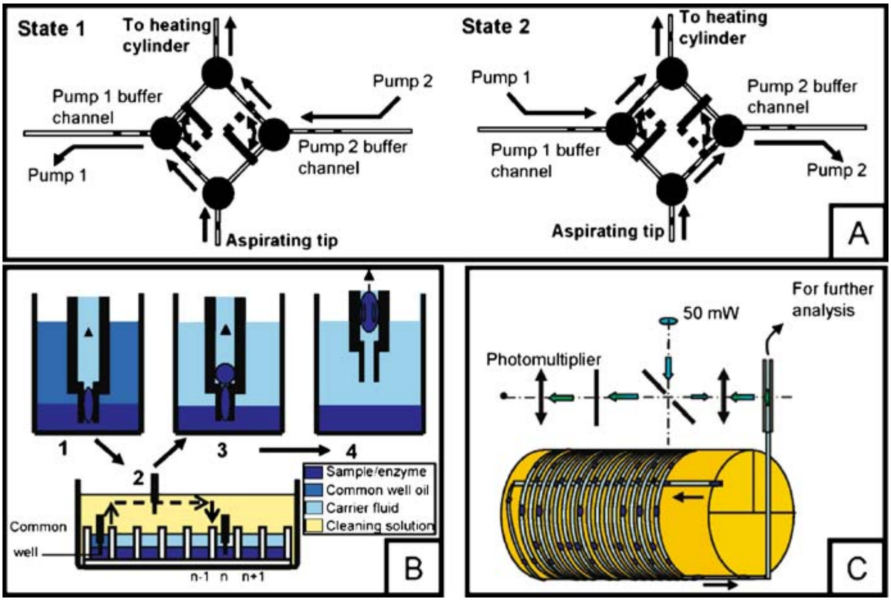


Fig. 24 Automated system performing continuous sampling, reagent mixing, and polymerase chain reaction in microdroplets transported in immiscible oil. **A** Combination of sequential injection with continuous flow using a two-state loop with two syringe pumps, each connected to a three-way pinch valve. *State 1*: pump 1 sequentially forms samples from the aspirating tip while pump 2 continuously pushes the droplets formed in state 2 of the last cycle to the heating cylinder. *State 2*: pump 1 continuously pushes the droplets formed from state 1 to the heating cylinder, while pump 2 sequentially forms a new train of droplets. **B** Steps 1-4 of the injection process from a MTP containing three liquid layers, using a dual-diameter aspirating tip. **C** Capillary wrapped 35 times around a three-zone heating cylinder, endpoint LIF detection, and output for further storage or analysis. Reprinted from [160], copyright 2006, American Chemical Society

same gene can be produced, thus permitting the production of more protein molecules and therefore their subsequent detection. An in-drop PCR technique was reported recently [28, 160–166]. Figure 24 shows an automated microdroplet platform for PCR.

4.6.3 Droplet Delivery

A droplet delivery system is necessary when droplets of interest are needed for further analysis using existing instruments such as a FACS machine or mass spectroscopy (MS). While there has been a great deal of attention on developing a delivery system from capillary electrophoresis samples to MS, such as the on-chip electrospray technologies, little work has been reported on delivering samples from a single droplet into MS. Such a delivery requires the separation of aqueous drops from the oil carrier medium and sprays the protein onto the MS collection window. The interface of the droplet-based microfluidic system to FACS would require double emulsion droplets to be produced where the carrier medium is aqueous solution instead of oil.

4.6.4 Micromixing

Mixing of reagents in droplets was mainly achieved by passing droplets through tortuous channels so that droplets could undergo various stages of stretching and deformation to induce mixing [12, 59, 167, 168].

5 Future Prospects

The quest for the ultimate microchip for protein studies is the drive for the development of various microfluidic droplet technologies. While substantial investigations have been carried out and have demonstrated the capability of droplet formation, fusion, splitting, reagent mixing, incubation, sorting, and PCR, the ultimate goal is to be able to integrate various components into one chip and to analyze the protein produced and, if necessary, to re-inject the protein of interest back to the initial line for second or third round evolution. It is envisaged that more and more studies will be focused on the coupling of existing analytical instruments (such as MS, FACS etc.) to the droplet platform.

The capability of combining cells and beads into droplets also opens new opportunities for genomics, proteomics, and cellomic applications. For example, the actual levels of activities of endogenous cellular enzymes may be directly monitored without the need for analyzing mRNA or protein expression levels through GFP fusions [7]. DNA, protein, and cell analysis may be carried out at high speed due to the capability of continuous and fast on-drop droplet generation. The high-throughout capability also allows new materials (e.g. nano-/microparticles) and biomolecules (e.g. new enzymes) to be synthesized and screened at a rate current techniques cannot achieve.

6 Concluding Remarks

In vitro compartmentalization (IVC) is a power tool for studying protein–protein reactions due to the high capacity and robustness of droplet technologies. There have been a number of emulsion droplet-based IVC studies for protein–protein interactions and the trend is to realize the droplet technology in a microfluidic chip. There has been increasing interest in microfluidic-based IVC, both in the fundamental understanding of on-chip drop generation and control, and in the device development for applications in chemical and biological reactions. It has been demonstrated from a number of studies that the aqueous microdroplets in microfluidic systems can be fused, split, sorted, and incubated. Contents of droplets can be mixed and analyzed, allowing enzymatic reactions and cell-free translation to be performed within the droplets. With the development of analytical tools such as droplet storage, delivery and so on, automated instruments could be developed, providing unprecedented power that current analytical instruments cannot match.

Acknowledgements Special thanks go to Dr. Sue Brown from CSIRO Entomology, Canberra, Australia and Dr. Peter Hudson from CSIRO Molecular Health Technologies, Parkville, Melbourne, Australia for their invaluable discussions in protein evolution and protein analysis.

References

1. Xie J, Schultz PG (2005) Adding amino acids to the genetic repertoire. Curr Opin Chem Biol 9:548–554
2. Tawfik DS, Griffiths AD (1998) Man-made cell-like compartments for molecular evolution. Nat Biotechnol 16:652–656
3. Aharoni A, Griffiths AD, Tawfik DS (2005) High-throughput screens and selections of enzyme-encoding genes. Nat Biotechnol 9:210–216
4. Amstutz P et al. (2001) In vitro display technologies: novel developments and applications. Curr Opin Biotechnol 12:400–405
5. Dower WJ, Mattheakis LC (2002) In vitro selection as a powerful tool for the applied evolution of proteins and peptides. Curr Opin Chem Biol 6:390–398
6. Griffiths AD, Tawfik DS (2000) Man-made enzymes – from design to in vitro compartmentalisation. Curr Opin Biotechnol 11:338–353
7. Griffiths AD, Tawfik DS (2006) Miniaturising the laboratory in emulsion droplets. Trends Biotechnol 24(9):395–402
8. Jestin J-L, Kaminski PA (2004) Directed enzyme evolution and selections for catalysis based on product formation. J Biotechnol 113:85–103
9. Kelly BT et al. (2007) Miniaturizing chemistry and biology in microdroplets. Chem Commun, pp 1773–1788
10. Leamon LH et al. (2006) Overview: methods and applications for droplet compartmentalization of biology. Nat Methods 3(7):541–543
11. Rothe A, Surjadi RN, Power BE (2006) Novel proteins in emulsions using in vitro compartmentalization. Trends Biotechnol 24:587–592
12. Song H, Chen DL, Ismagilov RF (2006) Reactions in droplets in microfluidic channels. Angew Chem Int Edit 45:7336–7356
13. DeMello A (2006) Control and detection of chemical reactions in microfluidic systems. Nature 442:394–402
14. Cristini V, Tan Y-C (2004) Theory and numerical simulation of droplet dynamics in complex flows – a review. Lab Chip 4:257–264
15. Atencia J, Beebe DJ (2005) Controlled microfluidic interfaces. Nature 437:648–655

16. Squires TM, Quake SR (2005) Microfluidics: fluid physics at the nanoliter scale. Rev Mod Phys 77:977–1026
17. Doktycz MJ, Simpson ML (2007) Nano-enabled synthetic biology. Mol Syst Biol 3:1–10
18. Arnold FH (1998) Design by directed evolution. Acc Chem Res 31:125–131
19. Arnold FH, Georgiou G (2003) Directed enzyme evolution. Methods Molec Biol 230
20. Besenmatter W, Kast P, Hilvert D (2004) New enzymes from combinatorial library modules. In: Robertson D, Noel JP (eds) Methods in enzymology, vol 388. Academic, New York, pp 91–102
21. Otten LG, Quax WJ (2005) Directed evolution-selecting today's biocatalysts. Biomolec Eng 22:1–9
22. Voigt CA, Kauffman S, Wang Z-G (2001) Rational evolutionary design: the theory of in vitro protein evolution. Adv Protein Chem 55:79–160
23. Ghadessy FJ, Holliger P (2004) A novel emulsion mixture for in vitro compartmentalization of transcription and translation in the rabbit reticulocyte system. Protein Eng Des Sel 17(3):201–204
24. Yang H et al. (2003) Evolution of an organophosphate-degrading enzyme: a comparison of natural and directed evolution. Protein Eng 16(2):135–145
25. Bernath K et al. (2004) In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. Anal Biochem 325:151–157
26. Johannes TW, Zhao H (2006) Directed evolution of enzymes and biosynthetic pathways. Curr Opin Microbiol 9:261–267
27. Miller OJ et al. (2006) Directed evolution by in vitro compartmentalization. Nat Methods 3(7):561–570
28. Williams R et al. (2006) Amplification of complex gene libraries by emulsion PCR. Nat Methods 3(7):545–550
29. Doi N, Yanagawa H (1999) STABLE: Protein–DNA fusion system for screening of combinatorial protein libraries in vitro. FEBS Lett 457:227–230
30. Sepp A, Choo Y (2005) Cell-free selection of zinc finger DNA binding proteins using in vitro compartmentalization. J Mol Biol 354:212–219
31. Yonezawa M et al. (2003) DNA display for in vitro selection of diverse peptide libraries. Nucleic Acids Res 31:e118
32. Doi N et al. (2004) In vitro selection of restriction endonucleases by in vitro compartmentalization. Nucleic Acids Res 32(12):e95
33. Pietrini AV, Luisi PL (2004) Cell-free protein synthesis through solubilisate exchange in water/oil emulsion compartments. ChemBioChem 5:1055–1062
34. Agresti JJ et al. (2005) Selection of ribozymes that catalyse multiple-turnover Diels–Alder cycloadditions by using in vitro compartmentalization. PNAS 102(45):16170–16175
35. Cohen HM, Tawfik DS, Griffiths AD (2004) Altering the sequence specificity of HaeIII methyltransferase by directed evolution using in vitro compartmentalization. Protein Eng Des Sel 17(1):3–11
36. Griffiths AD, Tawfik DS (2003) Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. EMBO J 22(1):24–35
37. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. PNAS 98(8):4552–4557
38. Bernath K, Magdassi S, Tawfik DS (2005) Directed evolution of protein inhibitors of DNAnucleases by in vitro compartmentalization (IVC) and nano-droplet delivery. J Mol Biol 345:1015–1026

39. Leemhuis H et al. (2005) New genotype–phenotype linkages for directed evolution of functional proteins. Curr Opin Struct Biol 15:472–478
40. Endo Y, Sawasaki T (2006) Cell-free expression systems for eukaryotic protein production. Curr Opin Biotechnol 17:373–380
41. Endo Y, Sawasaki T (2004) High-throughput, genome-scale protein production method based on the wheat germ cell-free expression system. J Strucz Funct Genomics 5:45–57
42. Shimizu Y, Kanamori T, Ueda T (2005) Protein synthesis by pure translation systems. Methods 36:299–304
43. Lee ER (2003) Microdrop generation. CRC, Boca Raton
44. Garstecki P et al. (2006) Formation of droplets and bubbles in a microfluidic T-junction – scaling and mechanism of break-up. Lab Chip 6:437–446
45. Husny J et al. (2006) The creation of drops in T-shaped microfluidic devices with the modified laser LIGA technique: I. Fabrication. Smart Matter Struct 15:S117–S123
46. Menetrier L, Tabeling P (2006) Droplet break-up in junctions: the concept of a critical length. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Society for Chemistry and Micro-Nano Systems (CHEMINAS), Japan, 2006, pp 98–101
47. Nisisako T, Okushima S, Torii T (2005) Controlled formulation of monodisperse double emulsions in a multiple-phase microfluidic system. Soft Matter 1:23–27
48. Nisisako T, Torii T, Higuchi T (2002) Droplet formation in a microchannel network. Lab Chip 2:24–26
49. Song H et al. (2003) Experimental test of scaling of mixing by chaotic advection in droplets moving through microfluidic channels. Appl Phys Lett 83(22):4664–4666
50. Tice JD et al. (2003) Formation of droplets and mixing in multiphase microfluidics at low values of the Reynolds and the capillary numbers. Langmuir 19:9127–9133
51. Van der Graaf S et al. (2005) Droplet formation in a T-shaped microchannel junction: a model system for membrane emulsification. Colloid Surf A 266:106–116
52. Kohler JM, Kirner T (2005) Nanoliter segment formation in micro fluid devices for chemical and biological micro serial flow processes in dependence on flow rate and viscosity. Sens Actuators A 119:19–27
53. Fuerstman MJ, Garstecki P, Whitesides GM (2007) Coding/decoding and reversibility of droplet trains in microfluidic networks. Science 315:828–832
54. He M et al. (2005) Selective encapsulation of single cells and subcellular organelles into picoliter- and femtoliter-volume droplets. Anal Chem 77:1539–1544
55. Thorsen T et al. (2001) Dynamic pattern formation in a vesicle-generating microfluidic device. Phys Rev Lett 86:4163–4166
56. Chen DL et al. (2007) Using three-phase flow of immiscible liquids to prevent coalescence of droplets in microfluidic channels: criteria to identify the third liquid and validation with protein crystallization. Langmuir 23:2255–2260
57. Zheng B et al. (2004) A droplet-based, composite PDMS/glass capillary microfluidic system for evaluating protein crystallization conditions by microbatch and vapor-diffusion methods with on-chip X-ray diffraction. Angew Chem Int Edit 43:2508–2511
58. Amici E et al. (2008) Alginate gelation in microfluidic channels. Food Hydrocolloids 2008:97–104
59. Huebner A et al. (2007) Quantitative detection of protein expression in single cells using droplet microfluidics. Chem Commun, pp 1218–1220
60. Sgro AE, Allen PB, Chiu DT (2007) Thermoelectric manipulation of aqueous droplets in microfluidic devices. Anal Chem 79:4845–4851

61. Ambravaneswaran B, Wilkes ED, Basarana OA (2002) Drop formation from a capillary tube: comparison of one-dimensional and two-dimensional analyses and occurrence of satellite drops. Phys Fluids 14:(8):2606–2621
62. Bogy DB (1979) Drop formation in a circular liquid jet. Ann Rev Fluid Mech 11:207–228
63. Cramer C, Fischer P, Windhab EJ (2004) Drop formation in a Co-flowing ambient fluid. Chem Eng Sci 59:3045–3058
64. Lister JR, Stone HA (1998) Capillary breakup of a viscous thread surrounded by another viscous fluid. Phys Fluids 10(11):2758–2764
65. Sugiura S et al. (2001) Interfacial tension driven monodispersed droplet formation from microfabricated channel array. Langmuir 17:5562–5566
66. Sugiura S et al. (2000) Preparation of monodispersed solid lipid microspheres using a microchannel emulsification technique. J Colloid Interf Sci 227:95–103
67. Taylor GI (1934) The formation of emulsions in definable fields of flow. Proc R Soc London A 146:501–523
68. Tomotika S (1935) On the instability of a cylindrical thread of a viscous liquid surrounded by another viscous fluid. Proc R Soc London A CL:322–337
69. Umbanhowar PB, Prasad V, Weitz DA (2000) Monodisperse emulsion generation via drop break off in a coflowing stream. Langmuir 16:347–351
70. Zhang DF, Stone HA (1997) Drop formation in viscous flows at a vertical capillary tube. Phys Fluids 9(8):2234–2242
71. Zhang X, Basaran OA (1995) An experimental study of dynamics of drop formation. Phys Fluids 7(6):1184–1203
72. Anna SL, Bontoux N, Stone HA (2003) Formation of dispersions using flow focusing in microchannels. Appl Phys Lett 82:364–366
73. Baroud CN, Robert de Saint Vincent M, Delville J-P (2007) An optical toolbox for total control of droplet microfluidics. Lab Chip 7:1029–1033
74. Dittrich PS, Jahnz M, Schwille P (2005) A new embedded process for compartmentalized cell-free protein expression and on-line detection in microfluidic devices. Communications 6:811–814
75. Joanicot M, Ajdari A (2005) Droplet control for microfluidics. Science 309:887–888
76. Luo C et al. (2006) Picoliter-volume aqueous droplets in oil: electrochemical detection and yeast cell electroporation. Electrophoresis 27:1977–1983
77. Shestopalov I, Tice JD, Ismagilov RF (2004) Multi-step synthesis of nanoparticles performed on millisecond time scale in a microfluidic droplet-based system. Lab Chip 4:316–321
78. Tan Y-C, Cristini V, Lee AP (2006) Monodispersed microfluidic droplet generation by shear focusing microfluidic device. Sens Actuators B 114:350–356
79. Tan Y-C et al. (2004) Design of microfluidic channel geometries for the control of droplet volume, chemical concentration, and sorting. Lab Chip 4:292–298
80. Tolosa L-I et al. (2006) Combined effects of formulation and stirring on emulsion drop size in the vicinity of three-phase behavior of surfactant-oil water systems. Indian Eng Chem Res 45:3810–3814
81. Utada AS et al. (2005) Monodisperse double emulsions generated from a microcapillary device. Science 308:537–541
82. Chan EM, Alivisatos AP, Mathies RA (2005) High-temperature microfluidic synthesis of CdSe nanocrystals in nanoliter droplets. J Am Chem Soc 127:13854–13861
83. Dreyfus R, Tabeling P, Willaime H (2003) Ordered and disordered patterns in two-phase flows in microchannels. Phys Rev Lett 90(14):144505

84. He M, Kuo JS, Chiu DT (2005) Electro-generation of single femtoliter- and picoliter-volume aqueous droplets in microfluidic systems. Appl Phys Lett 87:031916
85. Nisisako T, Torii T (2007) Formation of biphasic Janus droplets in a microfabricated channel for the synthesis of shape-controlled polymer microparticles. Adv Mater 19(11):1489–1493
86. Seo M et al. (2007) Microfluidic consecutive flow-focusing droplet generators. Soft Matter 3:986–992
87. Yang C-H et al. (2007) Using a cross-flow microfluidic chip and external crosslinking reaction for monodisperse TPP-chitosan microparticles. Sens Actuators B 124:510–516
88. Brouzes E et al. (2006) Droplet-based high-throughput live/dead cell assay. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Society for Chemistry and Micro-Nano Systems (CHEMINAS), Japan, 2006, pp 1043–1045
89. Baroud CN, Willaime H (2004) Multiphase flows in microfluidics. CR Physique 5:547–555
90. Hung L-H et al. (2006) Alternating droplet generation and controlled dynamic droplet fusion in microfluidic device for CdS nanoparticle synthesis. Lab Chip 6:174–178
91. Xu Q, Nakajima M (2004) The generation of highly monodisperse droplets through the breakup of hydrodynamically focused microthread in a microfluidic device. Appl Phys Lett 85(17):3726–3728
92. Ong W-L et al. (2007) Experimental and computational analysis of droplet formation in a high-performance flow-focusing geometry. Sens Actuators A 138:203–212
93. Garstecki P, Stone HA, Whitesides GM (2005) Mechanism for flow-rate controlled breakup in confined geometries: a route to monodisperse emulsions. Phys Rev Lett 94:164501
94. Husny J, Cooper-White JJ (2006) The effect of elasticity on drop creation in T-shaped microchannels. J Non-Newtonian Fluid Mech 137:121–136
95. Ward T et al. (2005) Microfluidic flow focusing: drop size and scaling in pressure versus flow-rate-driven pumping. Electrophoresis 26(19):3716–3724
96. Harvie DJE et al. (2006) Aparametric study of droplet deformation through a microfluidic contraction: low viscosity Newtonian droplets. Chem Eng Sci 61:5149–5158
97. Nguyen N-T, Lassemono S, Chollet FA (2006) Optical detection for droplet size control in microfluidic droplet-based analysis systems. Sens Actuators B 117:431–436
98. Rayleigh L (1879) On the capillary phenomena of jets. Proc R Soc London A 29:71–79
99. Zhu Y et al. (2007) Droplets transport in a microfluidic chip for in vitro compartmentalisation. 16th Australasian fluid mechanics conference, Queensland, 3–7 Dec 2007. University of Queensland, Brisbane
100. Adzima BJ, Velankar SS (2006) Pressure drops for droplet flows in microfluidic channels. J Micromech Microeng 16:1504–1510
101. Song H, Tice JD, Ismagilov RF (2003) A microfluidic system for controlling reaction networks in time. Angew Chem Int Edit 42:767–772
102. Tice JD, Lyon AD, Ismagilov RF (2004) Effects of viscosity on droplet formation and mixing in microfluidic channels. Anal Chim Acta 507:73–77
103. Eow JS et al. (2001) Electrostatic enhancement of coalescence of water droplets in oil: a review of the current understanding. Chem Eng Sci 84:173–192
104. Ahn K et al. (2006) Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. Appl Phys Lett 88:264105

105. Chabert M, Dorfman KD, Viovy J-L (2005) Droplet fusion by alternating current (AC) field electrocoalescence in microchannels. Electrophoresis 26:3706–3715
106. Link DR et al. (2006) Electric control of droplets in microfluidic devices. Angew Chem Int Edit 45:2556–2560
107. Priest C, Herminghaus S, Seemannc R (2006) Controlled electrocoalescence in microfluidics: targeting a single lamella. Appl Phys Lett 89:134101
108. Schwartz JA, Vykoukal JV, Gascoyne PRC (2004) Droplet-based chemistry on a programmable micro-chip. Lab Chip 4:11–17
109. Singh P, Aubry N (2007) Transport and deformation of droplets in a microdevice using dielectrophoresis. Electrophoresis 28:644–657
110. Armani M et al. (2005) Control of microfluidic systems: two examples, results, and challenges. Int J Robust Nonlinear Control 15:785–803
111. Cho SK, Moon H, Kim C-J (2003) Creating, transporting, cutting, and merging liquid droplets by electrowetting-based actuation for digital microfluidic circuits. J Microelectromech Syst 12:70–80
112. Jones TB et al. (2001) Dielectrophoretic liquid actuation and nanodroplet formation. J Appl Phys 89(2):1441–1448
113. Lee J, Kim C-JC (2000) Surface-tension-driven microactuation based on continuous electrowetting. J Microelectromech Syst 9(2):171–180
114. Paik P, Pamula VK, Fair RB (2003) Rapid droplet mixers for digital microfluidic systems. Lab Chip 3:253–259
115. Pollack MG, Shenderovb AD, Fair RB (2002) Electrowetting-based actuation of droplets for integrated microfluidics. Lab Chip 2:96–101
116. Washizu M (1998) Electrostatic actuation of liquid droplets for microreactor applications. IEEE Trans Ind Appl 34:732–737
117. Fidalgo LM, Abell C, Huck WTS (2007) Surface-induced droplet fusion in microfluidic devices. Lab Chip 7:984–986
118. Yan L, Thompson KE, Valsaraj KT (2006) A numerical study on the coalescence of emulsion droplets in a constricted capillary tube. J Colloid Interf Sci 298:832–844
119. Kohler JM et al. (2004) Digital reaction technology by micro segmented flow – components, concepts and applications. Chem Eng J 101:201–216
120. Liu K et al. (2007) Droplet-based synthetic method using microflow focusing and droplet fusion. Microfluid Nanofluid 3:239–243
121. Tan Y-C, Ho YL, Lee AP (2007) Droplet coalescence by geometrically mediated flow in microfluidic channels. Microfluid Nanofluid 3:495–499
122. Kotz KT, Noble KA, Faris GW (2004) Optical microfluidics. Appl Phys Lett 85(13):2858–2660
123. Darhuber AA, Troian SM (2005) Principles of microfluidic actuation by modulation of surface stresses. Ann Rev Fluid Mech 37:425–455
124. Shikida M et al. (2006) Using wettability and interfacial tension to handle droplets of magnetic beads in a micro-chemical-analysis system. Sens Actuators B 113:563–569
125. Heieh AT-H et al. (2006) Monodisperse liposomal gene carrier formulation in picoliter micro-reactor for consistent and efficient gene delivery. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Society for Chemistry and Micro-Nano Systems (CHEMINAS), Japan, 2006, pp 1369–1371
126. Link DR et al. (2004) Geometrically mediated breakup of drops in microfluidic devices. Phys Rev Lett 92:054503
127. Menetrier-Deremble L, Tabeling P (2006) Droplet breakup in microfluidic junctions of arbitary angles. Phys Rev E 74:035303

128. Ting TH et al. (2006) Thermally mediated breakup of drops in microchannels. Appl Phys Lett 89:234101
129. Arakawa T et al. (2006) Accurate high speed particles and biomolecules sorting microsystem using 3-dimensional sheath flow. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 512–514
130. Chang W-Y, Liu C-H (2006) A cell switching microsystem for single cell sorting application via enhanced dielectrophoresis design. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 1474–1476
131. Dittrich PS, Schwille P (2003) An integrated microfluidic system for reaction, high-sensitivity detection, and sorting of fluorescent cells and particles. Anal Chem 75:5767–5774
132. Fu AY et al. (2002) An integrated microfabricated cell sorter. Anal Chem 74:2451–2457
133. Fu AY et al. (1999) A microfabricated fluorescence-activated cell sorter. Nat Biotechnol 17:1109–1111
134. Gawad S, Schild L, Renaud P (2001) Micromachined impedance spectroscopy flow cytometer for cell analysis and particle sizing. Lab Chip 1:76–82
135. Grier DG (2003) A revolution in optical manipulation. Nature 424:810–816
136. MacDonald MP, Dholakia GCSK (2003) Microfluidic sorting in an optical lattice. Nature 426:421–424
137. McClain MA et al. (2003) Microfluidic devices for chemical analysis of cells. Anal Chem 75:5646–5655
138. Perroud TD, Patel KD (2006) Rapid fluorescence-activated cell sorting with optical-force deflection in a microfluidic device. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 984–986
139. Smith AE et al. (2006) Continuous flow particle sorting at low applied electric fields using electrodeless dielectrophoresis in ridged polymeric microstructures. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 1187–1189
140. Wang L et al. (2007) Dielectrophoresis switching with vertical sidewall electrodes for microfluidic flow cytometry. Lab Chip 7:1114–1120
141. Wang MM et al. (2005) Microfluidic sorting of mammalian cells by optical force switching. Nat Biotechnol 23:83–87
142. Wolff A et al. (2003) Integrating advanced functionality in a microfabricated high-throughput fluorescent-activated cell sorter. Lab Chip 3:22–27
143. Yi C et al. (2006) Microfluidics technology for manipulation and analysis of biological cells. Anal Chim Acta 560:1–23
144. Bang H et al. (2006) Microfabricated fluorescence-activated cell sorter through hydrodynamic flow manipulation. Microsyst Technol 12:746–753
145. Ahn K et al. (2006) Dielectrophoretic manipulation of drops for high-speed microfluidic sorting devices. Appl Phys Lett 88:024104
146. Gallardo BS et al. (1999) Electrochemical principles for active control of liquids on submillimeter scales. Science 283:57–60
147. Barbulovic-Nad I et al. (2006) DC-dielectrophoretic separation of microparticles using an oil droplet obstacle. Lab Chip 6:274–279
148. Baroud CN et al. (2007) Thermocapillary valve for droplet production and sorting. Phys Rev E 75:046302
149. Farahi RH et al. (2004) Microfluidic manipulation via marangoni forces. Appl Phys Lett 85(18):4237–4239

150. Glockner PS, Naterer GF (2005) Thermocapillary control of microfluidic transport with a stationary cyclic heat source. J Micromech Microeng 15:2216–2229
151. Chen CC et al. (2004) Design and operation of a microfluidic sorter for Drosophila embryos. Sens Actuators B 102:59–66
152. Tan Y-C, Lee AP (2005) Microfluidic separation of satellite droplets as the basis of a monodispersed micron and submicron emulsification system. Lab Chip 5:1178–1183
153. Cristobal G et al. (2006) On-line laser Raman spectroscopic probing of droplets engineered in microfluidic devices. Lab Chip 6:1140–1146
154. Kim H-S et al. (2006) Magneto-microfluidic device for apototic cell separation. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 416–418
155. Lemoff AV, Lee AP (2003) An AC magnetohydrodynamic microfluidic switch for micro total analysis system. Biomed Microdev 5(1):55–60
156. Burnham DR, McGloin D (2006) Holographic optical trapping of aerosol droplets. Opt Express 14(9):4175–4181
157. Kovac J, Voldman J (2006) Facile image-based cell sorting using opto-flucs (optofluidic cell sorting). In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 1483–1485
158. Tan W-H, Takeuchi S (2006) An optical retrieval microfluidic system for microarray applications. In: Proceedings MicroTAS2006, Tokyo, 5–9 Nov 2006. Transducer Research Foundation, Hilton Head, pp 509–511
159. Tan W-H, Takeuchi S (2007) A trap-and-release integrated microfluidic system for dynamic microarray applications. Proc Natl Acad Sci USA 104(4):1146–1151
160. Chabert M et al. (2006) Automated microdroplet platform for sample manipulation and polymerase chain reaction. Anal Chem 78:7722–7728
161. Diehl F et al. (2006) BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. Nat Methods 3(7):551–559
162. Kojima T et al. (2005) PCR amplification from single DNA molecules on magnetic beads in emulsion: application for high-throughput screening of transcription factor targets. Nucleic Acids Res 33:e150
163. Musyanovych A, Mailander V, Landfester K (2005) Miniemulsion droplets as single molecule nanoreactors for polymerase chain reaction. Biomacromolecules 6:1824–1828
164. Li M et al. (2006) BEAMing up for detection and quantification of rare sequence variants. Nat Methods 3(2):95–97
165. Dorfman KD et al. (2005) Contamination-free continuous flow microfluidic polymerase chain reaction for quantitative and clinical applications. Anal Chem 77:3700–3704
166. Gonzalez A et al. (2007) Gene transcript amplification from cell lysates in continuous-flow microfluidic devices. Biomed Microdev 9:729–736
167. Liau A et al. (2005) Mixing crowded biological solutions in milliseconds. Anal Chem 77:7618–7625
168. Zheng B, Tice JD, Ismagilov RF (2004) Formation of droplets of alternating composition in microfluidic channels and applications to indexing of concentrations in droplet-based assays. Anal Chem 76:4977–4982

Adv Biochem Engin/Biotechnol (2008) 110: 115–152
DOI 10.1007/10_2008_096
© Springer-Verlag Berlin Heidelberg
Published online: 17 April 2008

Large-Scale Analysis of Protein–Protein Interactions Using Cellulose-Bound Peptide Arrays

Ulrike Beutling[1] · Kai Städing[2] · Theresia Stradal[2] · Ronald Frank[1] (✉)

[1]Department of Chemical Biology, Helmholtz Centre for Infection Research,
Inhoffenstraße 7, 38124 Braunschweig, Germany
ronald.frank@helmholtz-hzi.de

[2]Research Group Infection and Motility, Helmholtz Centre for Infection Research,
Inhoffenstraße 7, 38124 Braunschweig, Germany

Abstract Peptide arrays for screening large numbers of peptide fragments and probing with large numbers of samples is discussed.

Keywords Epitope mapping · Microarrays · Peptide arrays · Solid phase peptide synthesis · SPOT-synthesis

Abbreviations

AA	Amino acid
Abu	α-Aminobutyric acid
Acm	Acetyl-aminomethyl
AP	Alkaline phosphatase
BCIP	5-Bromo-4-chloro-3-indolylphosphate *p*-toluidine salt
BPB	Bromophenol blue
BSA	Bovine serum albumin
CBS	Citrate-buffered saline
CDS	Color developing solution
DCM	Dichloromethane
DIC	N,N'-diisopropylcarbodiimide
ECL	Enhanced chemiluminescence
FITC	Fluorescein isothiocyanate
HOBt	*N*-hydroxybenzotriazole
HRP	Horse radish peroxidase
kd	Kilodalton
MBS	Membrane blocking solution
MTBE	*tert*-Butyl methylether
MTT	3-[4,5-Dimethylthiazol-2-yl]-2,5 diphenyltetrazolium bromide
NBT	Nitro blue tetrazolium chloride monohydrate
NMP	*N*-Methyl-2-pyrrolidinone
PBS	Phosphate-buffered saline
PEG	Polyethylene glycol
Pmc	2,2,5,7,8-Pentymethylchroman-6-sulphonyl
PP	Polypropylene
PVC	Polyvinyl chloride
rpm	Revolutions per minute
SC^2	Spotting compound-support conjugates
SDS	Sodium *n*-dodecyl sulphate
SDS-PAGE	Sodium *n*-dodecyl sulphate polyacrylamide gel electrophoresis
SM-A	Stripping mix A
SM-B	Stripping mix B
*t*Bu	*tert*-Butyl
TBS	Tris-buffered saline
TIBS	Triisobutylsilane
Trt	Triphenylmethyl

1 Introduction

Proteins interact via surface accessible interaction sites (Fig. 1 [1]), which involve amino acid side chain and backbone contacts along a linear segment of the protein chain (linear epitopes), or involve amino acid residues from two or more segments of the protein chain brought together by its folded secondary structure (conformational epitopes). Note that the term epitope is used here in its broadest sense for a protein interaction site and far beyond its default immunological meaning. Approaches that systematically study antibody antigen interactions can be directly applied to protein–protein interactions in general.

Linear epitopes can be copied by small peptide fragments that are readily amenable to chemical synthesis. This is also true for a significant part of conformational epitopes when linear components alone can contribute sufficient affinity or when they can be mimicked by linear "mimotope" peptides [2].

Many proteins, most prominently those of regulatory function, are built from smaller domains which are stably folded structural modules still displaying their specific functional property. The catalogue of such domains that recognize linear epitopes is rapidly growing (kringle, SH2, SH3, PH, EVH1, PDZ, WW, etc. [3]) indicating a more general ZIP code principle utilized by nature. These domains are found to be involved in various molecular organization and regulation phenomena.

Complementary to other biochemical approaches, such as large-scale analysis of protein complexes [4, 5] and molecular biology approaches such as

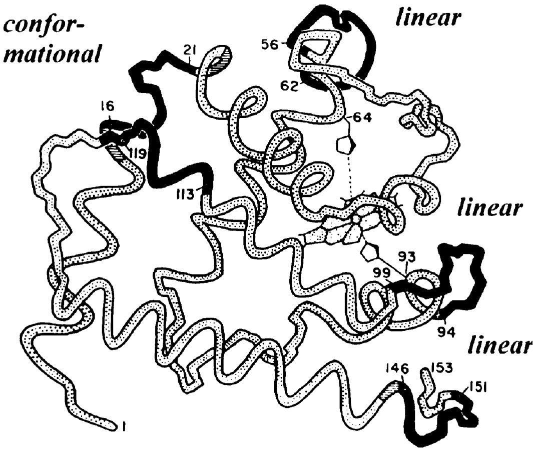


Fig. 1 A protein ligand presenting three linear loop-type and one conformational epitopes for interaction with target proteins. (The drawing of the protein chain was adapted from Atassi [1])

the yeast-two-hybrid method [6, 7], a peptide screening approach will immediately address functional protein interaction sites, leading to a detailed insight into the discovered molecular recognition events, placing them in the context of the whole genome and even allowing to rapidly decipher the chemical nature of these interactions [8]. This information can then be transferred into powerful small peptide tools that interfere with these interactions in vivo and help to link targets with phenotypes [9].

One important aspect of screening protein–protein interactions is to gain access to new targets for drug discovery. Thus, it is logical to set up a genome wide search for all "drugable" proteins and then validate these as relevant pharmaceutical targets by modern proteome analysis. It can be concluded that these drugable targets primarily belong to that repertoire of proteins that can bind small molecule ligands. Synthetic peptides are practical tools readily at hand to address this property. Although peptides themselves have lost attractiveness as pharmaceutical drugs, they are perfect molecular probes for the search of new pharmaceutical targets.

SPOT-synthesis [10] is an easy and very flexible technique for simultaneous parallel assembly of peptides on membrane supports (see Fig. 2). This method gives researchers rapid and low cost access to a large number of

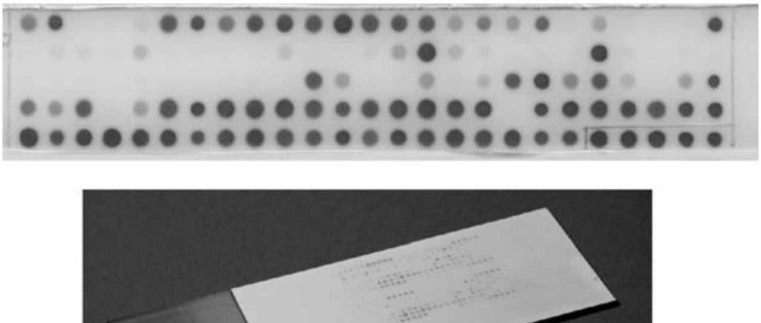


Fig. 2 Comparison of a membrane-bound macroarray with a cellulose-bound miniarray on a microscope slide (original size). (*top*) A macroarray of 120 peptides comprising a full single amino acid replacement set of the peptide Ac-NYGKYE-βAla was synthesized on a conventional AC-S01 cellulose membrane in lines of 25 spots each at a distance of 4 mm and assayed with monoclonal antibody 1D3 followed by an AP conjugated secondary antibody and color signal development with the BCIP/MTT substrates. (*bottom*) The same set of 120 peptides synthesized on an acid soluble cellulose membrane, processed through the SC^2-protocol, printed onto a plastic coated glass slide in duplicate in the same array layout and assayed as above

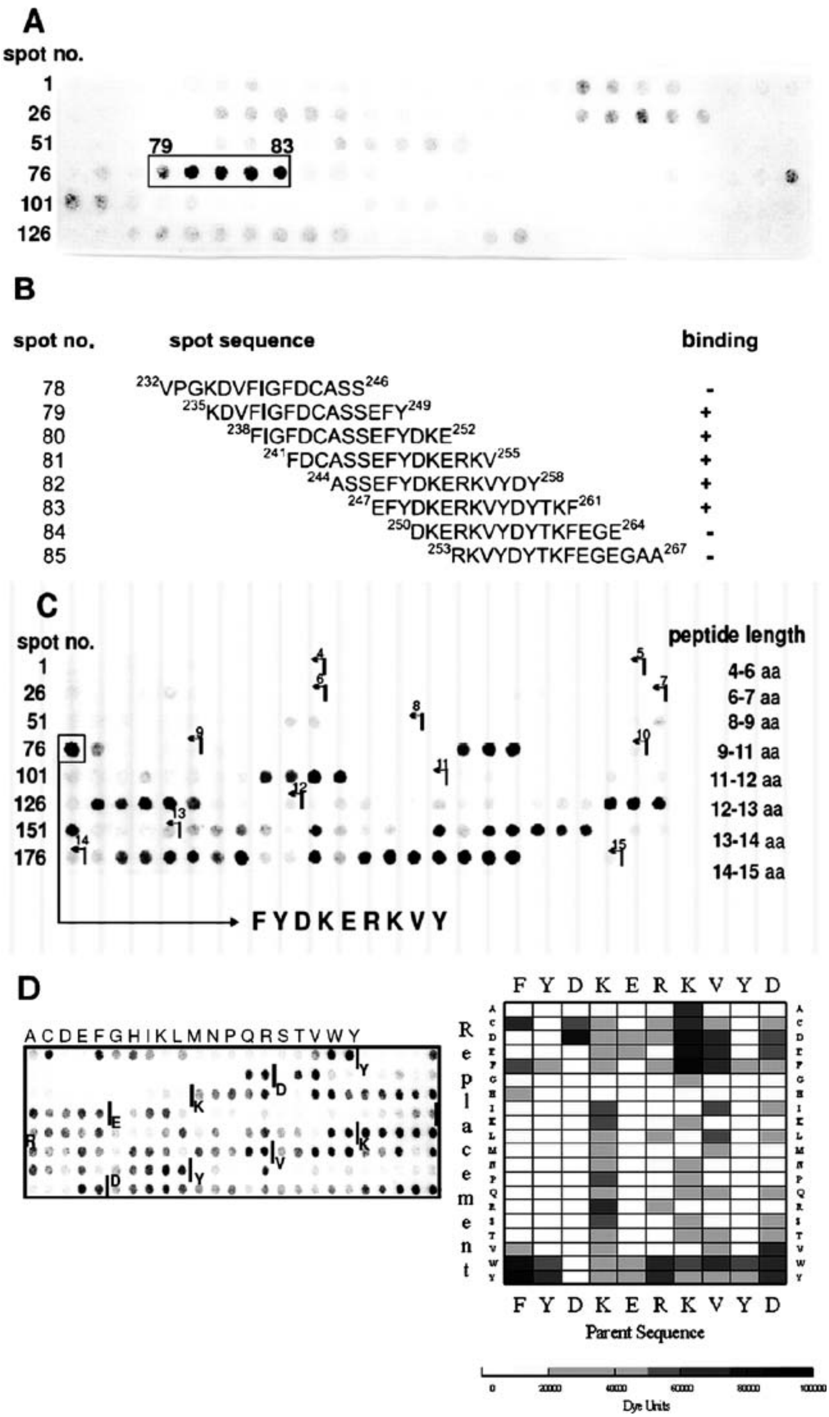
peptides both as solid phase bound and solution phase products for systematic epitope analysis. Each peptide is synthesized at a distinct site (spot) on a porous membrane. The final array of cellulose-bound peptides can be directly probed for protein binding in a western blot type overlay process. Membrane-bound peptide arrays manufactured by SPOT-synthesis are ideally suited for rapid screening through many protein sequences, however, they can only be reused a few times. Furthermore, SPOT-synthesis on porous membranes has its limitations when reducing the spot size below 1 mm and becomes costly and tedious when large numbers of copies of an identical array are required. We therefore have developed a special add-on to the SPOT-synthesis process for manufacturing and application of synthetic peptide/compound repertoires in the form of chemical mini- or, more sophisticated, microarrays. These maintain the advantageous features of cellulose-bound probe molecules but allows massive miniaturization and multiplication [11].

The process for manufacturing mini- or microarrays of synthetic peptides adds more experimental steps to the whole manufacturing process and, thus, requires additional effort. Therefore, this is only reasonable if indeed a great number of copies of the same array are required for a larger series of experiments such as profiling serum collections, hybridoma clones, genome spanning protein families (domains, kinases, etc.), recombinant protein variants, or for providing generic peptide libraries. The new chemical mini- and microarrays perform as reliably as the original, successful macro-SPOTs system on cellulose membranes. Single experiments or only a few serial experiments still should be preferably performed with the macroarray format, except if a limiting amount of sample makes miniaturization a strict prerequisite as with samples from small model organisms (like mouse or worm), tissue from patients or a few sorted cells; in this case, the surplus of effort is certainly justified.

2 Peptide Synthesis Strategies for Interaction Studies

2.1 Protein–Protein Interactions

In preclassifying a protein interaction as being mediated via a linear interaction site, it is a good indication if its interaction with a binding partner is detectable in a type of western blot analysis after denaturing SDS-PAGE; check both orientations of the analysis (protein A denatured and probed with protein B and vice versa). The thorough investigation of an interaction domain or an entire protein using SPOT peptide arrays involves three subsequent steps (Fig. 3 [12]). First, 15- to 20-mer peptide fragments, covering the



spot no.	spot sequence	binding
78	^{232}VPGKDVFIGFDCASS^{246}	-
79	^{235}KDVFIGFDCASSEFY^{249}	+
80	^{238}FIGFDCASSEFYDKE^{252}	+
81	^{241}FDCASSEFYDKERKV^{255}	+
82	^{244}ASSEFYDKERKVYDY^{258}	+
83	^{247}EFYDKERKVYDYTKF^{261}	+
84	^{250}DKERKVYDYTKFEGE^{264}	-
85	^{253}RKVYDYTKFEGEGAA^{267}	-

◀ **Fig. 3** An example for a classical binding site (epitope) analysis experiment for the analysis of the minimal binding site motifs of human plasmin(ogen) on α-enolase from *Streptococcus pneumoniae* [12]. **A** Membrane-bound array of 141 overlapping peptides of 15 amino acids each, with an offset of 3 amino acids, covering the 434 aa sequence of α-enolase was analyzed for binding to human plasminogen. Specific binding was detected for spots 79 to 83. Weak reactivity of other spots was due to non-specific binding of anti-plasminogen antibody and secondary antibody used. **B** Sequences of spots (78 to 85) and reactivity with plasminogen. **C** A membrane-bound array of 198 overlapping peptides, with peptide lengths from 4 up to 15 amino acid residues and an offset of only one residue analyzed with human plasminogen. The α-enolase sequence covered ranged from position 232 to 267 (VPGKDVDCASSEFYDKERKVYDYTKFEGEGAA). The peptide FYDKERKVY (spot 76) located between position 248 and 256 was identified as the minimal binding site for plasminogen, although there is some influence from neighboring N-terminal amino acid residues. The length of the peptides spotted on the membrane is indicated by numbers and arrows. **D** The spectral diagram display (*right side*) shows the results from the replacement scan membrane (*left side*). Each square represents the spot intensity obtained from the respective replacement peptide. In this particular example, residues Y2, D3, E5, Y9 are more critical for selective recognition of the binding site, a fact which would have escaped a "glycine or alanine walk" analysis. Dye units have an arbitrary scale for relative intensities

sequence with an offset of 3 to 5 amino acids, are synthesized to locate the binding site (Fig. 3A, B). Thus, a protein of 1000 amino acid residues (about 120 kd in size) is covered by 200 to 350 peptide fragments. If the capacity of the SPOT-synthesis is limited, for example when performing manual synthesis, it may be helpful to first narrow down the interaction site to a fragment or subdomain by using other experimental methods, for example by probing the interaction with deletion mutants applied in pull down assays. For shorter proteins or interaction domains, peptides with only one amino acid offset can be used as a starting point. When polyclonal sera are analyzed which may recognize overlapping epitopes bound by different antibodies, it can be helpful to shorten the length of the peptides to distinguish between the overlapping epitopes [13].

In a second step (sizing), the core binding motif is determined. A binding site identified in the previous mapping experiment is further characterized by using a series of overlapping peptides having an offset of only one amino acid and a stepwise reduced size (Fig. 3C). Ideally, only peptides carrying the core binding motif will react and one of the series will reveal one single spot which corresponds to the peptide with the minimal epitope. However, more complex results can be obtained when the core residues are not present in a single contiguous sequence. Furthermore, some proteins will bind with detectable affinity to several small "subepitopes".

The third step (analoging) determines the contribution of single amino acid side chains. In the past, this task was frequently addressed by "glycine- or alanine-walks", i.e., by exchanging only one amino acid per peptide with glycine or alanine, resulting in a small set of point mutated peptides covering the epitope. However, much more information and confidence on every amino acid position can be gained with a full replacement study. For this, a set of peptides with systematic single replacements of every amino acid residue in the core peptide sequence by all other genetically coded amino acids is probed (Fig. 3D). Obviously, analoging can be applied directly to an initial peptide hit if capacity permits, e.g., 15-mer full replacement amounts to 300 peptide spots.

Data obtained from these approaches can be used in a variety of follow up experiments. An independent verification of the interaction site identified is strictly recommended and can be pursued by either the construction of mutant protein analogues lacking essential residues of the core binding site or by the use of the identified peptides as soluble products in competition experiments. Amino acid replacement profiles of interaction sites as obtained in step 3 can also be used to search databases for potential cross-reactivity with other proteins. This is of particular importance for the identification of immunological cross-reactivity with antigens from pathogens, for example. Further analyses look at the contribution of post-translational modifications (e.g., phosphorylation or glycosylation) which can be studied by synthesizing sets of peptides with identical sequences, but different side chain modifications [14]. For binding motifs located at the N-terminus of a protein, the contribution of the N-terminal amino group to binding affinity can be assessed by synthesizing peptides with amino acid residues added in front of the N-terminal amino acid [15]. Additionally, using a modified linker strategy, peptides can also be presented with a free carboxy-terminal end; this is required for some interactions such as the binding to the PDZ domain [16].

If no linear peptide can be identified by peptide scanning, the interaction is most obviously truly conformationally defined. Assembly of branched double or triple peptide combinations on a single spot have been reported to be successful [17]. Linear peptide mimotopes may be identified by screening generic complete peptide libraries [18]. Strategic arrays of peptide pools that cover full libraries with several billions of peptides can be easily prepared for this purpose through incorporation of amino acid mixtures (Table 1). An example of a peptide library array probed with a monoclonal antibody is shown in Fig. 4 [20]. The amino acid residues of a mimotope sequence may be a guide to locate the conformational epitope in the primary sequence or, if available, the 3D-structure of the protein [21]. Peptide libraries have also been applied in studying peptide sequence preferences in protein–protein interactions [22] or enzyme substrate specificity [23, 24]. Furthermore, peptides selected from libraries as ligands/inhibitors for proteins that usually do not bind other peptides/proteins have been developed as tools in molecular biology such as the Strep-tag peptide binding to the biotin pocket of streptavidin [25]. Other approaches for a priori delineation of peptides include the screening of large series of individual, randomly selected peptides [26].

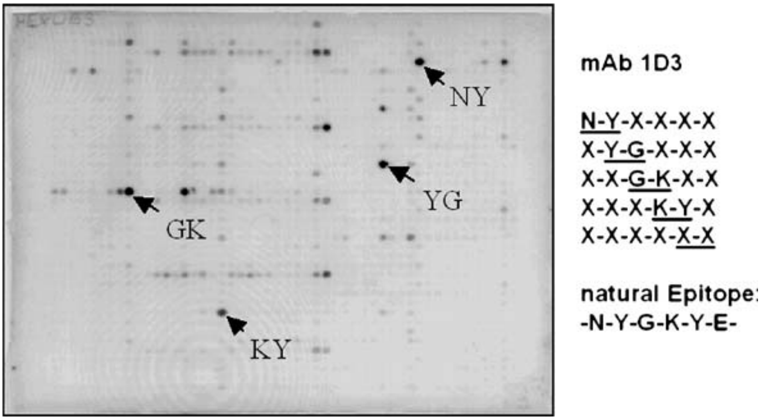


Fig. 4 A dual-positional scanning peptide library experiment (corresponds to Table 1, entry 3) with monoclonal antibody 1D3. Signal development was carried out with an alkaline phosphatase conjugated secondary antibody and the BCIP/MTT reagents. The results of this experiment allows for the determination of the epitope sequence a priori from the overlapping dipeptide signals [20]. The tyrosine (Y5) and glutamic acid (E6) in the natural epitope do not contribute significantly to specific recognition as they can be replaced by almost any other residue (data from the replacement analysis not shown)

Table 1 Strategies for the delineation of peptide sequences by activity screening of random peptide pools[a]

1) Iterative search starting with one or more defined positions, e.g., according to Geysen et al. [2]		
First generation	X-X-3-4-X-X	400 pools (Each 160 000 sequences)
Second generation	X-2-0$_3$-0$_4$-5-X	400 pools (Each 400 sequences)
Third generation	1-0$_2$-0$_3$-0$_4$-0$_5$-6	400 pools (Each 1 sequence)
2) Positional scanning with single fixed positions, one single screen according to Dooley and Houghten [19], see also Rodriguez et al. [24]		
1-X-X-X-X-X	20 pools	(Each 3.2×10^6 sequences)
X-2-X-X-X-X	20 pools	(Each 3.2×10^6 sequences)
X-X-3-X-X-X	20 pools	(Each 3.2×10^6 sequences)
X-X-X-4-X-X	20 pools	(Each 3.2×10^6 sequences)
X-X-X-X-5-X	20 pools	(Each 3.2×10^6 sequences)
X-X-X-X-X-6	20 pools	(Each 3.2×10^6 sequences)
3) Dual-positional scanning, one single screen according to Frank et al. [20]		
1-2-X-X-X-X	400 pools	(Each 160 000 sequences)
X-2-3-X-X-X	400 pools	(Each 160 000 sequences)
X-X-3-4-X-X	400 pools	(Each 160 000 sequences)
X-X-X-4-5-X	400 pools	(Each 160 000 sequences)
X-X-X-X-5-6	400 pools	(Each 160 000 sequences)

[a] Special codes to describe the pool compositions are 0_n = unvaried position in a particular screen occupied by single amino acid residues; $1, 2, 3\ldots$ = positions systematically varied by single amino acid residues in a particular screen; X = position occupied by a set of (e.g., all 20 L-) amino acid residues

Moreover, the chemistry allows incorporation of modified or not naturally occurring amino acids as well as artificial linkages. An overview is given in [9, 27]. See also the assembly of non-peptidic small organic molecules in [11].

2.2 Other Protein Ligand Interactions

Besides the molecular analysis of protein binding sites and antibody epitopes, as will be described in detail in the protocols below, a plethora of other protein interactions has been reported which were studied with modifications of the protocol steps specifically adopted to the particular requirements of the assays. These include:

- Mapping and analysis of T-cell epitopes through either MHC-binding or T-cell stimulation
- Enzyme substrate analysis and inhibitor design
- Protein/peptide interactions with nucleic acids
- Peptide nucleic acid (PNA) interactions with nucleic acids
- Peptide interactions with small ligands
- Chemical/enzymatic transformation of immobilized peptide
- De novo protein design
- Cell-based assay with cleaved, solution phase peptides

For a review on these applications see [27]. Moreover, solid phase peptide arrays can be used for affinity-capture of the protein of interest. For example, epitope specific antibodies can be isolated from polyclonal sera [28, 29], thus, combining monospecificity with the ease of rabbit serum preparation. More recent novel options include, e.g., the preparation of miniprotein (protein domain) arrays by combination of solid phase synthesis and chemical ligation [30] or the multiplexed biopanning of phage libraries for genome wide protein interaction mapping [31].

3 Guidelines for SPOT Peptide Synthesis

3.1 General Principle

The principle of the method involves each amino acid being added to a growing peptide chain, a coupling reaction started by dispensing a small droplet of the reaction mixture onto the membrane. The droplet gets absorbed and forms a circular spot. Using a solvent of low volatility containing activated amino acid monomers, such a spot forms an open reactor for chemical conversions involving reactive functions anchored to the membrane support, comparable to conventional solid phase synthesis. A large number of separate spots can be arranged as an array on a larger membrane sheet and the intermediate areas are chemically inactivated by acetylation. Each of these spots then can be separately manipulated by manual or automated delivery of the corresponding reagent solutions (Fig. 5). The volume dispensed and the absorptive capacity of the membrane determine the spot size, which can be adjusted to control the scale of synthesis. The spot size also controls the minimal distance between spot positions and thereby the maximum density of the array. Synthetic steps common to all spot reactors are carried out by washing the whole membrane with respective reagents and solvents. Fully automated instruments place the membranes on a porous plate and remove the reagent and solvent excesses by vacuum suction.

Because of their hydrophilic nature, cellulose membranes are particularly well suited for the presentation of immobilized peptides to a biological assay system. After SPOT-synthesis of the peptide array and incubation of the membrane with protein ligands, detection of proteins bound to individual spots is done in a manner analogous to an immunoblot (Fig. 2, top). Unspecific binding of biomolecules has only rarely been reported for the assay conditions given below. Assembly of peptides on the surface of the cellulose fibers in the membrane yield a quite high local concentration which allows for the capture of rather low affine binders (up to several 100 μM was reported [32]).

As long as the biological assaying of these macroarrays does not irreversibly transform the peptide probes, they can be reused many times upon stripping off all biologicals from the assay experiment. Depending on the biological assay, this stripping, however, can be quite insufficient and thus one

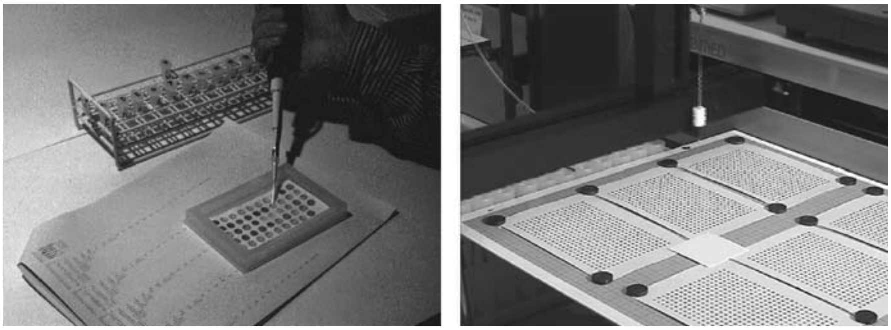


Fig. 5 (*left*) The minimal experimental set-up for manual SPOT-synthesis showing the reaction tray with a membrane displaying *blue* (*dark*) stained amino spots and *yellow* (*light*) stained coupled spots, the rack of small tubes containing the activated AA derivatives, the computer-generated listing of the pipetting operations and the manually operated pipette. (*right*) Platform of the AutoSpot robot showing six cellulose membranes (8 cm × 12 cm) with each an array of 17 × 25 (425) spot reactors

array may only be usable once. Furthermore, one synthesized array can be processed only serially through a set of experiments, which requires several identical synthetic arrays to proceed in parallel. Third, a conventional SPOT type array has rather large dimensions (min. 2 mm spot distance) and therefore we call it a macroarray; this requires considerably large volumes for the assay and, thus, the amount of available sample can become limiting.

We describe an easy add-on process that overcomes these limitations of cellulose membrane-bound macroarrays produced by SPOT-synthesis by transferring a synthetic membrane-bound macroarray to a multitude of microscope slide bound mini- or microarrays. The manufacturing of the peptide macroarray follow essentially standard SPOT-synthesis protocols with an array format adapted from the 384-well microtiter plates, except that a special, acid sensitive amino-cellulose membrane is used. Individual spots are separated post-assembly with the help of a 384-compatible punching device which delivers the cellulose-compound conjugate disc segments of 3 mm diameter into the wells of four 96-deepwell plates (Fig. 6). Then, the discs are treated with a TFA cocktail containing >80% TFA plus scavengers, as used in routine solid phase peptide synthesis. This treatment solubilizes the support itself with the compounds still covalently attached and simultaneously cleaves the acid sensitive side chain protecting groups. Precipitation with ether removes the bulk of acid together with the cleavage chemistry and the dried precipitate is then dissolved in DMSO. After appropriate dilution with DMSO, minute aliquots of these solutions of compound-support conjugates are transferred (printed) and adsorbed onto the target planar surfaces, usually glass microscope slides, with the help of a suitable pipetting device. We therefore call this process spotting compound-support conjugates: SC^2. One standard cellulose disc segment yields 0.5 mL of DMSO stock solution from which only nanoliter to picoliter aliquots can be used to print up to 10^6 mini- or 10^8 microarray copies.

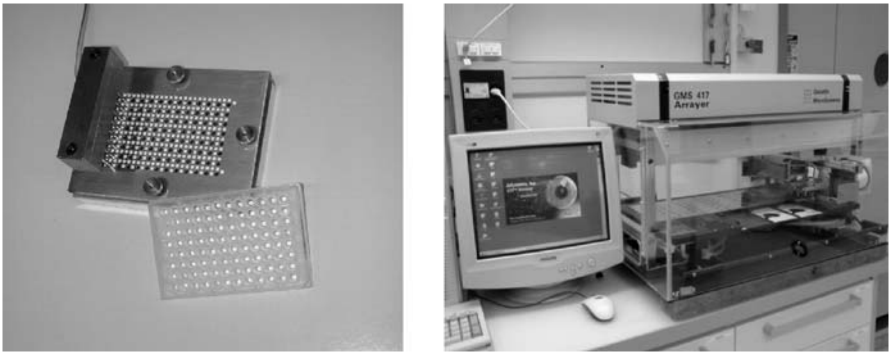


Fig. 6 (*left*) The 384→96 punching device used to separate peptide spots for the SC^2-process. (*right*) The model GMS 417 ring and pin Array Printer

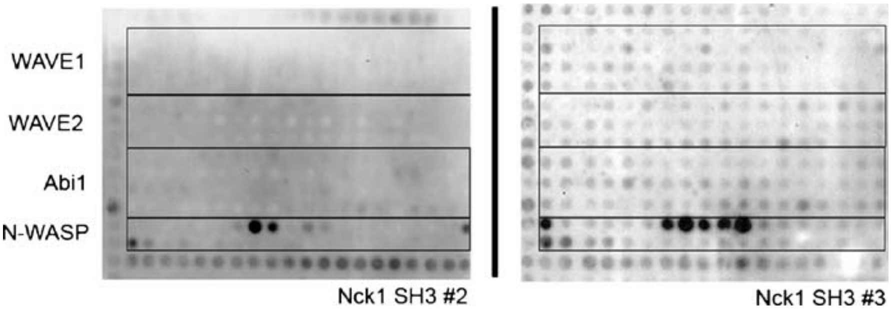


Fig. 7 SH3 domains are involved in many cellular signaling processes. They bind to specific proline-rich regions in target proteins. Nck1, an adaptor protein consisting of three SH3 and one SH2 domain, is known to bind to the N-WASP protein involved in actin polymerization and recruits it to sites of tyrosine phosphorylation. The 130 proline-rich peptide fragments of proteins indicated at the *left* were spotted as miniarrays onto glass slides and probed with recombinant SH3 domains (GST-fusions) coming from a series of proteins involved in actin-skeleton remodeling. Bound SH3 domains were detected using HRP-coupled anti-GST antibodies and chemiluminescence exposure of X-ray film

This SC^2-process maintains most of the beneficial properties of a cellulose-bound peptide array, in particular the low background binding and the high local peptide concentration. Thus, the new chemical mini- and microarrays perform as reliably as the original, successful macro-SPOTs system on cellulose membranes. This is demonstrated with the probing of low affinity recombinant SH3 domains with an array of polyproline-rich peptides utilizing exactly the same conditions as on a membrane array (Fig. 7).

3.2
Brief Introduction to Solid Phase Chemical Peptide Synthesis

Assembly of a peptide chain by chemistry starts at the C-terminal end (in contrast to biological synthesis, where the ribosome starts at the N-terminus). But peptide chemists write a peptide sequence following the same convention, this is the N-terminus on the left and the C-terminus at the right.

The amino acid building blocks are specially modified amino acids that carry protecting groups which assure a directed step-by-step assembly of the peptide chain. All reactive chemical functionalities at the amino acid side chains are blocked permanently throughout the whole assembly phase and are only removed in a final deprotection treatment. The terminal carboxylic acid function of the amino acid remains free and is chemically activated forming an active ester which then reacts in a coupling reaction to yield an amide bond (peptide bond) with a free terminal amino function of a growing peptide chain. The amino function of the amino acid is blocked by a temporary protecting group, which prevents self-coupling with its own activated carboxyl function.

Figure 8 outlines the basic steps in solid phase peptide assembly as they will be used in this section. The solid support material presents free amino functions covalently attached to its surface. The manufacturing of respective supports requires more sophisticated chemical expertise and equipment and,

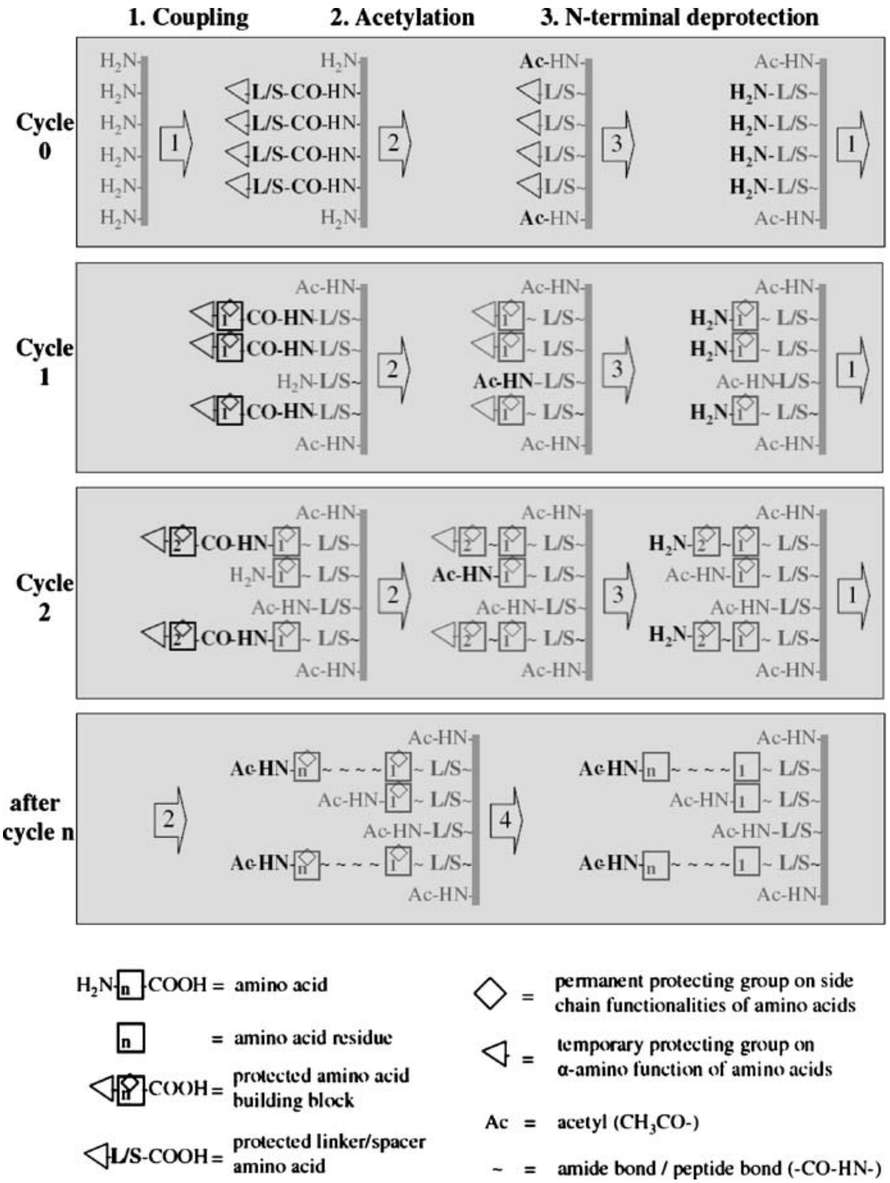

1. Coupling
2. Acetylation
3. N-terminal deprotection
Cycle 0
Cycle 1
Cycle 2
after cycle n
H2N-
L/S-CO-HN-
Ac-HN-
L/S-
H2N-L/S-
CO-HN-L/S-
Ac-HN-L/S-
1
2
3
4
n
H2N-[n]-COOH = amino acid
[n] = amino acid residue
protected amino acid building block
protected linker/spacer amino acid
= permanent protecting group on side chain functionalities of amino acids
= temporary protecting group on α-amino function of amino acids
Ac = acetyl (CH3CO-)
~ = amide bond / peptide bond (-CO-HN-)

Fig. 8 Outline of the series of chemical transformations during peptide assembly on a solid phase. Changes introduced at each step are highlighted in *black*

thus, the use of a quality checked commercial material is recommended. One cycle of amino acid addition consists of three steps which are repeated as often as amino acid residues are required for the target peptide sequence. Step 1 is the coupling of the protected amino acid building block to the free terminal amino functions presented on the support. All amino functions that have not reacted in step 1 will be blocked in step 2 by acetylation (also called capping reaction). This prevents these functions from reacting in a later step of the assembly process, which will result in peptide contaminants of wrong (deletion) sequences. Capping assures that inefficient coupling only gives rise to truncated peptides which are still fragments of the correct sequence. Then the N-terminal protecting groups at the growing peptide chains are removed in step 3 releasing the free amino function of the last coupled amino acid ready for the next coupling reaction. Cycle 0 is optional but recommended for array generation and involves the incorporation of a spacer or linker molecule preferably using the same chemistry as for amino acid coupling. A spacer amino acid increases the distance of the peptide to the support surface and will enhance access to these peptides by large protein acceptors. A linker amino acid introduces a special chemical entity with a bond that can be cleaved selectively after the assembly is completed to release the peptide from the support into solution. For more information see textbooks such as that from Chan and White [33]. After the final coupling cycle (*n*), the peptides are acetylated and in step 4 all side chain protecting groups are removed.

3.3
The SPOT Membrane Support

The membrane supports are of specially manufactured primary amino-cellulose paper and are optimized for proper performance in synthesis and bioassay. A large variety of such synthesis membranes are commercially available. Amino-PEGylated membranes are primarily used for the preparation of immobilized peptide arrays resulting in a very stable C-terminal attachment of the peptides. The recently described SynthoPlan APEG CE cellulose membrane (AC-S01 grade) is particularly stable against trifluoroacetic acid used in the final deprotection step and prolonged treatment (overnight) improves the peptide quality considerably [34]. Conversely, the SC^2-process requires an acid soluble membrane: membrane for CelluSpots from Intavis Bioanalytical Instruments AG, Cologne, Germany.

High quality arrays of spots providing suitable anchor functions for peptide assembly on cellulose membranes are most easily generated by spot-wise coupling to an evenly aminated membrane of a spacer Fmoc-amino acid such as Fmoc-β-alanine (Fig. 8, cycle 0). During this derivatization cycle, the array of spot reactors is generated and all residual amino functions between spots are blocked by acetylation (step 2). This array formation process requires very accurate pipetting. During peptide assembly (cycles 1 to *n*), slightly larger volumes are dispensed and the wettened areas then exceed those initially formed in order to avoid incomplete couplings at the edges.

The flexibility of SPOT-synthesis enables the investigator to easily vary the number of spots, the format of the array and the scale of each peptide synthesized. The arrays on the membrane supports are freely selectable to fit the individual needs of the experiment by variation of paper quality, thickness, specific anchor, loading and spot size [10], but this exercise goes beyond the protocol. Figure 9 demonstrates some array configurations made on the recommended AC-S01 paper membrane. The standard format used in manual SPOT-synthesis was adapted to the 8×12 array of a microtiter plate with 96 spots. However, to fully exploit the scope of the method, use of an automated SPOT-synthesizer such as the AutoSpot or MultiPep robot (from Intavis) is recommended. The AutoSpot instrument can handle up to four standard membrane sheets simultaneously or a whole DINA4 sheet of 210 mm×297 mm (Fig. 5, right). Moreover, automated spotting can be exploited to reduce the size of spots and, thus, increase the number of spots per area considerably. A 384 format (16×24) can be generated with, e.g., 0.1 μL spotting volume. A robust standard array format to fit the 8 cm×12 cm size of the membrane comprises 17 rows of 25 spots/row (425 spots). However, up to 2500 spots can be generated on the same standard membrane by pipetting as little as 30 nL volumes. This instrument only performs the pipetting work; all washing steps are carried out manually. The newer MultiPep instrument can perform fully automated SPOT-synthesis and has two types of membrane platforms, one for two standard sheets and one in the DINA4 format. Alternatively, MultiSynTech GmbH, Witten, Germany, offers an auxiliary tray for fully automated SPOT-synthesis on its Syro robots.

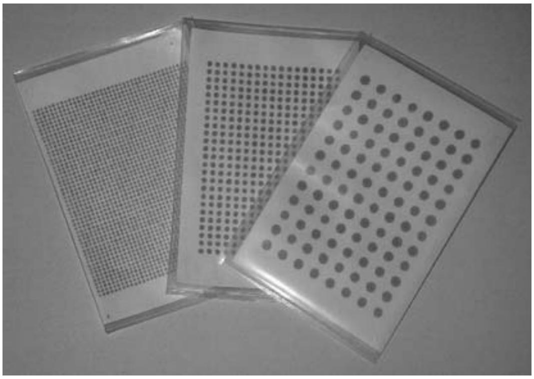


Fig. 9 Three array formats fitting onto an AC-S01 membrane of 8 cm×12 cm: 8×12 spots in the conventional microtiter plate format for manual spotting (spot distance 9 mm, spot volume 0.5 μL); 17×25 spots (spot distance 4 mm, volume 0.1 μL); 40×50 spots (spot distance 2 mm, volume 30 nL)

3.4 Peptide Assembly

Chemical and technical performance of this type of simultaneous parallel solid phase synthesis allows for the reliable assembly of arrays of peptide sequences up to a length of 20 amino acid residues utilizing conventional mild Fmoc/*t*Bu chemistry [35]. This chemistry employs the Fmoc protecting group for the amino function of the amino acid building blocks which is removed by treatment with piperidine base. The side chain functionalities are permanently protected with *tert*-butyl type groups which are removed only at the end of the synthesis by treatment with trifluoroacetic acid. Much longer peptide sequences are reported [36, 37] but the quality of such peptides strongly depends on the particular sequence and need to be preevaluated by case to case studies.

Free amino functions on the spots can be visualized by staining with bromophenol blue [38] after N-terminal deprotection (Fig. 8, step 3) and prior to the coupling reaction (step 1). This color staining allows the visual monitoring of the proper performance of all synthesis steps such as correct dispensing, quantitative coupling and acetylation (capping), and effective removal of piperidine from the Fmoc-deblocking steps. Thus, a standard membrane used for SPOT-synthesis displays an array of light blue spots on a white background (Fig. 9). In low density arrays, each spot may be marked by writing a number in pencil next to it. These numbers refer to the corresponding peptide sequences that are assembled on these spots and are a guide for rapid manual distribution of the solutions of activated amino acid derivatives at each elongation cycle (Fig. 5, left). For automated pipetting, no pencil marking is necessary as exact positioning of the membranes is assured by the perforation for the holder pins in the robot. The dry membranes are placed in a flat, chemically resistant trough or fixed on the platform of the synthesizer. As soon as the droplets of activated amino acid solutions are added to the spots, coupling proceeds with a conversion of free amino groups to amide bonds. After all amino groups have been consumed, the blue color of the spots changes to yellow indicating a quantitative reaction. The physicochemical properties of the growing peptide chains are very different, sometimes unfavorable and can slow down or even hinder the quantitative coupling of an amino acid building block. This is an inherent problem of solid phase peptide synthesis, however, it will be visible on the membrane when some of the spots keep their blue color and, thus, can be monitored and documented for later interpretation of results. The solvent within the spots slowly evaporates over the reaction time. For AC-S01 membranes, after approximately 15 min a spot is dry and the reaction will stop. However, additional drops may be added onto the same position without enlarging the spots and risking overlap with their neighbors. In this way, difficult coupling reactions can be pushed towards completion by double or triple couplings. Fully automated

SPOT-synthesis cannot profit from bromophenol blue staining because the instruments have no electronic image monitoring. However, we recommend to check regularly the quality of a synthesis by staining, for example every first cycle of the day. Nevertheless, non-reacted termini will be acetylated after each coupling step so that no false sequences will contaminate the product.

The introduction of randomized positions (X) within a peptide sequence assembled on a spot in order to prepare arrays of defined peptide mixtures (or pools) is quite reliably achieved by coupling with equimolar amino acid mixtures and applying these at a submolar ratio with respect to available amino functions on the spots [39, 40]. This is to allow all activated derivatives (also the slower coupling ones) to react quantitatively during a first round of spotting. All coupling reactions are then completed by three to four successive repeats of spotting. Using this coupling procedure, any position in a peptide sequence can easily be randomized without special considerations or increase in technical effort. Some current strategies for the delineation of peptide sequences by activity screening of random pools are given in Table 1.

4 Guidelines for Preparing Peptide Arrays by SPOT-Synthesis

The following protocol describes the parallel chemical synthesis of short linear peptides or peptide pools as arrays on modified cellulose membranes. Peptides are synthesized starting from their C-termini using Fmoc-amino acid derivatives. After completion of the assembly steps, acid-stable membranes are processed to cleavage of all side chain protecting groups, after which the peptide array is ready to be probed with the potential interaction partners. Alternatively, peptides synthesized on acid soluble membranes are separated after the assembly steps with the help of a puncher and then treated according to the SC^2-process described in Sect. 5.

4.1 Materials

- SPOT membranes. Acid-stable AC-S01 type amino-PEGylated membranes (manufactured by AIMS Scientific Products GmbH, Braunschweig, Germany) are recommended and available from AIMS itself or from Intavis AG, Cologne, Germany. Please note that the AutoSpot instrument requires a special format of the membranes with special perforation for the holder pins on the robot.
  Acid soluble membranes for CelluSpots are available from Intavis AG (order No. 32.105).
- Chromatography paper type 3MM (Whatman, Maidstone, UK).

- Bromophenol blue indicator. Prepare a stock solution of 10 mg per mL in DMF and keep at RT. This BPB stock should have an intense orange color and should be discarded when the color has turned to green.
- *N*,*N*-dimethylformamide (DMF). This should be free of contaminating amines and thus of highest affordable purity, such as the peptide synthesis grade DMF of Biosolve BV, Valkenswaard, The Netherlands. Amine contamination is checked by the addition of 10 μL of BPB stock to 1 mL of DMF. If the resulting color is yellow, then this batch can be used without further purification. Check each new batch.
- 1-Methyl-2-pyrrolidinone (NMP). This should be of highest purity available. Amine contamination is checked as above for DMF. If the resulting color is yellow, then the NMP can be used without further purification. Most commercial products, however, are not acceptable. To prepare a suitable quality, treat 1 L of NMP with 100 g of acidic aluminum oxide under constant vigorous shaking at RT overnight. Then, a 1 mL aliquot should give a yellow BPB test. Filter the slurry through a bed of dry silica gel (for flush chromatography, Mallinckrodt Baker BV, Deventer, The Netherlands) in a closed glass filter funnel (slight nitrogen pressure can speed up the process, but is not necessary). Divide the clear liquid into 100 mL portions and store tightly closed at –20 °C.
- *N*-hydroxybenzotriazole (HOBt). Anhydrous, ISOCHEM, Vert-Le-Petit, France. Store tightly closed at room temperature in a dry place.
- *N*,*N*′-diisopropylcarbodiimide (DIC), ≥98%.
- D(+)-Biotin, 99%, from Carl ROTH, Karlsruhe, Germany (order No. 3822.1).
- Fmoc-AA stock solutions. Fmoc-amino acid derivatives of all 20 L-amino acids as well as *β*-alanine and other special amino acid derivatives are available from several suppliers in sufficient quality (Novabiochem/Merck Biosciences, Schwalbach, Germany, or Bachem, Bubendorf, Switzerland). Side chain protecting groups should be Cys(Acm) or Cys(Trt), Asp(OtBu), Glu(OtBu), His(Trt), Lys(Boc), Asn(Trt), Gln(Trt), Arg(Pmc), Ser(tBu), Thr(tBu), Trp(Boc), and Tyr(tBu). It is necessary to prepare HOBt-esters of these amino acid derivatives in NMP for use throughout in spotting reactions. Dissolve 1 mMol of each Fmoc-AA in 5 mL NMP containing 0.25 M HOBt to give 0.2 M Fmoc-AA stock solutions. These stocks are kept in 10 mL plastic tubes that are closed tightly, flush frozen in liquid nitrogen, and stored at –70 °C. For use in coupling reactions with amino acid mixtures at randomized positions (X) in the peptide sequences, combine equal aliquots of Fmoc-AA stock solutions for the respective amino acids to be incorporated, dilute with threefold volume of NMP to give 50 mM solutions and store as described above.
- Special chemical derivatives. Free thiol functions of cysteine may be problematic because of post-synthetic uncontrolled oxidation. To avoid this, you may replace Cys by serine (Ser), alanine (Ala) or *α*-aminobutyric acid

(Abu). Alternatively, choose the hydrophilic Cys(Acm) and leave protected. For the simultaneous preparation of peptides of different size with free amino terminus, couple their terminal amino acid residues as αN-Boc derivatives so that they will not become acetylated during the normal elongation cycle. Boc is then removed during the final side chain deprotection procedure. Negative and positive control spots for the arrays on glass slides are very helpful. We have good experiences with β-alanine (see above) as negative and biotin as positive controls. Thus, always include some of these spots in your array. Biotin at 0.2 M is rather insoluble in DMF, but dissolves upon activation with DIC, just give more time for activation.

- Acetylation mix. This is a 2% solution of acetic anhydride ($\geq$99.5%) in DMF.
- Piperidine mix. This is a 20% solution of piperidine ($\geq$99%) in DMF. Please note that piperidine is toxic and should be handled only with gloves under a hood.
- Alcohol (methanol or ethanol) of technical grade (95%).
- Deprotection mix. This is trifluoroacetic acid (TFA, synthesis grade), dichloromethane (DCM), triisobutylsilane (TIBS) and water in a ratio of 80% TFA, 12% DCM, 3% TIBS and 5% water (mix in this order!). Please note that trifluoroacetic acid is very harmful and volatile, and should be handled with gloves under a hood.

4.2
Special Equipment

All equipment used for membrane synthesis should be resistant to organic solvents. Glassware or polypropyleneware should be exclusively used in all steps involving organic solvents. Standard micropipetting tips (Gilson, Eppendorf) can be employed.

- SPOT-synthesis kit. Software for the generation of peptide lists and pipetting protocols are included in the synthesis kit and in the operation software of the spotting robot. A freeware package is available from the authors.
- Flat reaction/washing troughs with a tightly closing lid made of chemically inert material (glass, teflon, polypropylene) with dimensions slightly larger than the membranes used.
- A spotting robot, model AutoSpot or MultiPep peptide synthesizer with spotting tray (Intavis AG).
- 1.5 mL plastic tubes (e.g., Eppendorf, safe twist) and appropriate racks as reservoirs for amino acid solutions.
- A rocker table.
- Two dispensers for DMF and alcohol adjustable from 5 to 50 mL. Handheld hair dryer with non-heating option.

- Appropriate bench space in a hood.
- A –70 °C freezer.

4.3 Methods

All volumes given below are for one standard AC-S01/CelluSpots membrane paper sheet of 8 cm × 12 cm and have to be adjusted for more sheets, or other paper qualities and sizes. Solvents or solutions used in washing and incubation steps are gently agitated on a rocker table at room temperature if not otherwise stated and are decanted after the time indicated. During incubations and washings the troughs are closed with a lid.

4.3.1 Preparative Work

1. Generate a list of peptides to be prepared. You may combine more than one list. Add them one after the other to fill up a complete array. The peptides can be separated after synthesis by simple cutting the membrane into corresponding sections.
2. Select the array(s) required for the particular experiment according to number, spot size and scale. For manual spotting you should adhere to a 8×12 format (spot distance 9 mm; spot volume 0.5 μL for array generation in cycle 0, 0.7 μL for elongation cycles). An array of 17 rows with 25 spots each (spot distance 4 mm, volume 0.1 μL during array generation and 0.2 μL for elongation cycles) is recommended for the AutoSpot.
3. Calculate the volumes of Fmoc-amino acid solutions required for each derivative and cycle; consider that a triple coupling procedure may be necessary and that each vial should contain a minimum of 50 μL. For example, in your list of peptides, alanine is required for twenty six peptides at cycle 1 and you will use a 17×25 array. Then for A1 you will need $26 \times 0.2 \times 3 = 15.6$ μL of Fmoc-Ala stock solution and you will take 50 μL for this vial. The SPOT software available can do this calculation for you.
4. Label a set of 1.5 mL plastic tubes with derivative and cycle code (e.g., A1) and distribute the Fmoc-amino acid stock solutions according to the calculated volumes required. Flush freeze in liquid nitrogen and store at –70 °C.

4.3.2 Generation of the SPOT Reactor Array

1. Mark the spot positions on the membranes with pencil dots for manual synthesis and place in the reaction trough. Alternatively, fix membranes on the platform of the SPOT robot for automated synthesis.

2. Take a 100 µL aliquot of the Fmoc-βAla stock from the freezer and bring to RT. Add 1 µL BPB stock. Add 4 µL DIC, mix, leave for 30 min and then spot aliquots (0.5 µL for 8×12 array or 0.1 µL for 17×25 array) of this solution to all positions according to the array configuration you have chosen. Let react for 60 min (cover the membranes on the spotter with glass plates). Please note for peptides longer than 20-mers it is recommended to reduce the loading of the spots by applying a mixture of the Fmoc-βAla stock and an *N*-acetyl-alanine stock (1 : 9). This will avoid molecular crowding of the larger peptide mass. You may also incorporate here a cleavable linker compound instead of β-alanine in order to cleave the peptides from the spots after assembly for solution phase assays; the safety-catch Frank-linker is recommended which yields peptides in physiological buffer solutions [41, 42].
3. Wash each membrane with 20 mL acetylation mix for 30 s, once again for 2 min and finally leave overnight in acetylation mix.
4. Wash each membrane with 20 mL DMF (three times for 10 min).
5. Incubate for 5 min with 20 mL piperidine mix.
6. Wash each membrane with 20 mL DMF (three times for 10 min).
7. Incubate each membrane with 20 mL of 1% BPB stock in DMF. Exchange the solution if traces of remaining piperidine turns the DMF solution into a dark blue solution. Spots should be stained only light blue!
8. Wash each membrane with 20 ml alcohol (three times for 10 min).
9. Dry with cold air from hair dryer between a folder of 3MM paper and store sealed in a plastic bag at –20 °C.

4.3.3 Assembly of the Peptides

1. Take the membranes from the previous step. Number the blue spot positions on the membranes with a pencil (H grade) for manual synthesis according to your peptide lists and place in separate reaction troughs. Alternatively, fix the non-numbered membranes correctly on the platform of the synthesizer. Number the membranes with a pencil and keep this arrangement through the whole synthesis. Note that you may now mark the cutting lines using a pencil for post-synthesis segmentation of the membrane into project specific sections. If bound protein will be eluted individually from single spot positions after having probed the spot membrane with a protein solution [28, 29] you should also mark the spots on those membranes used in automated synthesis. Pencil marking is quite stable during the synthesis procedure.
2. Take the set of Fmoc-amino acid stock aliquots for cycle 1 from the freezer, bring to RT and activate by addition of DIC (4 µL per 100 µL vial: ca. 0.25 M). Leave for 30 min. Then pipette aliquots of these solutions manually onto the appropriate spots on the membrane. Alternatively, place the

vials with the activated Fmoc-AA solutions into the corresponding location in the rack of the spotting robot and start cycle 1. Leave for at least 15 min. Repeat the spotting twice and then let react for 2 h (cover the membranes on the spotter with glass plates). If some spots stay dark blue, you may add additional aliquots. If most spots are yellow to green, then continue. Note that you add only 1 μL DIC to 100 μL Fmoc-AA mixture stock and repeat spotting four times for the efficient introduction of randomized X positions in the peptide sequences.

3. Wash each membrane with 20 mL acetylation mix for 30 sec and once again for 2 min. Then incubate a third time for about 10 min until all remaining blue color has disappeared.
4. Wash each membrane with 20 mL DMF (three times for 10 min).
5. Add 20 mL piperidine mix and incubate for 5 min.
6. Wash each membrane with 20 mL DMF (three times for 10 min).
7. Incubate with 20 mL of 1% BPB stock in DMF. Exchange the solution if traces of remaining piperidine turns the DMF solution into a dark blue solution. Spots should be stained only light blue! Due to the charge specific staining, BPB does not only bind to N-terminal amino groups. The side chains and protecting groups of other amino acids can strongly influence the staining intensity. The visible color of the peptides depends on the overall charge and therefore depends on the individual amino acid sequence.
8. Wash each membrane with 20 mL alcohol (three times for 10 min).
9. Dry with cold air from a hair dryer in between a folder of 3MM.
10. Start at step 2 for the next elongation cycle.

4.3.4 Terminal Acetylation

Synthetic peptides mimicking fragments of a longer continuous protein chain should be N-terminally acetylated to avoid an artificial charged terminus. Note that alternatively, special detection labels can be attached to the N-termini of peptides by spotting respective derivatives. This is useful, for example when peptides are applied as protease substrates and the enzyme activity followed through the change of the label upon cleavage of the peptide. We have successfully added biotin via its in situ formed HOBt-ester (normal activation procedure) or fluorescein via its isothiocyanate (FITC; 0.2 M) dissolved in DMF.

Continue after the final amino acid elongation cycle from the protocol above.

1. Incubate each membrane with 20 mL acetylation mix for at least 30 min until all remaining blue color has disappeared.
2. Wash each membrane with 20 mL DMF (three times for 10 min).

3. Wash each membrane with 20 mL alcohol (three times for 10 min).
4. Dry with cold air from a hair dryer in between a folder of 3MM.

4.3.5 Side Chain Deprotection of Membrane-Bound Peptide Arrays

After the peptide assembly is complete, it is necessary to remove all side chain protecting groups from the peptides. This must be performed under a hood as trifluoroacetic acid is very harmful! Note that this protocol is only applicable to acid-stable cellulose membranes such as AC-S01.

1. Prepare 40 mL of deprotection mix.
2. Place the dried membrane in the reaction trough, add deprotection mix, close the trough very tightly and agitate overnight. Note that this harsh treatment is required for complete cleavage of protecting groups [34, 43]. Cellulose membranes less resistant than AC-S01 will not survive.
3. Wash each membrane for 5 min with 20 mL DCM (four times).
4. Wash each membrane for 5 min with 20 mL DMF (three times).
5. Wash each membrane for 5 min with 20 mL alcohol (three times).
6. Wash each membrane for 5 min with 20 mL 1 M acetic acid in water (three times). Note that this is for removal of the Boc group from tryptophane.
7. Wash each membrane with 20 mL alcohol (three times for 5 min).

The membrane sheets may now be dried with cold air and stored at –20 °C or further processed as described in the next section.

5 Manufacturing Peptide Arrays on Glass Slides by the SC^2-Process

Compared to the synthesis of conventional peptide arrays on SPOT membranes described above (Sect. 4), there are a few but important changes. Use an acid soluble cellulose membrane (e.g., CelluSpots membrane from Intavis) for synthesizing the peptide array (see Sect. 4.1). The array dimensions should be 16×24 (384 spots at a distance of exactly 4.5 mm) to fit with the commercially available punching device (Fig. 6, left). After generation of the SPOT array with Fmoc-βAla and staining with BPB (Sect. 4.3.2), mark on every corner of the array at least 1 peptide spot position with a pencil by surrounding it. These marks are for adjusting later the membrane in the punching device. Follow the instructions described under Sects. 4.3.3 and 4.3.4 to assemble the peptides.

The dry membrane is then inserted in the punching device by placing it between two metal plates forming a plate-membrane-plate sandwich. Each metal plate has precisely drilled 3 mm holes mirroring the identical 384 spot grid of the cellulose membrane. One of four 8×12 tube racks is placed underneath the plates-membrane sandwich, and in this manner each cellulose-compound disc can be punched into its corresponding tube. Note that we recommend to use bar-coded Matrix tubes instead of deepwell microtiter plates which do seal better and can be handled also individually. You must adhere to a standard regime of correlating the four 96 daughter($'$) racks to the parent 384 array; we recommend the *z*-pattern, that is, A1 gets A$'$1, A2 gets A$''$1, B1 gets A$'''$1 and B2 gets A$''''$1.

Then, a strong acidic solution is added to each tube to both cleave the side chain protecting groups from the peptides and simultaneously dissolve the cellulose matrix to form a homogeneous solution. This acid treatment first swells the cellulose discs to double their previous thickness and usually after 0.1–24 h all cellulose-compound discs have disintegrated into a fine particulate suspension. Within 2–48 h most of the cellulose-compound discs should be completely dissolved. The solved cellulose-peptide conjugates are precipitated with ether, washed with ether and finally dissolved in DMSO to give the stock solutions. These stock solutions are stored at –70 °C, tightly closed to avoid trapping of water. Note that DMSO is quite hygroscopic and can take up more than 50% water from the air.

Up to 800 spots can be printed at a distance of 1 mm onto a standard microscope glass slide (2.5 cm×7.5 cm) by transferring about 10 nL from respective dilutions of the stocks in DMSO using pipetting robots equipped with liquid displacement syringes and a good enough *x*/*y*-precision (e.g., Slide Spotting Robot from Intavis). Alternatively, microarray instrumentation may be exploited. The use of piezo-dispensers is not recommended as the nozzles clog too easily; this can make life hard also with the split-pin instruments. Ring-and-pin instruments (e.g., the GMS 417 Arrayer in Fig. 10) are rather robust and reliable; we currently use this type of instrument with 500 µm solid pins to generate reliably miniarrays with a satisfying spot morphology.

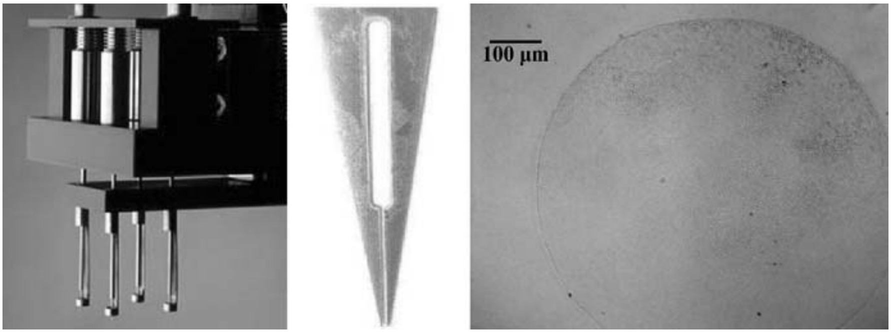


Fig. 10 (*left*) Ring-and-pin printing head of the GMS 417 Arrayer. (*middle*) Tip of a split pin. (*right*) Magnified image of a cellulose-peptide conjugate on a glass slide

The standard carrier for microarrays is the microscope glass slide. Cellulose-peptide conjugates generated by the SC^2-process readily adsorb to glass without the need of chemical fixing. However, the slides needs to be very clean and of suitable homogeneous hydrophilicity. Special products for the SC^2-arrays are commercially available. In this respect, it is also of interest that we successfully printed and assayed SC^2-spotting solutions on many other types of surfaces including plastic sheets made of PVC, PP and other polymers. This could be exploited to adopt SC^2-arrays to numerous alternative customized formats other than the conventional microscope slides. SC^2-arrays on special plastic coated slides are available as CelluSpots from Intavis AG; for home-made slides of this type contact the company.

5.1 Materials

- Acid soluble cellulose membrane for CelluSpots from Intavis Bioanalytical Instruments AG, Cologne, Germany (order No. 32.105).
- SynthoSlides for SC^2 from AIMS Scientific Products GmbH Braunschweig, Germany (order No. AG07-07/10) or coated slides for CelluSpots from Intavis (order No. 54.112).
- *tert*-Butyl methylether (MTBE), extra pure.
- Dimethyl sulfoxide (DMSO), 99.5% for synthesis.
- Dissolving solution. Trifluoroacetic acid (TFA, synthesis grade), dichloromethane (DCM) and triisobutylsilane (TIBS) in a ratio of 80% TFA, 13.7% DCM and 6.3% TIBS (mix in this order). Note that trifluoroacetic acid is very harmful and volatile, and should be handled with gloves under a hood.

5.2 Special Equipment

- A 384-hole punching device (Intavis).
- Matrix tubes. 1.4 mL 2D bar-coded TrakMates tubes (order No. 3711) are from Matrix Technologies (now Thermo Scientific Matrix), Hudson, NH, USA.
- CapMats (order No. 4431) from Matrix, not separable.
- Sepra Seal CapMats (order No. 4463) for sealing the Matrix tubes individually.
- MixMate. Benchtop mixing device from Eppendorf, Hamburg, Germany (order No. 5353 000.014).
- Temperature controlled ultrasonication bath such as the Sonorex Super 10P from Bandelin, Berlin, Germany.
- Suitable pipetting device for ether handling during the wash procedures. Serial Mate from Matrix Technologies, Hudson, NH, USA, is specially equipped by us with an exhaust device for pipetting ether.

- Suitable pipetting device for printing the miniarrays. Slide Spotting Robot from Intavis is equipped with a 500 µL syringe and a teflon coated needle for distributing solutions in nL ranges (order No. 54.000) or GMS 417 Arrayer from Affymetrix (Santa Clara, CA, USA) equipped with 500 µm pins.
- Microtiter plates made from polypropylene to fit into the sample holder of the printing device.
- For slide storage a microscope slide holder box (order No. K540.1) from Carl ROTH, Karlsruhe, Germany.

5.3 Methods

5.3.1 Preparation of Stock Solutions in DMSO of Cellulose-Bound Peptides

1. Prepare 120 mL dissolving solution.
2. Pipette 300 µL of the dissolving solution in every tube containing a cellulose-compound disc.
3. Seal tube-racks with a non-separable CapMat.
4. Shake/vortex racks for 1 h.
5. Sonicate racks for 1 h.
6. Repeat shaking/vortexing and sonication until all cellulose is dissolved.
7. Add 500 µL MTBE to each tube to precipitate the cellulose-compounds. The overall volume is now c800 µL per tube. Shake/vortex the rack for 5 min at 2000 rpm.
8. Place the tube racks for 15 min in a fridge at –20 °C; a white precipitate should be visible in the tubes.
9. Centrifuge the racks in a cooled centrifuge (4 °C) at 3000 rpm for 10 min.
10. Remove the supernatant from each tube, preferably by a pipetting device. Note that you leave ca. 25 µL of the supernatant in order not to perturb the precipitated pellet.
11. Wash the remaining pellets 3 times with each 500 µL MTBE by repeating steps 7 to 10.
12. Cover the racks with the opened tubes with a double sheet of 3MM paper. Remove very carefully remains of ether by tilting the racks aside so that liquid can slowly drip out onto the 3MM paper. Do not disturb the pellets!
13. Leave the tubes open to air for 1 h (max). Note the cellulose pellet must not totally dry out!
14. Dispense 500 µL DMSO into each tube.
15. Seal tube-racks with Sepra Seal CapMats (separable).
16. Dissolve the cellulose-compound pellets by a combination of shaking/vortexing and sonication at 40 °C until a clear, colorless solution for each cellulose-compound conjugate is obtained.

17. Flush freeze the stock solutions with liquid nitrogen and store at –20 °C until use.

5.3.2 Printing Miniarrays on Microscope Glass Slides

1. Prepare 1 : 20 dilutions with DMSO from the cellulose-peptide stock solutions in a polypropylene microtiter plate. Arrange peptides to fit the layout of the final array and the pipetting scheme used by the printing device. Include the positive and negative controls! Remember not to leave the DMSO solutions standing unsealed for a period of time longer than needed for dilution or printing, as the DMSO will trap considerable amounts of water from the air.
2. Print/spot the diluted cellulose-peptide solutions onto glass slides corresponding to the manual from the manufacturer of the printing device. The recommended type of slides enables the placement of spots at a distance of 1 mm by transferring about 10 nL solution. Note that when using coated slides (e.g., from Intavis) it could be necessary to spot up to 40 nL per spot to get a satisfying spot morphology. The spot distance should than be set to 1.2 mm.
3. Leave the slides sitting in the printing device until the DMSO is evaporated (approx. 60 min).
4. Place the slides in a microscope slide holder box and put the box with an open cover for 30 min in an oven at 60 °C.
5. Remove the object holder box from the oven, let it cool down, close the top and store the printed slides at –4 °C. Note that these slides should be usable for at least 12 months.

6 Guidelines for Probing the Peptide Arrays

6.1 General Considerations

The membrane segments from Sect. 4 or printed slides from Sect. 5 are now ready to be incubated with solutions of the protein acceptor such as an antiserum, body fluid, cell extract, recombinant protein, etc.

Depending on the protein of interest and the chosen detection method, different blocking conditions can be compared to obtain an optimal signal-to-noise ratio. The following blocking solutions of increasing "stringency" may be tested: (1) 3% BSA in PBS, (2) 2% (w/v) skim milk powder in TBS, (3) 2% (w/v) skim milk powder, 0.2% (v/v) Tween 20 in TBS, (4) MBS, (5) MBS with 50% (v/v) horse serum. The blocking conditions are quite critical and one may have to attempt several conditions for optimization. In our lab, blocking solution (3) works best for most membrane array applications, while (1) is recommended for fluorescence detection on glass slides.

Detection of antibody/protein molecules bound to the peptide spots can be achieved in a variety of ways. Besides the antibody-based immunoblotting [44], many other labeling techniques, e.g., with radioisotopes or fluorescent dyes are also fully compatible. At least one of the interaction partners has to be labeled for detection on a blot. Biotinylation of the probe protein followed by detection with a streptavidin-AP conjugate is a convenient option. If an antibody is the interaction partner, secondary antibodies or labeled protein A or G are recommended for detection. Precheck incubations using only the detection agents are always required, especially when enzyme-labeled animal sera are used, since unspecific binding or cross-reaction to antibody or enzyme may give rise to false positive reactions. It is further possible that specific antibodies to a certain antigen are present in the secondary antisera, for example when proteins are analyzed which originate from *E. coli* or other organisms naturally in contact with the donor animal. Generally, avoid detection procedures of several steps in order to loose sensitivity by washing off the protein binders. Antibodies bind bivalently and, thus, show high avidity and low off-rates. This, however, can be very unfavorable with low affinity monovalently binding proteins. The choice of a detection system should ensure that peptide spots will not become chemically or otherwise irreversibly modified, because peptide arrays on cellulose membranes can be reused many times (more than 20 times) when treated properly. Therefore, alkaline phosphatase is recommended over horse radish peroxidase as enzyme label because the latter requires addition of hydrogen peroxide which also oxidizes the peptides; NBT-based color development is not removable from the membrane, therefore MTT is recommended. A comprehensive collection of relevant publications [45] lists many useful detection procedures.

In case of low-affinity interactions with higher off-rates of the protein peptide complex or in cases of significant perturbation by unspecific background from, e.g., detection antibody, it is recommended to perform the detection of the proteins bound to the spots after electrotransfer to a secondary (mostly nitrocellulose) membrane. This also allows the use of a wider range of other detection procedures [22].

Signal patterns obtained from peptide arrays on spots can be documented and quantitatively evaluated utilizing modern image analysis systems as used with other 2D analysis media such as electrophoresis gels, blotting membranes and microarrays.

Fluorescence-labeled detection reagents enable discrimination in one experiment controls and several target proteins by applying a mixture of target-specific labeled reagents with different distinguishable fluorophores. Fluorescence detection, however, can be obscured on membranes by quenching or background fluorescence, but is very effective with arrays on glass microscope slides.

6.2 Materials

- Tris-buffered saline (TBS). 8.0 g NaCl, 0.2 g KCl and 6.1 g Tris-base in 1 L water. Adjust pH to 7.0 with HCl. Autoclave and store at 4 °C.
- T-TBS. TBS buffer plus 0.05% Tween20.
- Phosphate-buffered saline (PBS). 8.0 g NaCl, 0.2 g KCl, 1.43 g $Na_2HPO_4 \cdot 2H_2O$ and 0.2 g KH_2PO_4 in 1 L water. Adjust pH to 7.0 with HCl. Autoclave and store at 4 °C.
- Citrate-buffered saline (CBS). 8.0 g NaCl, 0.2 g KCl, and 10.51 g citric acid ($\times 1H_2O$) in 1 L water. Adjust pH to 7.0 with NaOH. Autoclave and store at 4 °C.
- Membrane blocking solution (MBS). Mix 20 mL casein-based blocking buffer concentrate (No. B6429; Sigma-Genosys Inc.), 80 mL T-TBS (pH 8.0) and 5 g sucrose; the resulting pH will be 7.6. Store at 4 °C.
- Horse serum (Gibco-Invitrogen, Carlsbad, CA, USA).
- Alkaline phosphatase (AP) conjugated detection antibodies.
- Alkaline phosphatase (AP) conjugated streptavidin.
- Color developing solution (CDS). Dissolve 50 mg 3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyltetrazolium bromide (MTT) in 1 mL of 70% DMF in water. Store at –20 °C. Dissolve 60 mg 5-bromo-4-chloro-3-indolylphosphate *p*-toluidine salt (BCIP) in 1 mL DMF. Store at –20 °C. Prepare CDS always fresh: to 10 mL CBS add 50 μL 1 M magnesium chloride, 40 μL BCIP and 60 μL MTT. Remember to never use NBT instead of MTT, since the developed color cannot be removed from the membrane.
- Immun-Star chemiluminescent kit (No. 170-5018, Bio-Rad Laboratories, Hercules, CA, USA).
- Horse radish peroxidase (HRP) conjugated detection antibodies.
- ECL western blotting detection reagents (Amersham Biosciences UK Ltd, Buckinghamshire, UK).
- Protan Nitrocellulose Transfer Membrane (Schleicher & Schuell, Dassel, Germany).
- Transfer buffer for western blotting. 25 mM Tris-Cl (pH 7.6), 192 mM glycine, 20% methanol, 0.03% sodium dodecyl sulfate (SDS).
- Stripping mix A (SM-A). 8 M urea, 1% SDS in PBS; store at room temperature. Add 0.5% 2-mercaptoethanol prior to use and adjust pH to 7.0 with acetic acid.
- Stripping mix B (SM-B). 10% acetic acid, 50% ethanol and 40% water; store at room temperature.
- Fluorescence-labeled streptavidin: Streptavidin Alexa Flour 647 conjugate from Molecular Probes (Eugene, OR, USA) (1.0 mg per mL). Note that

the fluorescence label of the streptavidin should be detectable at another wavelength than the labels used for the secondary antibodies.

- Horse radish peroxidase (HRP) conjugated streptavidin for chemiluminescence detection.
- Fluorescence-labeled detection antibodies (for label see above).
- BSA-blocking solution for slides (albumin, bovine serum, Fraction V, approx. 99%; available from Sigma-Aldrich, St. Louis, MO, USA; order No. A3059): 3% BSA in PBS buffer, pH 7.4.

6.3 Special Equipment

- Polystyrene cell culture plates (12 cm × 12 cm) with covers (Greiner Bio-One, Frickenhausen, Germany; order No. 688102).
- Flat glass tray to hold at least one membrane.
- A sonication bath with temperature control.
- A digital recording device (scanner or CCD camera) for documentation of signal patterns on membranes plus analysis software for quantification of signals. In case of chemiluminescence detection, autoradiography (X-ray) films can be used.
- Plastic bags and sealing device.
- A blotting apparatus (Biometra-Fast-Blot, Göttingen, Germany, B337593) or others.
- Thin transparent plastic wrap (Saran Wrap).
- Slide handling equipment (e.g., from Carl ROTH, Karlsruhe, Germany).
- Incubation or staining chamber (order No. HL98.1), staining box (order No. HA44.1), microscope slide holder (order No. HA49.1), cover slips (order No. H878.1), microscope slide holder box (order No. K540.1).

6.4 Methods

6.4.1 Probing Peptide Arrays on SPOT Membranes

This basic procedure is worked out for use with AP conjugated detection antibody and a color signal development. HRP-labeled detection agents require hydrogen peroxide and gradually destroy the peptides! More sensitive detection can be achieved with a chemiluminescent substrate of AP (e.g., Immun-Star). In this case follow the instructions of the supplier for steps 9 to 12. Alternatively, chemical biotinylation of proteins is quite popular. These can be detected on the membrane using AP conjugated streptavidin under the same conditions as with the AP secondary antibody. If radioactive labeled reagents are used, adopt steps 5 to 12 accordingly. As an easy alternative to chemical labeling, in vitro coupled transcription/translation systems (TNT, Promega) readily produce sufficient radio-labeled proteins from cDNA cloned into plasmids with a T7 promoter. The resulting reaction mix can be directly applied in membrane probing [46].

Prior to probing your protein with the peptide spots on the membrane, always apply this protocol first by omitting steps 5 and 6. This is necessary to assess unspecific signals from components of the detection process or remaining proteins from a previous experiment on the same membrane. In case the proteins are electrotransferred and detected on a secondary nitrocellulose membrane (method B, below), this precaution does not apply.

Protein Binding Assay Protocol

Method A

1. Place the membrane in a polystyrene plate and wet with a few drops of alcohol. Please note that this is to enhance rehydration of some hydrophobic peptide spots. The peptide locations should not be visible as white spots! If this happens, treat with alcohol in a sonication bath at room temperature until spots have disappeared.
2. Wash membrane for 10 min with 10 mL TBS (three times).
3. Incubate overnight with 10 mL MBS.
4. Wash membrane once for 10 min with 10 mL T-TBS.
5. Incubate for 2 to 4 h with probe antibody (or protein) diluted in 8 to 10 mL MBS. For monoclonal antibodies or pure proteins use approximately 4–5 µg of purified antibody per milliliter incubation volume. When using a polyclonal serum, we recommend a dilution of 1 : 100. Note that it is not necessary to use a large volume of protein solution for the incubation. However, make sure that the membrane is completely covered and prevent drying out by using a lid or seal in a plastic bag.
6. Wash membrane for 10 min with 10 mL T-TBS (three times).
7. Incubate for 1 to 2 h with AP conjugated secondary antibody diluted in 10 mL MBS.
8. Wash membrane for 10 min with 10 mL T-TBS (twice).
9. Wash membrane for 10 min with 10 mL CBS (twice).
10. Transfer the membrane to a flat glass tray and add 10 mL of CDS. Incubate without agitation until good signals are obtained. For individual peptides on spots this usually takes 10 to 30 min; peptide pools may require longer incubations (2 h to overnight). Stop reaction by washing twice with PBS (1×30 s, 1×3 min). Keep membrane wet. For storage, leave at 4 °C in a container with PBS or cover with plastic wrap. (Please note if the membrane dries out, proteins may denature and become difficult to remove.) After successful documentation of signals by photography or electronic scanning, continue with membrane stripping.

Method B

If weak binding of the test protein is anticipated, or when too high background from the detection reagents is observed in method A, the electrotransfer of bound test protein onto a secondary nitrocellulose membrane may help. Here, any appropriate detection system on the nitrocellulose can be used (e.g., HRP conjugates), as the peptides will not be affected. Proceed first as above steps 1 to 6.

7a. Briefly equilibrate both the peptide membrane and a sheet of nitrocellulose, trimmed to fit the peptide membrane, in transfer buffer.

8a. Electrotransfer the proteins bound to the peptide spot membranes onto nitrocellulose for 1 h using 0.85 mA cm^{-2}. Due to the denaturation by SDS, all proteins should have acquired a negative charge. Therefore, the nitrocellulose should be placed towards the positive electrode. Depending on the chemical properties of the protein ligands, the time required for the transfer might differ and, therefore, has to be determined empirically.

9a. Block the nitrocellulose membrane with MBS for 2 h at room temperature.

10a. Incubate the nitrocellulose membrane for 75 min with an AP or HRP conjugated detection antibody or AP-/HRP-streptavidin for biotinylated proteins diluted in MBS. Use dilutions comparable to those employed in immunoblots after SDS-PAGE.

11a. Wash the nitrocellulose membranes three times for 5–10 min with T-TBS, subsequently followed by washing three times for 5–10 min with TBS.

12a. Remove excess buffer from the nitrocellulose membrane by gently placing the membrane on a sheet of 3MM paper. To avoid damage to the adsorbed protein do not wipe or press tissue onto the membrane.

13a. Detect the spots by using a chemiluminescence detection kit according to the instructions of the suppliers. Note the following: (1) If no signal can be detected after 30 min of exposure, check the detection system with a positive control from the kit. If detection reagents are functional, use less stringent blocking. If no binding occurs, this may indicate a discontinuous binding site or very low affinity binding. (2) In case of unspecific signals and a high background, increase the stringency of the blocking conditions and make sure that your primary binding partner and detection reagent (e.g., antibody) are of high purity and are used in the highest possible dilution.

Membrane Regeneration (Stripping)

A peptide spot membrane that has been processed through a protein binding assay can be used again for probing another protein if all remains from the assay can be removed completely (stripping). Principally, membranes can be regenerated up to 50 times without loss of signal intensity, because the peptides are very stably immobilized. But in some cases proteins resist elution from the spots, and the membranes can only be used once for method A (on-spot membrane detection). This has to be checked by probing a regenerated spot membrane first with the detection system (see protein binding assay protocol). Alternatively, method B of the protein binding assay protocol can be applied.

1. Wash the spot membrane for 10 min with 20 mL of water (twice).
2. Incubate with 20 mL DMF until the blue color of spot signals has dissolved (usually about 10 min; incubate in a sonication bath at 40 °C if necessary). Remove the solution and wash once again for 10 min with 20 mL DMF. This step can be omitted if other than a dye precipitation detection was used.
3. Wash the spot membrane for 10 min with 20 mL of water (three times).
4. Wash the spot membrane for 10 min with 20 mL SM-A in a sonication bath at 40 °C (three times).
5. Wash the spot membrane for 10 min with 20 mL SM-B (three times).
6. Wash the spot membrane for 10 min with 20 mL alcohol (three times).
7. Go to step 2 of the protein binding assay protocol for the next binding assay or dry the membrane with cold air from a hairdryer in between a folder of 3MM and store at –20 °C sealed in a plastic bag.

6.4.2
Probing Peptide Arrays on Glass Slides

All washing steps are reduced to 3 min and carried out under gentle agitation on a rocker table.

1. Place the slides in the microscope slide holder of a staining box and put the holder in the staining box. Add approx. 100 mL ethanol into the box until the slide working areas are covered completely with alcohol. Wash for 3 min. Note that this is to enhance rehydration of some very hydrophobic peptide spots.
2. Wash the slides with TBS (three times).
3. Incubate overnight with BSA-blocking solution. Note that depending on the protein of interest and the detection method, different blocking conditions can be compared to obtain an optimal signal-to-noise ratio. Avoid impure, fluorescent reagents when using fluorescence detection.
4. Wash once with T-TBS.
5. Place the slides in an incubation chamber.
6. Prepare incubation solutions. Per slide dilute 1 μL of the sample to be probed in 100 μL BSA-blocking solution. Pipette immediately 60 μL incubation solution on the respective slide. Prevent drying out of the slide surface. Place a glass cover slip slowly onto the surface of the droplet to spread the antibody solution. Locking of air bubbles under the cover slip must be avoided. Incubate for 2 to 4 h. Note that for monoclonal an-

tibodies or pure proteins use approximately 4–5 μg per mL incubation volume. When using a polyclonal serum or cell extract, we recommend to start with a dilution of 1 : 100. It is not necessary to use a larger volume of protein solution for the incubation. Depending of the viscosity of the incubation solution volumes between 60 μl and 100 μl per slide are required.

7. Remove the cover slips carefully from the slide surfaces by washing off with T-TBS. Transfer the slides back into the microscope holder and place in the staining box.
8. Wash the slides with T-TBS (three times).
9. Prepare the labeled detection reagents such as secondary antibodies diluted in BSA-blocking solution corresponding to your selection of target proteins. This second incubation solution should contain the labeled streptavidin to detect the biotin controls! Follow the instructions from entry 5. Incubate for 1 to 2 h. Note that commercial detection reagents that contain 1 mg per 1 mL stock solution should be diluted to 1 : 400 or higher in BSA-blocking solution.
10. Wash the slides with T-TBS (three times).
11. Signal read out for fluorescence label:
    - Wash the slide with Milli-Q water (three times).
    - Dry the slide in a nitrogen stream or by centrifugation (1000 rpm for 10 min).
    - Place slides into slide reader instrument for scanning.
12. Signal read out for chemiluminescence detection:
    Keep slides under TBS buffer until processed by a chemiluminescence procedure in a dark room or scanning device. Slides must not dry in order to maintain activities of the HRP or AP enzymes.
    A simple procedure includes:
    - Place a piece of X-ray film on a clean surface.
    - Cover with a thin transparent plastic wrap (Saran Wrap).
    - Add a droplet of 200 μL of a chemiluminescence substrate solution central to the covered X-ray film. For the substrate solution follow the instructions of the supplier of your chemiluminescence kit.
    - Remove excess buffer from the slide by letting it run off from one edge onto a piece of paper towel for about 2 s; place the slide top down onto the droplet for the time of exposition needed. Avoid trapping of air bubbles. Do not use more substrate solution, otherwise the slide will swim and yield an unfocused image.
    - Place the slide back to a reservoir of TBS buffer for further expositions and process X-ray film.

7 Concluding Remarks

The protocols are optimized to help any researcher, even if not trained in chemistry, to prepare high quality low cost synthetic peptide arrays for a variety of biological screening experiments, most prominently for the detailed molecular study of protein–protein interactions. These protocols worked successfully also under extreme conditions such as the tropical summer of Argentina with lab temperatures of about 35 °C, the dichloromethane almost boiling and a humidity of over 90%. The authors are happy to give advice in case of problems with the procedures or changes to the procedures for other applications. We welcome comments, corrections and suggestions.

Acknowledgements The protocols given here reflects the expertise of many colleagues developed over the past years. We thank them for their input, improvements to the method and extension of applications. This work is supported by Grants No. 01GR0102 and 01GR0474 of the German National Genome Research Net (NGFN). We thank Dr. Heinrich Gausepohl and Dr. Ole Brandt from Intavis Bioanalytical Instruments AG for their helpful discussions to improve the SC^2-process and make it a commercial product, as well as Dr. Norbert Zander from AIMS Scientific Products for support in the development of suitable synthesis membranes and modified microscope slides.

References

1. Atassi MZ (1977) The complete antigenic structure of myoglobin: Approaches and conclusions for antigenic structures of proteins. In: Atassi MZ (ed) Immuno Chemistry of Proteins. Plenum, New York, p77
2. Geysen HM, Rodda SJ, Mason TJ (1986) Mol Immunol 23:709
3. Cesarini G, Gimona M, Sudol M, Yaffe M (eds) (2004) Modular Protein Domains. Wiley, Weinheim
4. Gavin A C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz et al. (2002) Nature 415:141
5. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P et al. (2002) Nature 415:180
6. Fields S, Song OA (1989) Nature 340:245
7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR et al. (2000) Nature 403:623
8. Tong A HY, Drees B, Nardelli G, Bader GD, Brannetti B et al. (2002) Science 295:321
9. Frank R (2002) Comb Chem High Throughput Screening 5:429
10. Frank R (1992) Tetrahedron 48:9217
11. Dikmans A, Beutling U, Schmeisser E, Thiele S, Frank R (2006) QSAR Comb Sci 25:1069
12. Bergmannn S, Wild D, Diekmann O, Frank R, Bracht D, Hammerschmidt S (2003) Mol Microbiol 49:411
13. Blüthner M, Mahler M, Müller DB, Dünzl H, Bautz FA (2000) J Mol Med 78:47
14. Mukhija S, Germeroth L, Schneider-Mergener J, Erni B (1998) Eur J Biochem 254:433

15. Kneissel S, Queitsch I, Petersen G, Behrsing O, Micheel B, Dübel S (1999) J Mol Biol 288:21
16. Schultz J, Hoffmüller U, Ashurst J, Krause G, Schmieder PJ, Macias M, Schneider-Mergener J, Oschkinat H (1998) Nature Struct Biol 5:19
17. Espaniel X, Wälchli S, Rückle T, Harrenga A, Huguenin-Reggiani M, van Huijsduijnen RH (2003) J Biol Chem 278:15162
18. Geysen HM, Mason TJ (1993) Bioorg Med Chem Lett 3:397
19. Dooley CT, Houghten RA (1993) Life Sciences 52:1509
20. Frank R, Kieß M, Lahmann H, Behn C, Gausepohl H (1995) Combinatorial synthesis on membrane supports by the SPOT technique. In: Maia LS (ed) Peptides 1994. ESCOM, Leiden, p 479
21. Oggero M, Frank R, Etcheverrigaray M, Kratje R (2004) Mol Divers 8:257
22. Rüdiger S, Schneider-Mergener J, Bukau B (2001) EMBO J 20:1042
23. Dostmann WRG, Taylor MS, Nickl CK, Brayden JE, Frank R, Tegge WJ (2000) Proc Natl Acad Sci 97:14772
24. Rodriguez M, Li SSC, Harper JW, Songyang Z (2004) J Biol Chem 279:8802
25. Schmidt TGM, Koepke J, Frank R, Skerra A (1996) J Mol Biol 255:753
26. Reineke U, Ivascu C, Schlief M, Landgraf C, Gericke S, Zahn G, Herzel H, Volkmer-Engert R, Schneider-Mergener J (2002) J Immunol Method 267:37
27. Frank R (2002) J Immunol Method 267:13
28. Valle M, Kremer L, Martínez C, Roncal F, Valpuesta JM, Albar JP, Carrascosa JL (1999) J Mol Biol 288:899
29. Billich C, Sauder C, Frank R, Herzog S, Bechter K, Takahashi K, Peters H, Staeheli P, Schwemmle M (2002) Biol Psychiatry 51:979
30. Töpert F, Knaute T, Guffler S, Pires JR, Matzdorf T, Oschkinat H, Schneider-Mergener J (2003) Angew Chem Int Ed 42:1136
31. Bialek K, Swistowski A, Frank R (2003) Anal Bioanal Chem 376:1006
32. Hoffmüller U, Russwurm M, Kleinjung F, Ashurst J, Oschkinat H, Volkmer-Engert R, Koesling D, Schneider-Mergener J (1999) Angew Chem Int Edit 38:2000
33. Chan WC, White PD (eds) (2000) Fmoc Solid Phase Peptide Synthesis: a practical approach. Oxford University Press, Oxford, UK
34. Zander N (2004) Mol Divers 8:189
35. Fields GB, Noble RL (1990) Int J Pept Protein Res 35:161
36. Töpert F, Pires R, Landgraf C, Oschkinat H, Schneider-Mergener J (2001) Angew Chem Int Edit 40:897
37. Gail R, Frank R, Wittinghofer A (2005) J Biol Chem 280:7107
38. Krchnák V, Vágner J, Safár P, Lebl M (1988) Collect Czech Chem Commun 53:2542
39. Kramer A, Volkmer-Engert R, Malin R, Reineke U, Schneider-Mergener J (1993) Peptide Res 6:314
40. Frank R (1994) Spot-Synthesis: An easy and flexible tool to study molecular recognition. In: Epton R (ed) Innovations and Perspectives in Solid Phase Synthesis 1994. Mayflower Worldwide, Birmingham, p 509
41. Hoffmann S, Frank R (1994) Tetrahedron Lett 35:7763
42. IRIS BioTech (2007) http://www.iris-biotech.de. last visited: 9 April 2008
43. Kramer A, Reineke U, Dong L, Hoffmann B, Hoffmüller U, Winkler D, Volkmer-Engert R, Schneider-Mergener J (1999) J Peptide Res 54:319
44. Harlow E, Lane D (1988) Antibodies – A laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, p 319–358

45. Frank R, Schneider-Mergener J (2002) SPOT-Synthesis – Scope of applications. In: Koch J, Mahler M (eds) Peptide Arrays on Membrane Supports: A Laboratory Manual. Springer, Berlin Heidelberg New York, p 1
46. Niebuhr K, Ebel F, Frank R, Reinhard M, Domann E, Carl UD, Walter U, Gertler FB, Wehland J, Chakraborty T (1997) EMBO J 17:5433

Adv Biochem Engin/Biotechnol (2008) 110: 153–175
DOI 10.1007/10_2008_101

Published online: 5 June 2008

Antibody Microarrays as an Experimental Platform for the Analysis of Signal Transduction Networks

Ulrike Korf[1] (✉) · Frauke Henjes[1] · Christian Schmidt[1] · Achim Tresch[2] · Heiko Mannsperger[1] · Christian Löbke[1] · Tim Beissbarth[1] · Annemarie Poustka[1]

[1]Division of Molecular Genome Analysis, B050, Deutsches Krebsforschungszentrum,
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany
u.korf@dkfz.de

[2]Department of Chemistry and Biochemistry, Ludwig-Maximilians-Universität
München, Feodor Lynen Str. 25, 81377 München, Germany

Abstract A significant bottleneck for the time-resolved and quantitative description of signaling networks is the limited sample capacity and sensitivity of existing methods. Recently, antibody microarrays have emerged as a promising experimental platform for the quantitative and comprehensive determination of protein abundance and protein phosphorylation. This review summarizes the development of microarray applications involving antibody-based capture of target proteins with a focus on quantitative applications. Technical aspects regarding the production of antibody microarrays, identification of suitable detection and capture antibody pairs, signal detection methods, detection limit, and data analysis are discussed in detail.

Keywords Antibody array · Protein microarray · Protein phosphorylation · Quantitative biology · Systems biology

1 Introduction

Signal transduction is mediated via complex networks of interacting proteins. Understanding how signals flow through these pathways, and how healthy and diseased tissues differ in intracellular signaling, requires tools to decipher the cellular network in a comprehensive way. Antibody-based methods have a long tradition in analyzing the turnover of proteins and their posttranslational modifications, but established approaches, such as Western blotting or enzyme-linked immunosorbent assay (ELISA), are limited to the detection of single proteins at a time. In contrast to traditional antibody-based approaches, protein microarrays possess the capacity to multiplex the detection of several target proteins. Thus, different key players of signal transduction can be monitored in parallel, and therefore this experimental platform has emerged as an appropriate tool to study crosstalk between different signal transduction modules.

A conventional ELISA requires substantial quantities of antibody to coat the bottom of a multititer plate. The comparatively high amount of capture antibody can result in the depletion of target protein from the sample, thus reducing the accuracy and sensitivity of the measurement. To overcome these shortcomings, immunoassay miniaturization was introduced in the mid-1980s and initially applied to the quantification of small molecules, such as hormones [1]. Besides improving the sensitivity and accuracy, the miniaturized format of antibody microarrays cuts down on the costs of consumables. The capacity of protein microarrays to multiplex several proteins reduces sample consumption and assay time. Quite recently, antibody microarrays were also used as an experimental platform to demonstrate the finely tuned character of signal transduction by monitoring the dose-dependent phosphorylation of key signal transduction molecules [2].

The visualization of captured proteins can be achieved either directly (Fig. 1A) or indirectly as a one-step (Fig. 1B) or two-step procedure (Fig. 1C). For direct detection, the sample must be labeled with a suitable dye before incubation on the array (Fig. 1A). This labeling procedure potentially influences the antigenic properties of a protein by masking epitopes, and therefore can disturb the final readout. For indirect detection, the antibody carries either a detectable label (Fig. 1B) or is recognized by a secondary antibody (Fig. 1C). Antibody pairs for so-called sandwich detection (Fig. 1C) must consist of antibodies from two different animal species, for example, from mouse and rabbit. A two-step indirect detection increases the specificity of the antibody-based recognition, because two different antibodies will rarely reveal identical

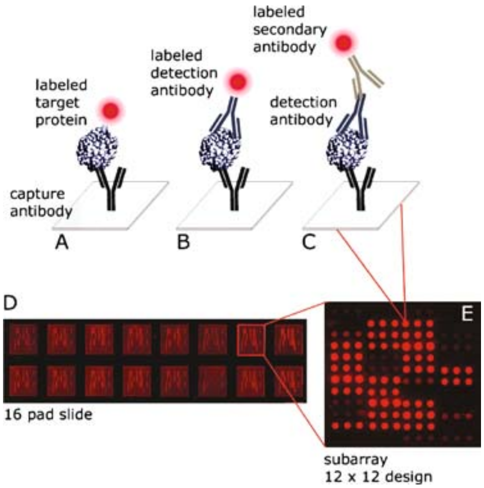


Fig. 1 The majority of antibody microarray approaches use fluorescent dyes for signal detection. These dyes are covalently coupled to protein, either the sample itself or antibodies used for detection; direct detection of a labeled probe (*A*), indirect detection with a labeled detection antibody (two-step procedure) (*B*), indirect detection with a labeled secondary antibody (three-step procedure) (*C*). Image of a multipad slide after signal detection in the near-infrared (NIR) range at 700 nm (*D*). A single pad is shown to illustrate the spot morphology (*E*)

profiles of unspecific interactions with unrelated proteins. Fluorescent dyes are well established for signal detection on protein microarrays, although other approaches were explored as well. In general, protein microarray data reveal excellent signal-to-noise ratios [3], and are therefore appropriate for systems biology approaches. Since measurements on a multiplexed scale result in considerable amounts of data, suitable strategies and software tools for automated spot detection, determination of signal intensity, background correction, and the final data analysis were introduced in parallel.

2
Protein Profiling on Antibody Microarrays

2.1
Overview of Semiquantitative Approaches

Numerous antibody microarray approaches were explored as an analytical tool for protein profiling on a multiplexed scale. Most approaches share the idea of employing the slide format as solid support for the immobilization of antibodies, but differ in the type of microarray content, and in strategies for the immobilization of antibodies and detection of target proteins (Table 1).

Table 1 Noncommercial antibody microarray applications

Author	Array design	Array production	Microarray coating	Detection mode	Signal analysis	Application
Haab [4]	115 antibodies	Steel tips	Poly-L-lysine-coated glass slides	Cy3/Cy5 indirect detection	Ratiometric readout	Proof-of-principle approach
Screecumar [6]	146 antibodies	Contact Gene machines (San Carlos, CA)	Poly-L-lysine- or superaldehyde-coated glass slides	Cy3/Cy5 direct detection	Ratiometric readout	Protein profiling
Nielsen [7]	3 antibodies	Contact GMS 417 arrayer (Affymetrix)	BSA-coated glass slides	Cy3/Cy5 direct vs indirect	Ratiometric readout	Analysis of protein phosphorylation
Ivanov [9]	3–360 antibodies	Manually (Schleicher & Schüell)	Poly-L-lysine-coated glass slides; PVDF membrane	HRP+ECL/film (1) indirect Fluorescent dyes (2) direct	Ratiometric readout	Analysis of posttranslational modifications
Gembitsky [8]	35 antibodies	Contact Microgrid II (Genomics Solutions)	Hydrogel-coated glass slides	Cy5 direct + indirect	Ratiometric readout	Protein profiling
Usui-Aoki [10]	182 antibodies	Contact Microgrid II (Genomics Solutions)	Bare gold affinity chips	Label-free direct detection	Ratiometric readout	Protein profiling
Duffy [11]	32 antibodies	Biodot Array (BioRad)	PVDF membrane	ECL indirect	Yes/no answer	Discovery tool for protein–protein interactions
Chen [12]	36–48 antibodies	Noncontact, Biochip Arrayer (PerkinElmer Life Sciences)	PATH slides, ultrathin nitrocellulose on glass slides	Phycoerythrin	Ratiometric readout	Analysis of glycan structures

Table 1 (continued)

Author	Array design	Array production	Microarray coating	Detection mode	Signal analysis	Application
Korf [2]	4 antibodies	Noncontact, Biochip Arrayer (PerkinElmer Life Sciences)	Nitrocellulose-coated multipad slides	Alexa 680 (NIR) indirect	Absolute quantification	Time-resolved analysis of protein phosphorylation

In 2001 Haab and coworkers [4] demonstrated that the microarray format is suitable for the highly sensitive detection of different proteins. Using a set of 115 different antibody/antigen pairs and a direct detection strategy, the sensitivity of antibody and protein microarrays was compared. Antibody microarrays revealed a tenfold lower detection limit when put side by side with the protein microarray format, and enabled a detection limit of 1 ng/ml for most proteins down to a partial concentration of 10^{-6} [5]. Thus, antibody microarrays were the more promising platform for the detection of low-abundance proteins in crude mixtures.

Miniaturized antibody microarrays were also employed for comparative protein profiling to monitor the impact of ionizing radiation on a cancer cell line. In this instance, a dye-swap experiment was performed for fluorescence detection on antibody microarrays with 146 different features. Several apoptotic markers were found to be significantly upregulated after radiation, indicating that this platform is potentially useful to analyze treatment outcome [6].

The critical steps in antibody microarray fabrication were also examined by Nielsen and coworkers in 2003 [7]. In measurements with recombinant proteins and of crude biological mixtures, the sensitivity of direct and indirect detection were compared. The phosphorylation of epidermal growth factor receptors was chosen as a biological model and monitored in different cell lines. Recombinant proteins could be detected with greater sensitivity in a direct detection approach. However, when profiling the activation of endogenously expressed cell surface receptors, the indirect detection was significantly more sensitive, indicating that the sandwich approach might be more suitable for the analysis of complex biological samples. The introduction of a covalent fluorescent tag as required for direct detection possibly performs differently in complex lysates than in samples with purified recombinant proteins, which might explain the observed discrepancy. The indirect detection approach was employed to profile the abundance of two different cell surface receptors, EGFR and ERBB2 (members of the epidermal growth factor receptor family), and their activation in response to ligand binding. Receptor phosphorylation, as readout for receptor activation, was detected by using specific phosphotyrosine antibodies labeled with fluorescent dyes. The analysis of time-course experiments reflected the fast dynamics of signal transduction through EGFR. In addition, the abundance of the transferrin receptor was taken as a measure to calibrate the amount of sample loaded on the slide. The impact of a small molecule inhibitor employed at different concentrations was used to disturb signal transduction through the EGFR to demonstrate the utility of protein microarrays for drug discovery research.

The deregulation of tyrosine phosphorylation has been implicated in many types of cancer. Mutations frequently result in the upregulation of certain kinases, which quickly moved into focus for the development of targeted therapies. In 2004, Gembitsky and coworkers [8] described an antibody microarray approach to profile the posttranslational modification patterns of proteins. The protocol was optimized to allow profiling of tyrosine phosphorylation for up to 35 different proteins in response to growth factor receptor mediated activation. Depending on the tissue type, 1000 to 100 000 cells were needed for sample preparation. The readout was ratiometric and compared relative amounts of protein abundance. Results from the inhibition of the BCR-ABL kinase by Gleevec indicate that antibody microarrays are useful to dissect signaling pathways, and to profile the activity of target-specific anticancer drugs. Ivanov and coworkers [9] employed a similar approach to profile posttranslational modification patterns of proteins involved in intracellular regulation. Selected proteins were immunoprecipitated and fluorescently labeled. The precipitated proteins were subsequently examined with antibody microarrays recognizing specifically posttranslational modifications, e.g., phosphotyrosine, ubiquitin, or acetyllysine. The resulting signal intensities reflected relative levels of posttranslational modifications. A step ahead toward label-free detection on antibody microarrays was made by coupling surface plasmon resonance technology as an optical sensor to detect binding of target proteins in real time. In this proof-of-principle approach, 382 antibodies against proteins from the mouse KIAA clone collection were spotted on gold affinity chips for protein expression profiling. Differential protein expression patterns were detected in mouse tissues [10].

A simplified prototype of an antibody array was produced by immobilizing antibodies on PVDF membrane with a vacuum-driven filtration setup; 32 different antibodies were selected to identify binding partners of a scaffolding protein [11]. Antibody microarrays were also employed to monitor specific changes in the glycan structure of proteins by examining the variation of selected oligosaccharide structures with glycan-specific lectins [12]. In this experiment, capture antibodies were printed on ultrathin nitrocellulose-coated glass slides. The analysis of glycan structure required the chemical derivatization of existing capture antibody-bound oligosaccharides. Lectin profiling revealed the cancer-associated increase of the sialyl Lewis acid structure on the tumor-associated antigens MUC1 (mucin-1) and CEA (carcino embryonic antigen).

Moreover, commercial antibody arrays were also exploited for the relative comparison of protein abundance. Biological samples were mostly labeled with the fluorescent dyes Cy3 and Cy5, mixed, and the respective signal intensities compared after microarray incubation [13–15]. In summary, the results reveal that antibody microarrays are useful for protein profiling, as well as for monitoring changes in the pattern of posttranslational modifications. However, all approaches were merely based on comparative profiling and did not allow the precise quantification of target proteins.

2.2 Introduction of Antibody Microarrays for the Absolute Quantification of Protein Phosphorylation

The exact quantification of proteins and posttranslational modifications requires well-characterized standards, antibodies, or antibody pairs, and suitable software tools for the analysis of measurements resulting from multiplexed calibration slopes. None of the previously published or commercial approaches includes a calibration step for the calculation of absolute numbers on the turnover of proteins [16, 17]. Thus, these arrays deliver a relative readout at best. The absolute quantification of protein phosphorylation on a multiplexed scale was recently realized by combining the sandwich format with the detection of fluorescent signals in the near-infrared (NIR) range. Signaling through cytokine receptors was examined by monitoring the phosphorylation of Erk1/2 (extracellular-signal regulated kinase) and Stat3 (signal transducers and activators of transcription) [2]. Several new steps were introduced to the experimental design of the first proof-of-concept approaches. First, the precise quantification of a certain posttranslational modification required well-characterized standard proteins, and the phosphorylation rate of standard proteins was determined by liquid chromatography/mass spectrometry (LC/MS). Antibody pairs compatible for quantitative detection in a multiplexed setting were identified as being crucial for a robust quantitative readout. For this reason, quality measures for accuracy and dynamic range of different antibody pairs were introduced and summarized as an "antibody-pair plot" (Fig. 2). This plot can be used to estimate specifically the extent of cross reactivity in combinations of different detection antibodies, and can also be employed for the identification of antibody pairs in a multiplexed setting. The miniaturized format allows routine time-resolved measurements

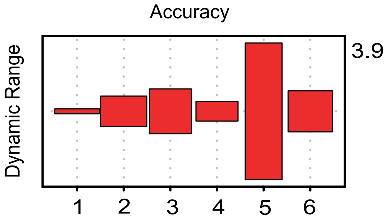


Fig. 2 Antibody-pair plot summarizing the quality measures of dynamic range and accuracy for six different capture antibodies. The first capture antibody (1) recognizes Stat3, the second capture antibody (2) pStat3. Phosphorylated ERK1/2 is detected with antibody 3 and antibody 5; total ERK1/2 is detected with antibodies 4 and 6. The detection of captured protein was performed with a mixture of two Stat3-specific detection antibodies and two ERK1/2-specific detection antibodies. The antibody-pair plot illustrates the dynamic range of signal intensities on a log2 scale, presented as the length of the box. The accuracy is presented as the width of the box

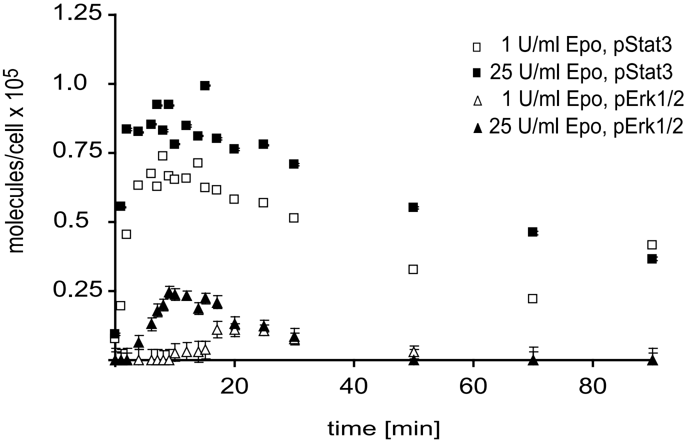


Fig. 3 Erythropoeitin (Epo)-mediated activation of Stat3 and Erk1/2 phosphorylation in a mouse cell line transiently expressing a hybrid receptor for Epo (EpoR:gp130) [2]. Pathway activation was expressed as molecules of phosphoprotein per cell after stimulation with two different concentrations of Epo. *Error bars* are indicated

from primary cells with high sensitivity requiring only a few thousand cells. The accuracy and sensitivity of this quantitative antibody microarray approach enables differentiation between major events, e.g., the activation of Stat3 through a cytokine receptor, and minor events, such as the activation of Erk1/2. The dose–response rates illustrated that the duration and total amount of protein phosphorylation were clearly concentration-dependent (Fig. 3). In summary, quantitative antibody microarrays emerged as a promising tool to examine the fine-tuning of signal transduction, and to analyze the crosstalk between signaling modules.

2.2.1
Generation of Standard Curves on Multipad Slides

The quantification of individual target proteins requires a standard curve based on the measurements of suitable calibrator proteins (Fig. 4). In the Quantpro software the data describing the correlation between standard and the corresponding signal intensities was called a *calibration series* [2]. A linear regression was fitted on the calibration series, and the slope of the regression curve ($S = \Delta\text{intensity}/\Delta\text{concentration}$) was characteristic for a certain antibody combination. The signal intensities resulting from time-resolved measurements were summarized as a *measurement series*. The concentration of the different proteins in a sample was calculated based on the linear regression of the calibration series. Bootstrap analysis can be used to evaluate the quality of these estimates, i.e., error introduced by the linear regression of standards, the linearity of measurements, and signal reproducibility.

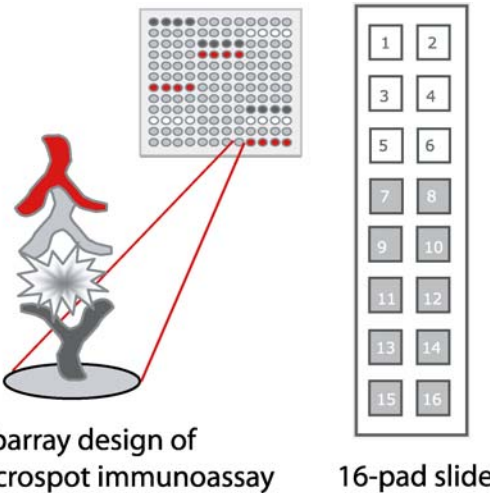


Fig. 4 Array layout for quantitative measurement in the multiplexed microspot immunoassay format. The 12 × 12 design allows accommodation of up to 15 different capture antibodies. Pad 1–6 was employed for incubation with six different concentrations of calibrator proteins. Samples were loaded on pad 7–16

3 Printing Antibody Microarrays for Quantitative Applications

3.1 Instrumentation

3.1.1 Noncontact Spotting

Modern microdispensing technologies exploit the principle of contact-free sample delivery. This way, sample delivery can be performed with high precision, and accurate and reproducible printing is absolutely mandatory for quantitative microarray applications. The microdispensing instrument releases the droplets in a predefined position at a distance of 0.4 and 0.6 mm from the surface of the microarray.

The majority of the noncontact spotters exploit a phenomenon known as the *converse piezoelectric effect*. In this instance, an electric pulse applied to a crystal results in its mechanical deformation. The mechanical stress produced by a piezo crystal in a liquid-filled glass capillary can force the liquid to form a droplet at the tip of the capillary. The exact spot size depends on the type of capillary, the orifice diameter, and experimental parameters, such as sample viscosity. Small adjustments of the drop size can be made by modification of the voltage and pulse duration applied to the piezo crystal. The small drop size, typically between 100 and 600 pl, makes this approach ideal for microdispensing. However, the small orifice of the capillaries, usually in the range between 50 and 100 μm, can easily be clogged by small particles derived from cell debris or dust. Consequently, piezo effect-based microdispensing systems require very pure samples as well as a housing to protect against environmental dust.

Multichannel microdispensing systems offer two different modes for sample delivery (Fig. 5). In the *simultaneous mode* all channels are spotted at the same time, thus producing independent subarrays in a distance defined by the offset of the capillaries. Thus, a certain capillary prints all samples within a subarray. On the contrary, in the *sequential mode* a single capillary is used to deliver a certain sample to a certain position of every subarray. The second capillary continues with printing the next sample, and so on. Printing with all capillaries in parallel in the simultaneous mode is faster than sample delivery in the sequential mode, but requires more material. In this mode, each sample of the subarray requires a corresponding position on the source

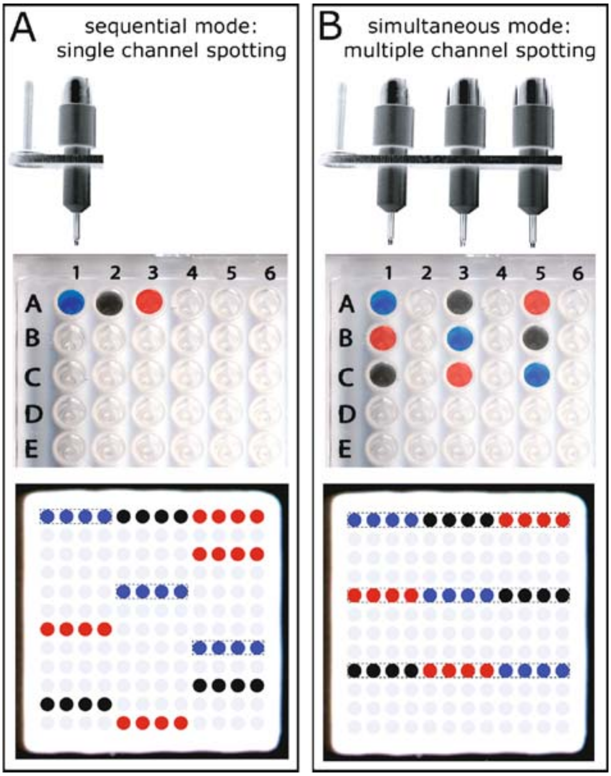


Fig. 5 Printing antibody arrays in the sequential mode (**A**) and in the simultaneous mode (**B**)

plate, since an individual needle cannot address a certain well several times. In contrast, a certain sample well can be addressed several times in the sequential mode, and thus allows spotting of many replicate spots per subarray from a single sample. However, the positional adjustments made by the instrument to realign the position of the needles over and over again consume substantial working time. In summary, printing in the sequential mode with a multichannel instrument takes about the same time as printing the same array with a single-capillary instrument. An additional important maintenance issue is the alignment of the capillaries in multichannel instruments. Only a correct alignment ensures that the spots of the microarray form a correct rectangular grid that will, e.g., match to the gal file of the analysis software.

Incorrect dispensing can result in smeared droplets, dislocated droplets, or drops with satellites, thus making the subsequent data analysis difficult or even impossible. Satellites are detected as multiple, but smaller, spots close to the actual spot position and reduce the quality of the subarray. Online pressure control within the piezo element is therefore crucial to guarantee that identical arrays are printed on all microarrays. Therefore a microarray printer must determine and maintain optimal pressure for sample delivery throughout the printing process. The BioChip Arrayer System (PerkinElmer, Boston, USA; las.perkinelmer.com) keeps the optimal pressure through an electronically controlled pressure transducer in the fluidic system. Clogged capillaries are immediately detected. Other piezo microdispensing systems maintain the optimal pressure through gravity control, and the sample delivery process is recorded online using a camera (Scienion AG, Berlin, Germany; www.scienion.com). A nano-plotting instrument employs a silicon glass-based capillary that is connected via a metal shaft to the fluidic system (GeSiM GmbH, Großerkmannsdorf, Germany; www.gesim.de). Piezo-spotters are available with up to 16 channels to accommodate the corresponding number of capillaries, and some manufacturers offer a flexible number of capillaries. However, a high number of channels increases the complexity of the printing process, and the need to undertake adjustments concerning the alignment of the capillaries for accurate sample delivery. Furthermore, the traditional inkjet technology has been exploited to build microdispensing instruments suitable for the production of protein microarrays (Arrayjet, Dalkeith, Scotland, UK; www.arrayjet.co.uk). This instrument is distinguished by a multichannel print head, in contrast with the limited number of glass or silicon capillaries from the formerly mentioned instruments.

Maintenance of Noncontact Microdispensing Systems

Microdispensing systems require regular maintenance, as well as observation of the rules for sample preparation and reagent quality. Samples have to be free of any particles that could possibly clog the capillaries, and for

Table 2 Liquids for microdispenser maintenance

Purpose	Buffer composition
Flushing the liquid path [a]	Isopropyl alcohol/chromatography water (70 : 30)
Flushing the liquid path [a]	Methanol/chromatography water (100 : 0; 80 : 20; 20 : 80)
Cleaning capillaries [b]	2 M Sodium hydroxide
Silane coating solution [b]	PlusOne Repel-Silane ES (Amersham Bioscience AB, Uppsala, Sweden)
Piezo element cleaning buffer 1	Tween 20/phosphate-buffered saline solution (5 : 95)
Piezo element cleaning buffer 2	DMSO/chromatography water (10 : 90)

[a] Care must be taken that the liquid completely fills the liquid system. Capillaries must be disconnected to prevent them from being clogged by any particles released through the cleaning process.

[b] Technically, glass tips of the capillaries are dipped into the cleaning solution to aspirate twice as much cleaning solution as sample is taken up for spotting runs. Next, the piezo element of the capillaries is turned on to generate an ultrasonic pulse to support the effectiveness of the cleaning solution. This procedure results in an extremely hydrophilic surface of the capillaries, which needs to be silanized to prevent adhesion of proteins. Silanization is best done by aspirating 100 μl methanol, followed by 100 μl air. The air is dispensed within 20 s while the capillaries are dipping for 10 s into a suitable silane coating solution. Dispensing air prevents any silane from moving into the capillaries. After drying the silane coating for 5 min the capillaries are flushed with water to wash away surplus silane.

the same reason highly purified water must be used in the fluidic system. The fluidic system must also be flushed regularly to prevent biological contamination, e.g., by algae. Flushing also removes gas bubbles resulting from evaporation (Table 2). Silanizing the glass capillaries can improve sample delivery, and thus reduce the risk of satellite formation. However, different and mostly customized strategies exist, and antibody arrays can successfully be printed with non-silanized needles, or with a silane coating on the inner and outer surface, or just on the outer surface of the capillary. Silanization requires cleaning of the capillaries under harsh conditions to remove any material deposited on the glass (Table 2). The capillary alignment and droplet morphology must be controlled after silanization of multichannel microdispensing instruments.

3.1.2
Contact Spotting

Contact spotting relies on steel pins for sample delivery. Pins can be categorized into two groups: solid pins and so-called split pins. The channel in a split pin serves as reservoir for the sample fluid, which is retained by capillary forces. Solid metal pins have to take up new sample after spotting of single drops, but split pins allow sample delivery to multiple slides before the next sample is taken up. As a major advantage, contact spotting is compatible with highly viscous samples, and samples of very different viscosity. The volume of the droplet strongly depends on the viscosity of the sample fluid. Thus, printing samples with different viscosities increases the spot-to-spot volume variation and complicates signal analysis. To overcome this problem, the spotted sample volumes can be quantified on the slide with an additional assay and the readout can be normalized [18]. The print option is always the *sequential mode*. The drop size depends on the pin type and size, and is in the range between high pico- and low nanoliters per spot. Generally, contact spotting is preferred as a technically less demanding approach, and when the high accuracy of noncontact spotting is not required. Pin and ring techniques are an additional spotting option. A ring is dipped into and immediately removed from the sample to form a liquid layer covering the area given by the ring diameter. A solid pin pushes the liquid film to a solid surface to form a spot. As its major advantage, pin and ring spotting is compatible with a wide variety of different sample viscosities and surface chemistries, and is therefore useful for explorative strategies.

Instrument Maintenance in Contact Spotting

The major maintenance issue in contact spotting is cleaning and controlling the integrity of the spotting pins. Before the next sample is taken up in a new spotting run, the pins are cleaned and dried automatically using an ultrasonic device. However, split pins have to be checked manually from time to time. Residual sample can remain within the channel of the split pins, which can be detected using a magnifier. A frequent problem is the deformation of pin tips due to mechanical stress from the printing procedure.

3.2 Slide Formats

Antibodies are either printed on single-pad slides for medium-scale protein profiling, or on multipad slides for the exact quantification of a limited number of proteins (Table 1). Discovery-type experiments focus on printing a high number of different antibodies in duplicate in the microarray format. However, precise quantitative analysis requires a higher number of replicate spots for the robust calculation of calibration slopes, and capture antibodies must be spotted in at least six replicate spots per subarray [2]. Identical subarrays are mostly printed on multipad nitrocellulose slides for their subsequent use in a multiplexed multititer plate format (Fig. 4).

Antibody arrays can also be printed in multititer plates with a capacity of 18 spots per well [7]. With respect to the number of different proteins in a proteome, the miniaturization of the protein microarray format to a nanodesign is desirable. Up to now, several strategies have been explored. For example, the production of novel nanostructured supports using electron-beam lithography was reported [19], and similar strategies were recently reviewed [20]. The proof-of-principle approaches demonstrated that antibody-based assays can be performed in vials ranging between 6 and 4000 al in volume. However, despite the fact that nanostructuring allows the density of antibody arrays to be increased, the identification of suitable antibody probes still presents the most limiting step before this approach can be up-scaled to match the size of the cellular proteome.

3.2.1 Surface Coatings

Surface coatings for the production of antibody microarrays should maintain the immobilized antibodies functional until use. In general, different architectures exist to generate a protein-friendly coating on a glass surface. Direct coating simply introduces functionality to the glass surface to improve the immobilization of proteins. In contrast, other materials employed for coating form three-dimensional structures on the glass surface, which were optimized to take up proteins and preserve them in a functional form. Planar and three-dimensional coatings differ mainly in their protein binding capacity. In any case, the antibody binding capacity of the membrane should be consistent over the complete surface to give reproducible results of good quality. Another important issue is the inactivation of the surface coating to block against nonspecific binding, and a variety of different blocking options exist.

Numerous surface coatings, detection strategies, and protocols were tested as solid supports for antibody microarrays. Angenendt and coworkers printed antibodies on arrays by contact spotting, and detection was performed with fluorescent dyes (Cy3/Cy5) in the visible range [21, 22]. The lowest detection limit was observed using polyacrylamide-coated slides, and antibody quantities in the low femtomole range were detected [21]. Certain surface coatings, poly-L-lysine and activated polystyrene, revealed a much higher detection limit for antibody microarrays than for protein microarrays [22]. This difference reflects the fact that some surface coatings bind proteins with high capacity but fail to preserve the functionality. Other surface coatings, e.g., nitrocellulose slides and dendrimer-coated slides, revealed comparable detection limits for antibody and protein microarrays, indicating that they are potentially useful for antibody microarray applications. However, signal detection involving nitrocellulose coatings is restricted by the strong autofluorescence of nitrocellulose in the visible range. This finding was confirmed by Guilleaume and coworkers [23] by correlating autofluorescence intensity with the thickness of a nitrocellulose coating. To detect fluorescence signals in the visible range, polyacrylamide-coated slides are better suited than those with nitrocellulose coatings. Also, glass slides with aldehyde silane, poly-L-lysine, or aminolysine consistently produced superior results using fluorescence detection in the visible range [24]. The potential of agarose-coated slides for antibody microarrays in the multiplex immunoassay format was explored as an alternative to polyacrylamide-coated glass slides. Agarose-coated slides are easy to prepare in constant quality. The sandwich detection of the chemokine MCP-1 (macrophage/monocyte chemotactic protein 1) was performed in the visible range with the fluorescent dye Cy3 [25]. Signals were uniform and of good reproducibility with respect to the intra- and interarray variation. Efforts to design the next generation of solid supports for antibody microarrays were undertaken by Wingren and coworkers [26]. They introduced a silicon-based macroporous solid support, and reported in terms of improved sensitivity, spot morphology, dynamic range, and reproducibility compared to nitrocellulose-coated glass slides.

3.2.2 Signal Detection

Wingren and coworkers also evaluated different combinations of solid support and fluorescent label by comparing the properties of three different fluorescent dye pairs: Cy3/Cy5, Alexa-647/Alexa-555, and ULS-biotin/ULS-flu [27]. All three fluorescent dye pairs revealed a discrepancy between the ratio of fluorescence signal intensity of the individual dyes. This observation makes a two-color approach less favorable for the relative comparison of two samples. Instead, Wingren and coworkers recommend a one-color approach for the analysis of complex samples on antibody microarrays [27]. The highest sensitivity was achieved on black polymer Maxisorb slides as solid support with ULS-biotin/NHS-biotin labeling (Table 3).

Indeed, fluorescent dyes are the most common strategy for signal detection in the field of antibody microarrays. A large number of fluorescent dyes are available as chemically activated compounds for labeling reactions, or linked to secondary antibodies and small molecules such as biotin. However, optimization is necessary to choose the best combination of surface coating and label. The widely used fluorescent dyes Cy3 and Cy5 can be detected with a standard microarray scanner. An alternative detection range is the NIR range at 700–900 nm, and suitable NIR dyes are available [28]. The low autofluorescence of biological compounds and of nitrocellulose in the NIR region reduces the background and thus increases the sensitivity [29]. Furthermore, fluorescence signals can also be quantified by exploiting the planar waveguide technology known for its high sensitivity [30].

Table 3 Comparison of surface coatings for the production of antibody microarrays

Author	Coating	Label	Assay[a]	Detection limit
Angenendt [21]	Hydrogel	Cy3/Cy5	Protein detection	Femtomole/spot
Angenendt [22]	Dendrimer nitrocellulose	Cy3/Cy5	Protein detection	Femtomole/spot
Steinhauer [26]	Nitrocellulose	Cy5	Protein capture	Femtomole/spot
Guilleaume [23]	Hydrogel Nexterion H	Alexa-647 (hydrogel) Alexa-532 (Nexterion H)	Protein detection	Picomole/spot
Seurynck-Servoss [24]	Aldehyde-poly-L-lysine aminosilane	Cy3	Protein capture	pg/ml
Lv [25]	Agarose-coated slides	Cy3	Protein capture	ng/ml
Wingren [27]	Black polymer maxisorb Nexterion H	NHS-biotin, ULS-biotin	Protein capture	Femtomolar
Korf [2]	Nitrocellulose	Alexa 680 (NIR)	Protein capture	pg/ml

[a] In *protein detection* assays the amount of protein deposited on the slide is detected using a fluorescently labeled antibody. In *protein capture* assays a fluorescently labeled protein is captured from the supernatant requiring functional antibodies on the solid surface.

3.3 Signal Analysis

In summary, specific signals on antibody arrays were mostly determined by employing fluorescent dyes, which are measurable in the visible or NIR range. The signals can be detected using scanning instrumentation with a resolution sufficient for microarrays consisting of spots of size 100–300 μm. Signal intensities of single spots can be quantified using standard software, and the mean or median signal intensities are directly correlated with target protein expression.

Various software tools, e.g., GenePix, Quantarray, or ScanAlyze, can be used for the analysis of signal intensities. The results of the image analysis are usually stored in text format files;, the GenePix software calls them ".gpr" files (Table 4). The information on how individual antibodies are assembled within the grid of capture antibodies is summarized as a "gal" file. This gal file is used to connect the information on the spot localization (row, column) with a certain capture antibody name. Practically, the image analysis software places a grid on top of the microarray image with circular features matching the position of each single spot. Before individual spot intensities can be calculated, the circular features must be aligned to match with not perfectly positioned spots. In addition, the signal analysis software of-

Table 4 Quantitative analysis of measurements on multipad slides with *Quantpro*[a]

Input	Data format	Information
Antibody information	txt file	Capture antibody ID (AB1) Detection antibody (AB2) Calibrator protein
Array design	gal file	Capture antibody ID (AB1) Capture antibody position (row/column) Multipad (block) assignment
Microarray signal analysis	gpr file	Slide images gal file
Slide description	txt file	Purpose (standard or sample) Concentration (for standards) Dilution (for samples) Detection antibody (AB2) Time points Link to gpr file data

[a] http://www.dkfz.de/mga/Quantpro

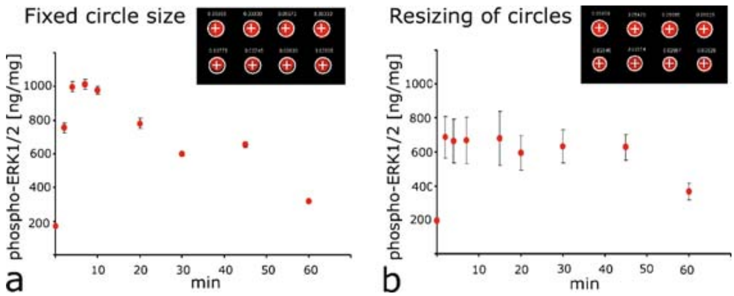


Fig. 6 Microarray data analysis with fixed circular features (**b**) compared to resized circular features (**a**). A subarray image is shown. Formation of phospho-Erk1/2 was calculated from the signal intensities of the same slide analyzed according to **a** or **b**. *Error bars* are indicated

fers an option to resize the circular features to cover the strongest pixels of each single spot. However, the automated resizing option of the software is not useful for the quantification of weak signals, or spots of irregular morphology (Fig. 6). Instead, employing a fixed circle size defined by the area of the large spots of the array can improve the final quality of the data, in particular the standard deviation. The fixed circle method was also employed by other groups [27]. In general, the median is a good approach to compensate for the inhomogeneous distribution of signal intensities, especially those relating to confounding factors like dust, scratches, or spot size variation, which impact the mean more strongly. The median offers a robust assessment of the majority of pixels in a spot, and thus serves as a basis to calculate the corresponding protein concentration. Although local background can be subtracted from the median of the signals, this correction can increase the noise level, especially for signals of very low or very high intensity. For instance, very bright spots often affect the neighboring background readings, and therefore the local background often correlates with signal intensity. An additional option is to increase the number of replicate spots to improve the robustness of the readout [31].

4
Software Tools for Data Analysis

4.1
Converting Raw Data into Protein Concentration

To analyze antibody microarray data obtained by time-resolved quantitative measurements on multipad slides, a tailored software tool, *Quantpro*, was developed [2]. This software package is based on the statistical computing environment of R (http://www.r-project.org). Features of the software include the quantification of the proteins of interest, the visualization of time-resolved measurements, and also the evaluation of different antibody pairs. The *Quantpro* tool uses a simple graphical user interface (GUI) that allows access to the most important functions, and is available at the website (http://www.dkfz.de/mga/Quantpro). The preparation of data and experimental information for the software package has to be made via several text files which are read and analyzed by the *Quantpro* package. Experimental information and capture antibody layout are summarized as tab-delimited files (Table 4), and are required along with the gpr files from the microarray image analysis. Information on sample assignment, standard concentration, dilution factor, or detection antibody in a tabular format is summarized as the so-called *slide description file*. Furthermore, information on capture and detection antibodies is annotated as the *antibody combination file*. Commercial software is also available for protein microarray data analysis (www.vigenetech.com).

4.2 Evaluation of Antibody Pairs

The signal intensities from a calibration series can be employed for the estimation of the reproducibility of individual measurements. For this purpose, a bootstrap-like approach was developed to reduce the number of measurements employed for the calculation of a calibration series by a certain percentage of the spots. The data of these spots are considered as artificial measurements to calculate the predictive power of the reduced calibration series. After calculating the concentration of the artificial measurements, the absolute deviation from theoretically expected concentration is recorded. This is repeated 100 times by remixing the values for the calibration series and measurement series. The resulting average deviation is reported as the *accuracy* of the antibody pair. Furthermore, the *dynamic range* (i.e., order of magnitude of the fluorescence intensity readout, as computed by the slope of the fitted linear regression times the calibration range) is used as an additional quality score. As a part of the *Quantpro* software package, this approach was implemented to computationally assess the quality of different capture and detection antibody combinations, and to evaluate their efficiency in quantitative measurements. The antibody-pair plot tool (Fig. 2) summarizes different quality measures as a visual output, such as dynamic range and accuracy. Both parameters are based on the signal intensities of a certain calibrator protein measured in a dilution series, and they are specific for a certain detection/capture antibody combination.

5
Outlook

Recent progress in the field of antibody microarrays has demonstrated the potential of this platform for the analysis of signal transduction pathways. Two principally different types of antibody arrays were employed for proteome analysis. The first approach is based on medium-density antibody microarrays with a few hundred up to a thousand different antibodies. Medium-density arrays, also available from commercial vendors, were applied in the exploratory phase of a project for proteome profiling. The other type of antibody microarray approach comprises only a few different antibodies for the focused analysis of selected signaling pathways. This type of array is beneficial for the validation of data from large-scale experiments, and for the quantitative analysis of time-resolved changes of protein abundance and protein phosphorylation. In terms of technological aspects, robust protocols relying on standard equipment were established, and the antibody microarray technology has matured to a widely applicable platform. Both platforms are valuable to complement other technologies, such as mass spectrometry, but will also be useful for the validation of functional high-throughput screens. However, the availability of specific and well-characterized antibodies will continue to play the most important role in this field.

The analysis of protein microarray data still presents an issue of major importance. Microarray software tools were originally developed for the field of DNA microarrays and later also employed for the analysis of protein microarray data. However, there are fundamental differences between DNA and protein microarrays. First, the density of DNA microarrays is significantly higher and the readout is always based on the relative comparison of two differently labeled samples per spot. For this reason, the analysis routine was adapted to the specific needs of protein microarrays, especially when a quantitative readout is generated. A software package, *Quantpro*, facilitates quantitative measurements in a multiplexed sandwich format, data processing, and graphical display of experimental data. *Quantpro* also includes an additional tool for the evaluation of the accuracy and dynamic range of single antibody pairs and of antibody pairs in a multiplexed setting. Guidelines for the validation of binding reagents in protein microarrray experimentation are necessary, as well as qualitative and quantitative standards to validate the performance of antibodies in a multiplexed setting. The introduction of the antibody-pair plot tool was a first step in this direction [2].

Antibody microarrays using fluorescence detection will also dominate future applications. However, alternatives to both strategies exist and further advancements with respect to miniaturizing the array format to a nanoarray design are promising. Considerable progress was also made in the field of label-free detection, although this type of microarray readout is currently not realized in routine applications. In summary, quantitative antibody microarray applications have matured to a routine tool and are now available for the quantitative analysis of signal transduction networks.

References

1. Ekins R (1996) J Clin Ligand Assay 19:145
2. Korf U, Derdak S, Tresch A, Henjes F, Schumacher S, Schmidt C, Hahn B, Lehmann W, Poustka A, Beissbarth T, Klingmueller U (2008) Proteomics (in press)
3. Nielsen UB, Geierstanger BH (2004) J Immunol Methods 290:107
4. Haab BB, Dunham MJ, Brown PO (2001) Genome Biol 2: RESEARCH0004. Epub 22 Jan 2001
5. Paweletz CP, Charboneau L, Bichsel VE, Simone NL, Chen T, Gillespie JW, Emmert-Buck MR, Roth MJ, Petricoin EF, Liotta LA (2001) Oncogene 20:1981
6. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, Lawrence TS, Chinnaiyan AM (2001) Cancer Res 61:7585
7. Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G, Sorger PK (2003) Proc Natl Acad Sci USA 100:9330
8. Gembitsky DS, Lawlor K, Jacovina A, Yaneva M, Tempst P (2004) Mol Cell Proteomics 3:1102
9. Ivanov SS, Chung AS, Yuan ZL, Guan YJ, Sachs KV, Reichner JS, Chin YE (2004) Mol Cell Proteomics 3:788
10. Usui-Aoki K, Shimada K, Nagano M, Kawai M, Koga H (2005) Proteomics 5:2396
11. Duffy HS, Iacobas I, Hotchkiss K, Hirst-Jensen BJ, Bosco A, Dandachi N, Dermietzel R, Sorgen PL, Spray DC (2007) J Biol Chem 282:9789
12. Chen SM, LaRoche T, Hamelinck D, Bergsma D, Brenner D, Simeone D, Brand RE, Haab BB (2007) Nat Methods 4:437
13. Bartling B, Hofmann HS, Boettger T, Hansen G, Burdach S, Silber RE, Simm A (2005) Lung Cancer 49:145
14. Celis JE, Moreira JMA, Cabezon T, Gromov P, Friis E, Rank F, Gromova I (2005) Mol Cell Proteomics 4:492
15. Smith L, Watson MB, O'Kane SL, Drew PJ, Lind MJ, Cawkwell L (2006) Mol Cancer Ther 5:2115
16. Wang J, Laschinger C, Zhao XH, Mak B, Seth A, McCulloch CA (2005) Biochem Biophys Res Commun 330:123
17. Lin HJ, Hsieh FC, Song H, Lin J (2005) Br J Cancer 93:1372
18. Löbke C, Laible M, Rappl C, Sahin Ö, Arlt D, Wiemann S, Poustka A, Sültmann H, Korf U (2008) Proteomics 8:1586–1594
19. Ghatnekar-Nilsson S, Dexlin L, Wingren C, Montelius L, Borrebaeck CAK (2007) Proteomics 7:540
20. Wingren C, Borrebaeck CAK (2007) Drug Discov Today 12:813
21. Angenendt P, Glokler J, Murphy D, Lehrach H, Cahill DJ (2002) Anal Biochem 309:253
22. Angenendt P, Glokler J, Sobek J, Lehrach H, Cahill DJ (2003) J Chromatogr A 1009:97
23. Guilleaume B, Buness A, Schmidt C, Klimek F, Moldenhauer G, Huber W, Arlt D, Korf U, Wiemann S, Poustka A (2005) Proteomics 5:4705
24. Seurynck-Servoss SL, White AM, Baird CL, Rodland KD, Zangar RC (2007) Anal Biochem 371:105
25. Lv LL, Liu BC, Zhang CX, Tang ZM, Zhang L, Lu ZH (2007) Electrophoresis 28:406

26. Steinhauer C, Ressine A, Marko-Varga G, Laurell T, Borrebaeck CAK, Wingren C (2005) Anal Biochem 341:204
27. Wingren C, Ingvarsson J, Dexlin L, Szul D, Borrebaeck CAK (2007) Proteomics 7:3055
28. Schutz-Geschwender A, Zhang Y, Holt T, McDermid D, Olive MD (2004) http://www.licor.com/bio/PDF/IRquant.pdf
29. Loebke C, Sueltmann H, Henjes F, Wiemann S, Poustka A, Korf U (2007) Proteomics 7:558
30. Pawlak M, Schick E, Bopp MA, Schneider MJ, Oroszlan P, Ehrat M (2002) Proteomics 2:383
31. Smyth GK, Michaud J, Scott HS (2005) Bioinformatics 21:2067

Adv Biochem Engin/Biotechnol (2008) 110: 177–194
DOI 10.1007/10_2008_104
© Springer-Verlag Berlin Heidelberg
Published online: 2 August 2008

Using Aptamers to Study Protein–Protein Interactions

Parag Parekh · Jennifer Martin · Yan Chen · Dalia Colon · Hui Wang ·
Weihong Tan (✉)

Department of Chemistry and Department of Physiology and Functional Genomics,
Shands Cancer Center, Center for Research at the Bio/Nano Interface,
Genetics Institute and McKnight Brain Institute, University of Florida,
Gainesville, FL 32611-7200, USA
tan@chem.ufl.edu

Abstract The emerging science of systems biology focuses on the systematic study of complex interactions in whole biological systems. A systemic, or integrative, methodology is employed as the chief means of discovering new properties and understanding the aggregate of processes that occur in a biological system. Accordingly, the Human Genome Project has provided a complete map of genes and resultant proteins corresponding to their function. Protein–protein interactions are important pieces of this biological tapestry, and understanding how they work cooperatively in a cell will result in a better understanding of the whole organism. To accomplish this objective, we report the use of DNA/RNA aptamers as a novel tool for the study and elucidation of protein–protein interactions, both in vivo and in vitro.

Keywords Aptamers · Fluorescence anisotropy · FRET · Protein–protein interactions

1 Introduction

All proteins are interconnected by networks within cells. The interaction between them involves electrostatic forces, Van der Waals forces, hydrogen bonds and hydrophobic effects, as well as water-mediated interactions between amino acid residues on the surface of the proteins [1]. These interactions may occur on identical (homo-oligomer) or non-identical (hetero-oligomer) proteins. Moreover, protein–protein interactions are dependent on various factors, including the shape and surface of proteins. Therefore, from a systems biology point of view, understanding protein–protein interactions is an essential step towards an integrative understanding of the whole organism.

Several methods of measurement are used to study protein–protein interactions. These include affinity purification-based methods and fluorescence-based FRET and BRET assays. These methods, together with nuclear magnetic resonance (NMR), or STINT-NMR, which is a new method for mapping in-cell interactions, can all be used in vitro or in vivo. Other methods include quantitative surface plasmon resonance (SPR), crystal structure determination, calorimetric study for quantitative analysis of protein interactions, atomic force microscopy (AFM) for detection and analysis and protein microarrays for detection and selectivity of protein interactions can only be used in vitro. Genetic test systems involving yeast 2 hybrid assays (Y2H), or the mating-based split ubiquitin system (mbSUS), are used solely for in vivo measurements [2]. This chapter focuses on the use of oligonucleotide aptamers in the study of protein–protein interactions and looks at their applications, both in vivo and in vitro.

2
Aptamer Fundamentals

Aptamer (from the Latin *aptus*, meaning fitting) is a term applied to oligonucleotides that specifically bind to a target. The target can range in size from small molecules, proteins, cells and tissues to a whole organism. Aptamer–target interactions are based on molecular recognition events due to unique tertiary binding structure of the aptamers based on their sequence. These interactions are dependent on Van der Waals forces, hydrogen binding and electrostatic forces between the target and its aptamer. The binding affinity of aptamers to its specific target varies in micro to picomolar range. In contrast to antibodies, aptamers are the more promising alternative for target identification and validation since they can be easily labeled and modified during chemical synthesis. This gives molecular engineers the advantage of constructing aptamers with various signal transduction mechanisms, thus enabling a broad range of sensing strategies. Hence, unlike antibodies, aptamers can act as both the selective target-capture agent and the signal-transduction agent. As such, aptamers, which have batch-to-batch reproducibility and an unlimited shelf life, can be generated for targets that do not generate an immune response, such as toxins and explosives like ricin, tri-nitro toluene (TNT), etc.

Identification of oligonucleotide aptamers for a specific target is accomplished by a process known as in vitro selection, or SELEX (systematic evolution of ligands by exponential enrichment), a selection strategy independently developed in 1990 by Szostak and Gold [3,4]. SELEX (Fig. 1) involves a library of random nucleotides (DNA or RNA) with defined primer sequences at the 5′- and 3′-end regions of the sequences. The library contains approximately 10^{14} to 10^{15} possible different starting sequences; thus, each member of the library is a unique single-stranded sequence. This library is incubated with the molecule of interest under conditions as defined by the target. The nucleic acid ligands adopt different conformations and will interact correspondingly with the target. The low-affinity binding species are washed away, and the molecules bound to the target are eluted and retained. This eluted pool of nucleic acid species enriched with higher affinity

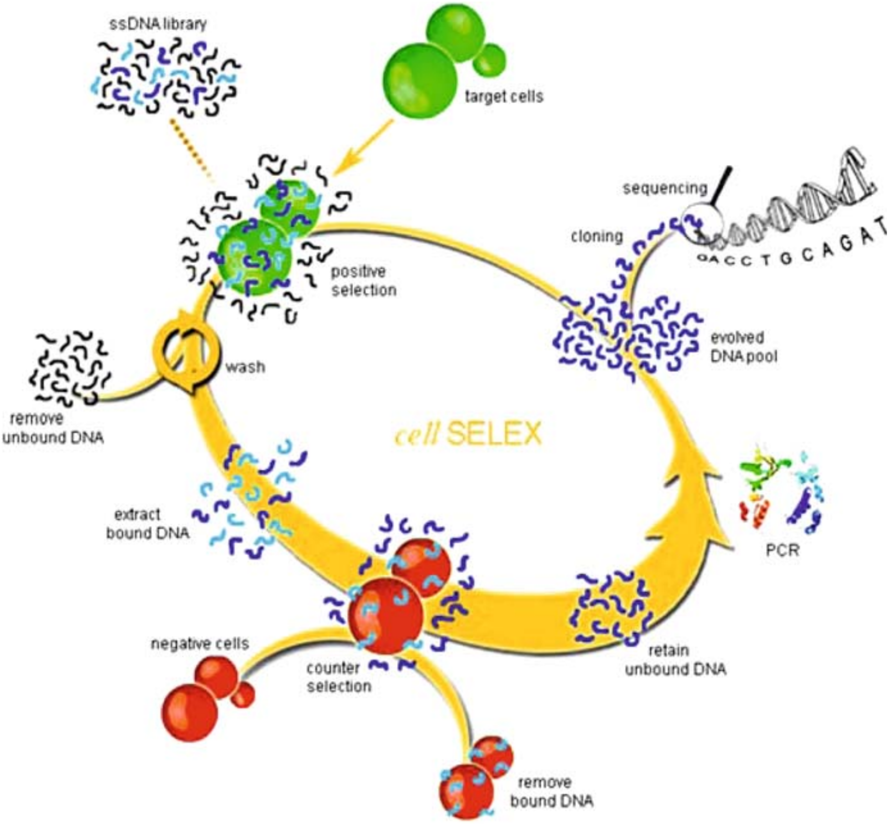


Fig. 1 Schematic representation of the cell-based aptamer selection. Briefly, the ssDNA pool was incubated with CCRF-CEM cells (target cells). After washing, the bound DNAs were eluted by heating to 95 °C. The eluted DNAs were then incubated with Ramos cells (negative cells) for counterselection. After centrifugation, the supernatant was collected, and the selected DNA was amplified by PCR. The PCR products were separated into ssDNA for next-round selection or cloned and sequenced for aptamer identification in the last-round selection [5]

sequences are amplified with a PCR or RT-PCR for DNA or RNA, respectively, and used as a starting point for the next cycle of the selection process. In early cycles, the oligonucleotides with no affinity for the target are eliminated. Later cycles are performed under increasingly stringent conditions. This causes a competition among the sequences for particular binding sites on the target, leading to a systematic evolution of the nucleic acid species with higher affinity for the target after each cycle. Aptamers with high specificity for the target generally require 15–25 cycles of selection. As indicated previously, this depends on the stringency of the conditions utilized during each round of selection and the affinity interaction between target and potential aptamer candidates. Aptamer sequence information is obtained by cloning and sequencing after the final round of selection. Different biochemical methods are then employed to verify and quantitate the affinity and specificity of the aptamers generated from the SELEX process.

3 Examples of Aptamers as Tools for the Study of Protein–Protein Interactions

3.1 In vitro

In order to implement aptamers in the study of protein–protein interactions, the first step requires us to develop the capacity of aptamers to detect proteins. To accomplish this, our initial work used two very important protein targets, namely, platelet-derived growth factor (PDGF) and thrombin, as a proof of principle towards the detection of proteins in a complex biological environment. PDGF is a growth factor that regulates cell growth and division and plays an important role in blood vessel formation. Its variants have been implicated in various cancers and embryonic developmental disabilities. PDGF is a dimeric glycoprotein composed of ligands A–D. Four homodimers and a heterodimer, AB, are active isoforms of PDGF.Thrombin is a coagulation protein that has many effects in the coagulation cascade. It is a serine protease that converts soluble fibrinogen into insoluble strands of fibrin, as well as catalyzing many other coagulation-related reactions.

The PDGF aptamer previously identified by Green et al. [6] (Fig. 2) was used to demonstrate a high affinity for PDGF-BB. It also had distinguishable affinities for three homodimeric PDGF isoforms. We developed a fluorescence quenching assay to detect PDGF in biological samples [7]. To accomplish this, the sample is presented to multiple reporter binding sites. Each binding site is comprised of two partially hybridized molecules. In our case, these are represented by a fluorophore and a quencher attached to opposite ends of an aptamer, which we term molecular aptamer beacon, or MBA. This MBA

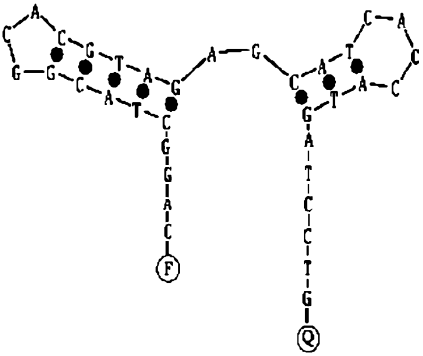


Fig. 2 Structure of the PDGF aptamer: the stable conformation of the aptamer under physiological conditions in the absence of PDGF. Upon PDGF–protein binding, the aptamer forms a close-packed, tight structure, which reduces the distance between the two termini of the aptamer and causes fluorescence quenching. Different molecular variants of PDGF dimers will bind with the aptamer with differences not only in binding affinity, but also in binding avidity. The avidity can easily be determined in the signaling step based on FRET between the fluorophore (F) and the quencher (Q) [7]

has fluorescence in buffer when the two ends are apart. In the presence of PDGF, its target, these two ends come into close proximity. This action, which is caused by fluorescence based resonance energy transfer (FRET), rapidly quenches the fluorescence. FRET is fluorescence energy resonance transfer. FRET between the fluorophore and quencher takes place when the distance between them is less than 5 nm. The aptamer has different affinities for various isoforms of PDGF, this was also reflected in the quenching assay shown in Fig. 3. It is also possible to distinguish the molecular variants of PDGF using a single-step MBA fluorescence quenching assay. In fact, at a very basic level, this is a protein–protein interaction (homo- and heterodimerization) in and of itself, utilizing MBA to distinguish different isoforms of target protein.

The fluorescence quenching assay to detect PDGF in biological samples was, by in large, successful. It is generally accepted that fluorescence enhancement assays are more sensitive than quenching assays; therefore, we next developed a novel method of using two separate single fluorophore-labeled aptamers. The working principle here is that an enhancement in fluorescence is observed due to FRET using a pair of dyes. The process may be described as follows. Cy3 dye (red) was conjugated on the 3′-end of one aptamer, and Cy5 (far-red) was conjugated on the 5′-end of other single-labeled aptamer. These single-labeled aptamers were then mixed in an equimolar ratio. In this method, if the two aptamers having different labels are bound to the same target PDGF molecule, the FRET pair will be in close proximity. Under these conditions, the donor Cy3 fluorophore is quenched by the acceptor Cy5 when excited at the peak wavelength of Cy3 absorption, but a simultaneous increase in Cy5 fluorescence is monitored [8]. A compari-

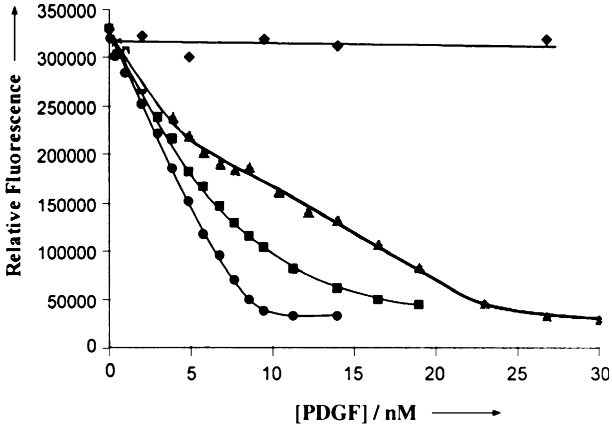


Fig. 3 Dose-response curves of PDGF variants: fluorescence signals of MBA for PDGF-AA (*triangles*), PDGF-AB (*squares*), PDGF-BB (*circles*), and denatured PDGF-BB (*diamonds*). The concentration of the MBA was 10 nM [7]

sion of fluorescence enhancement (Fig. 4B) using two singly labeled aptamers and quenching assays (Fig. 4A) shown by the standard double-labeled MBA reveals the sensitivity of the assays employed.

In fact, this assay has now been developed even further and can be used to detect PDGF in real-time monitoring without the washing and separating steps. This was achieved using an aptamer labeled with pyrene at both ends. When bound to PDGF, a light switching excimer is formed that changes its fluorescence from 400 to 485 nM. An additional benefit of using this strategy is the elimination of most biological background using time-resolved measurements. This results from the fact that the pyrene excimer has a longer fluorescence lifetime (~ 40 ns) compared to background (~ 5 ns) [9] (Fig. 5).

A large number of proteins in their active form are oligomers, either homo- or hetero- oligomers. Also, various proteins exist in different isoforms. Application of same principles utilizing aptamers can be applied to study multimeric proteins.

Finally, the use of DNA-based applications is challenged by one major drawback. Specifically, nuclease degradation of the DNA probe results in the tendency to false positives. This can be overcome by improvement in the stability of MBAs, using modified DNA bases like locked nucleic acids (LNA) or L-DNA bases which are not recognized by nucleases. The assay was further developed with modified DNA backbone without affecting its specificity. The easy modification of DNA allows for the possibility of multiplexing the assay using various known fluorophore–quencher systems.

Fundamentally, the work involving PDGF detection strongly indicates that these aptamers do possess the ability to recognize protein targets. We have shown that our aptamers also have potential as tools for more complex

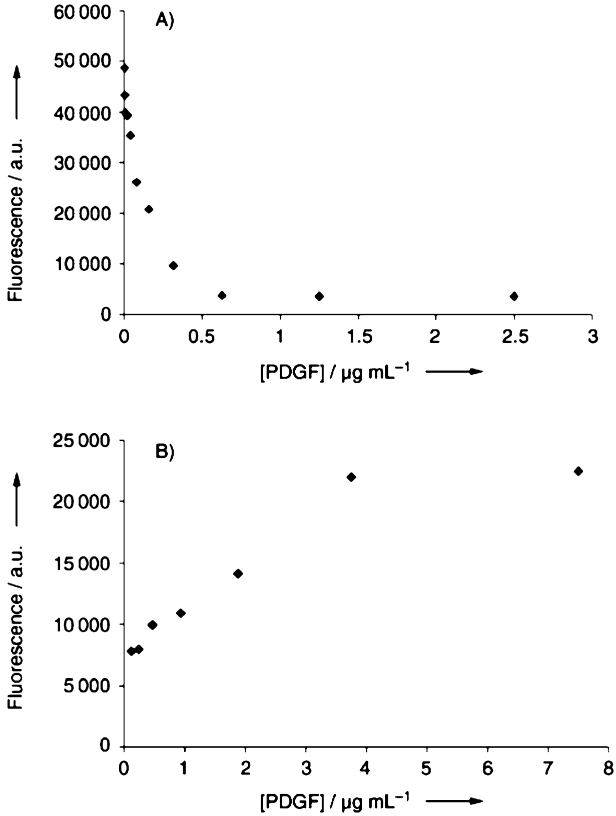


Fig. 4 A comparison of quenching and enhancement formats of the MBA-based FRET assay. Increasing concentrations of PDGF-BB were incubated in the standard fluorescence–quenching assay buffer (final volume 100 l) with **A** an MBA labeled with Dabcyl quencher at the 3′-end and fluorescein at the 5′-end (fluorescence measurements were made at the excitation and emission maxima for fluorescein) or **B** with an equimolar mixture of the PDGF aptamer labeled with Cy3 at the 3′-end and a PDGF aptamer labeled with Cy5 at the 5′-end (fluorescence measurements were made at an excitation maximum of Cy3 and emission maximum of Cy5) [8]

protein–protein interactions. One very important example of this would be the detection of thrombin protein–protein interactions in the blood coagulation cascade, which involves a complex series of biological reactions resulting in the formation of a blood clot. The importance of this application lies in the inhibition of a specific target in this cascade, which has been the major theme in the development of safe anticoagulation drugs. Most drugs target thrombin or factor Xa in this cascade, and selection of a DNA aptamer inhibiting thrombin was first reported in 1992 [10]. Over the next 15 years, thrombin aptamers have been an indispensable tool for elucidating the protein–protein interactions in this very important biological signal cascade.

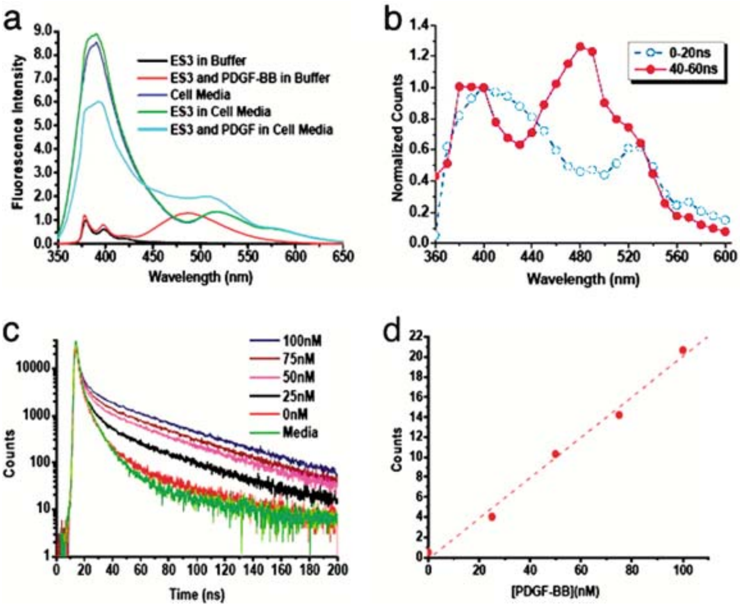


Fig. 5 Monitoring PDGF in dyed cell media. **a** Steady-state fluorescence spectra of cell media: 200 nM ES3 in cell media; 200 nM ES3 and 50 nM PDGF-BB in cell media; 200 nM ES3 in Tris·HCl buffer; and 200 nM ES3 with 50 nM PDGF-BB in Tris·HCl buffer. **b** Time-resolved fluorescence spectra of 200 nM ES3 and 50 nM PDGF-BB in cell media at different time windows after the excitation pulse, 0–20 ns (*blue*) and 40–60 ns (*red*). **c** Fluorescence decays of 200 nM ES3 in cell media with various concentrations of PDGF-BB. **d** The response of fluorescence intensity to the change of protein concentration [9]

We used this aptamer probe as a tool to study these protein–protein interactions by affinity capillary electrophoresis (ACE) [11]. ACE refers to a collection of techniques in which high affinity is used with capillary electrophoresis (CE) to determine analytes. To explain, thrombin has two positive-charged sites, termed exosite 1 and exosite 2, on opposite sides of the protein. Thrombin was labeled with 6-carboxyfluorescein, and the thrombin–protein interaction was probed in a competitive assay in which an antigen (thrombin) was mixed with fluorescently labeled antigen (thrombin 6-carboxyfluorescin). Then, a limiting concentration of thrombin aptamer was employed to study different interactions between thrombin and anti-thrombin III. This aptamer-based ACE assay was used to quantify and monitor the thrombin–anti-thrombin III interaction in real time.

In the ACE assay, a 15 mer Thrombin DNA aptamer, which specifically binds exosite I, was selected. This aptamer adopts a G-quadruplex structure when bound to thrombin. In the presence of K^+ and Ba^{2+}, it can be separated into two peaks in CE, which correspond to (1) linear aptamer (L-apt) and (2) thrombin binding G-quadruplex structure (G-apt). Increasing the concentration of anti-thrombin, AT III, increased the area of G-apt peak (Fig. 6). This means that binding of AT III might cause a conformational change in thrombin such that the binding with aptamer at exosite 1 would be rendered

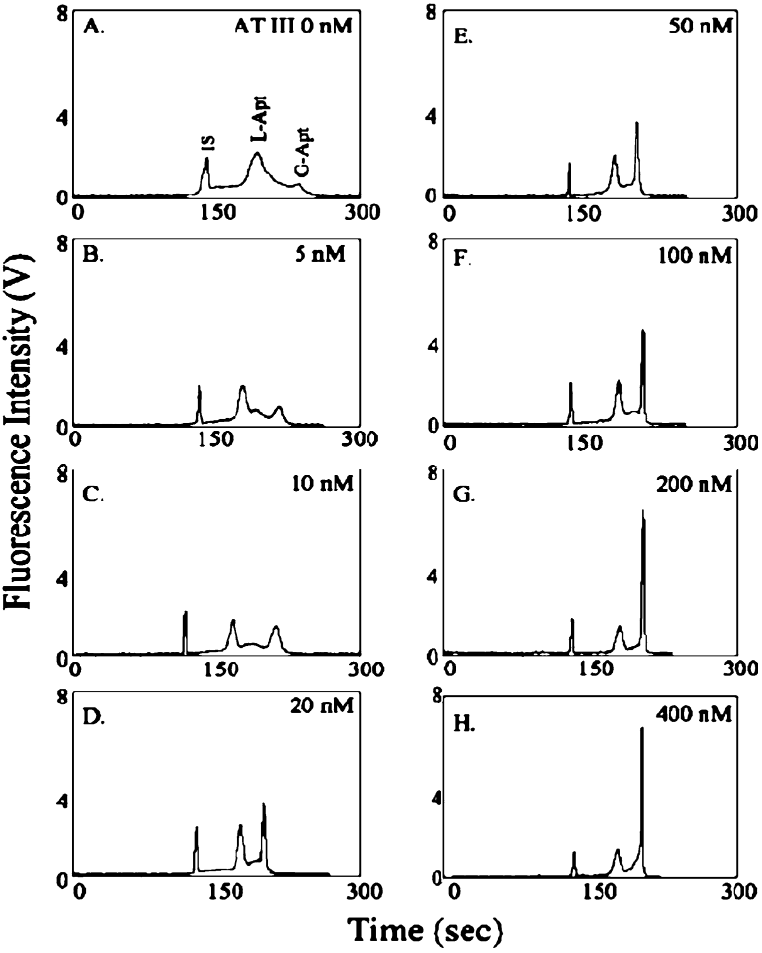


Fig. 6 Determination of AT III using aptamer-based ACE. Electropherograms obtained for 200 nM aptamer with 200 nM thrombin and various concentrations of AT III. In electropherograms (**A**–**H**), AT III concentrations were 0, 5, 10, 20, 50, 100, 200, and 400 nM, respectively. Aptamers were mixed with thrombin and incubated for 60 min at room temperature. The desired concentrations of AT III were mixed with aptamer/thrombin complex solutions. The resulting samples added fluorescein as an internal standard to 10 nM with incubation for another 60 min. Separation was carried out at a constant electric field of 500 V/cm. The samples were injected into the capillary (total length, 50 cm; effective length, 25 cm) hydrodynamically for 10 s; voltage of 500 V/cm was applied to drive the separation. Peak areas were corrected for variations in injection volume by dividing by the area of the internal standard peak [11]

unstable. The reaction of thrombin and AT III was completed within 10 min and was monitored in real time (Fig. 7a). The limit of detection (LOD) was determined to be 2.1nm based on a calibration curve of peak areas of free G-apt vs. AT III concentration (Fig. 7b).

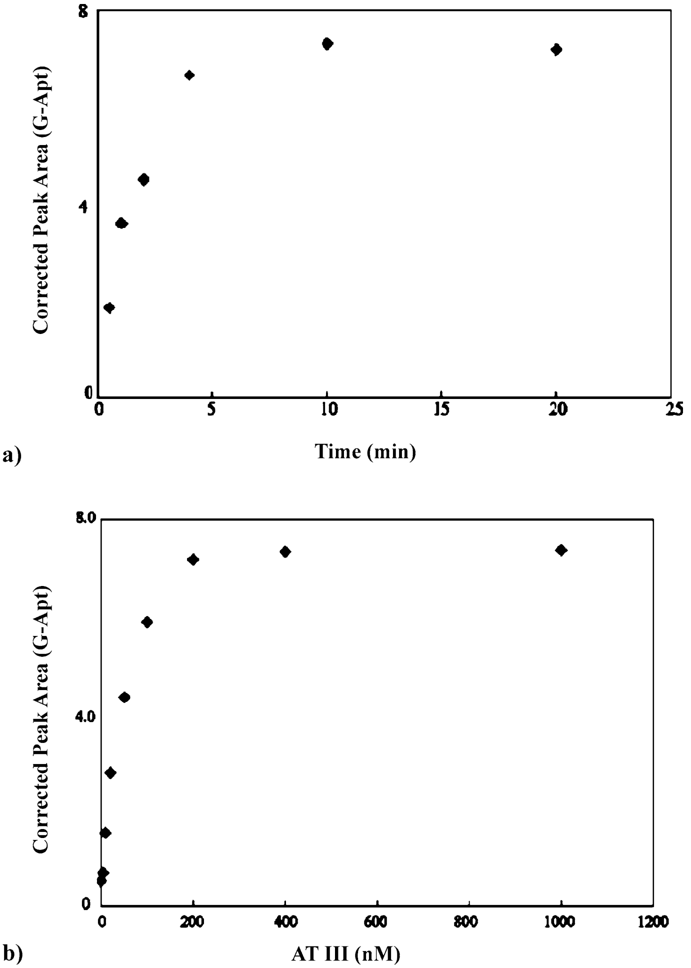


Fig. 7 **a** Using thrombin binding aptamer to monitor thrombin–AT III interaction. A solution of mixed aptamer and thrombin was incubated in electrophoresis buffer for 60 min. Then AT III was mixed with the aptamer/thrombin complex solution and incubated for another 60 min. The final concentration of aptamer, thrombin, and AT III was 200 nM. After the resulting sample incubation for different times, rapid injection into the capillary and the separation were carried out at a constant electric field of 500 V/cm. **b** Calibration curve constructed using samples containing 200 nM aptamer, 200 nM thrombin, and various concentrations of AT III (0–1.0 M). Peak area of G-Apt was corrected for variations in injection volume by dividing by the area of the internal standard peak [11]

The same model system was later studied in real time with aptamer-based assays without any modifications to the two interacting proteins. This allowed true real-time monitoring of interactions between two proteins in their unaffected biological state. In this assay, there is a "bait" protein, which is the aptamer binding protein, and a second "prey" protein, both unmodified. Two signal transduction strategies are used to monitor the changes in the labeled aptamer in order to reveal the protein–protein interactions: fluorescence based resonance energy transfer (FRET) and fluorescence anisotropy.

For the FRET assay, dual-labeled aptamers with a fluorophore and a quencher were used. Binding of aptamer to the "bait" protein caused a quenched fluorescence. Then, the subsequent binding of the "prey" protein to the "bait" protein resulted in a restoration of fluorescence (sequential binding). This amounts to the inhibition of aptamer binding, thus preventing quenching (co-incubation) (Fig. 8a). These signaling MBA's can easily be synthesized to form intramolecular FRET. On the other hand, through careful design and engineering, a few bases can be added to aptamer sequences to introduce conformational changes on binding to its target, which results in protein-dependent fluorescent changes in FRET [12, 13]. FRET-based assay uses the aptamer/thrombin as the "bait" solution and a sulfated fragment of C-terminal 13-residue of Hirudin (HirF), which is a peptide with anticoagulation properties, as the "prey" protein. The addition of HirF causes

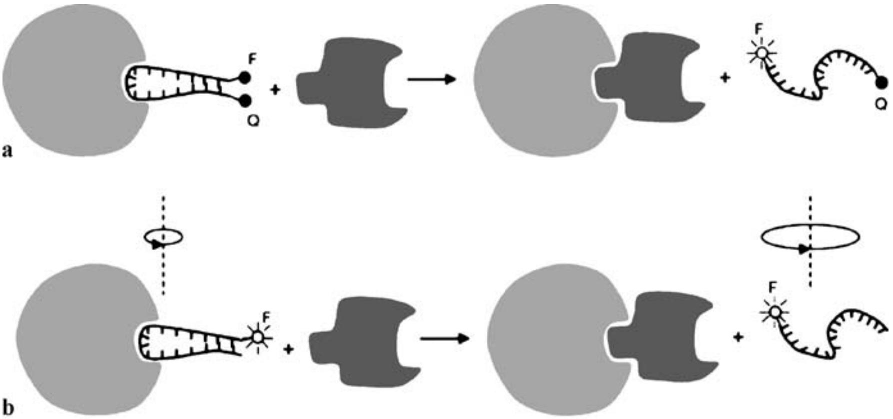


Fig. 8 Dye-labeled protein-binding DNA aptamers reporting protein–protein interactions. **a** Dual-labeled aptamer with a fluorophore and a quencher. The folded form of the aptamer results in a quenched fluorescence when it binds to the bait protein. The bait–prey protein interaction causes release of the aptamer from the bait protein, leading to a restored fluorescence. **b** Single-labeled aptamer. When bound to the much larger bait protein, the aptamer displays slow rotational diffusion. The interaction between bait and prey proteins displaces the aptamer. The unbound aptamer has much faster rotational diffusion. The change in the rotation rate is reported by fluorescence anisotropy of the dye molecule [13]

a sharp increase in fluorescence within seconds, indicating that both bind to the same site of thrombin (Fig. 9). Addition of another antibody, AHT, causes no significant change in fluorescence, indicating that it binds thrombin, but not at exosite I. Another protein, AT III, showed a slow trend toward increasing signal, which confirms that binding is a multi-step, covalent bond forming process.

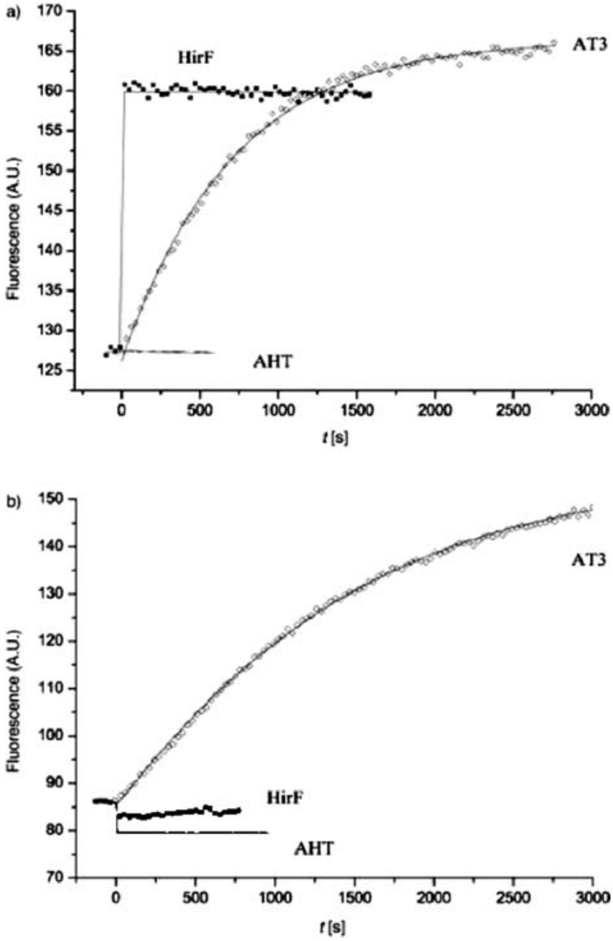


Fig. 9 Dual-labeled aptamers (FQ = fluorescence–quencher labeled) for thrombin/protein interactions. **a** In a solution of mixed 100 nm FQ-15Ap and 100 nm a-thrombin, 200 nm AT3 (■), 500 nm HirF (◇) or 300 nm AHT (△) was added at 0 s, and fluorescence of 6-FAM was continuously monitored. **b** In a solution of mixed 100 nm FQ-27Ap and 100 nm a-thrombin, 300 nm AT3 (■), 500 nm HirF (◇) or 300 nm AHT (△) was added at 0 s, and fluorescence of 6-FAM was continuously monitored [13]

Alternatively, changes in single-labeled aptamers or aptamer complexes can be monitored in real time with fluorescence anisotropy (Fig. 8b). In a fluorescence anisotropy assay, the fluorophore is excited by a polarized light, and linearly polarized components of emission are detected. This reveals information about the size, shape and flexibility of the fluorophore linked to the macromolecule. Compared to the target proteins, aptamers are relatively small; thus, upon binding the protein, significant changes in its molecular weight are reflected by significant changes in its rotational dif-

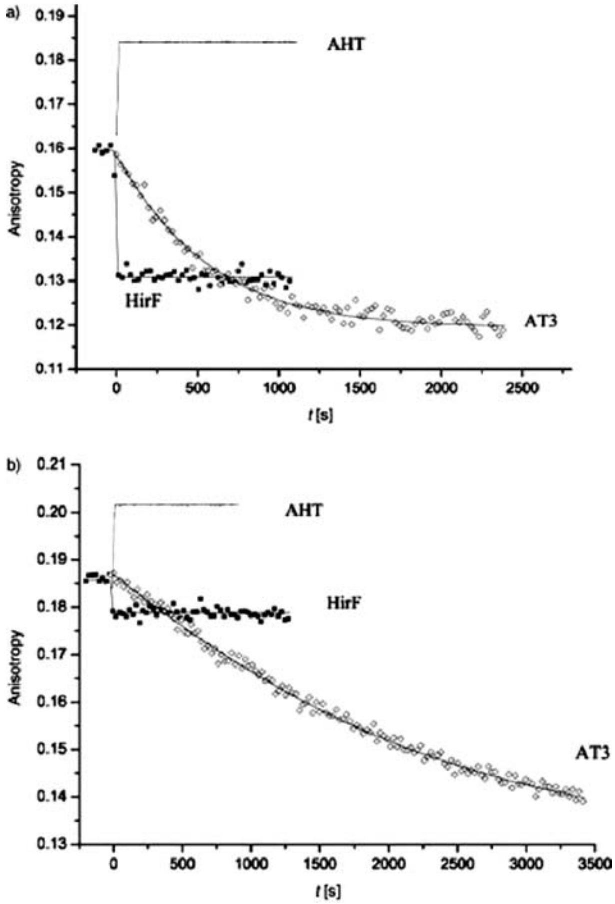


Fig. 10 TAMRA-labeled aptamers for a-thrombin/protein interactions based on fluorescence anisotropy. **a** In a solution of mixed 100 nm T-15Ap and 100 nm a-thrombin, 200 nm AT3 (◇), 500 nm HirF (■) or 300 nm AHT (△) was added at 0 s, and anisotropy of TAMRA was recorded in real time. **b** Same experiments as in (**a**) using the T-27Ap aptamer where 200 nm AT3 (◇), 500 nm HirF (■) or 300 nm AHT (△) was added to the aptamer/a-thrombin mixture solution at 0 s [13]

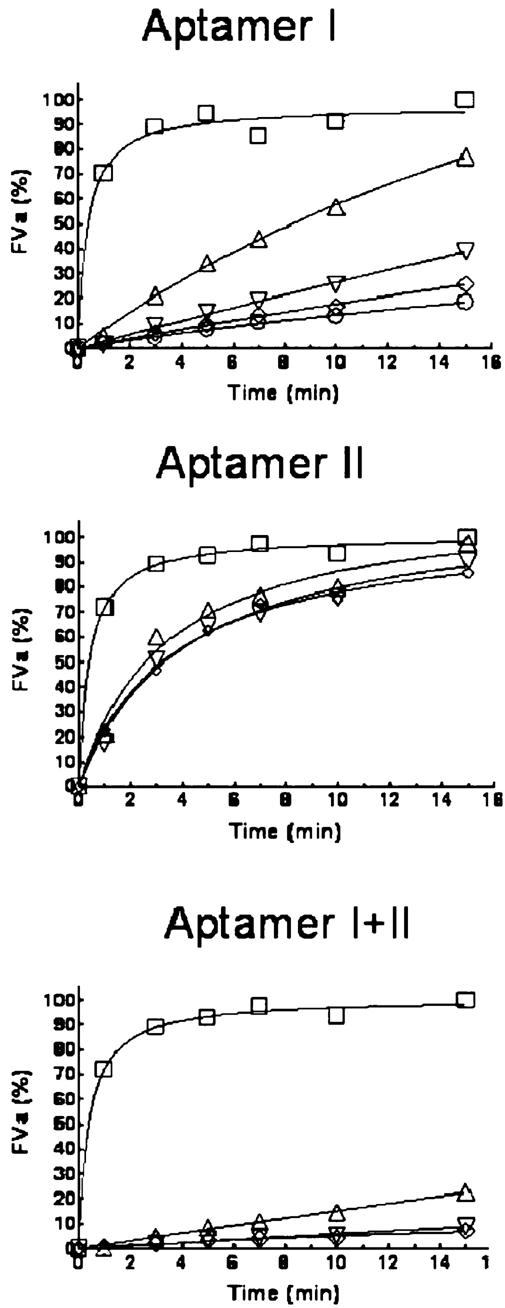
fusion rates. No conformational change is required before or after protein binding in a fluorescence anisotropy assay. This overcomes the handicap of complex instrumentation required to monitor and differentiate polarized excitation and emission. These assays are highly useful for aptamers without any knowledge of structure and conformation changes of the target [12, 13]. Fluorescence anisotropy addressed the details of displacement of aptamer in AT III-thrombin interaction, revealing a mechanism in which AT III slowly attacks thrombin, while at the same time forcing the aptamer to release from the binding site (Fig. 10).

Aptamers generated for multiple sites on the same protein were used to elucidate the role of thrombin exosites I and II in the activation of the blood coagulation cascade. Thrombin activates Factor V through proteolysis at Arg^{709}, Arg^{1018} and Arg^{1545}. Although both exosites are implicated in FV activation, their individual role in recognition of the cleavage sites was elucidated with the help of aptamers by monitoring the time-course of activation of FV cleavage site mutants when only one thrombin cleavage site is available [14]. These time courses were recorded in the presence of aptamer 1, aptamer 2 or both (Fig. 11). Cleavage at Arg^{709} was completely blocked in the presence of the exosite I aptamer and had a minor effect with the exosite II aptamer. Arg^{1018} cleavage also showed the same results. Arg^{1545} cleavage was significantly inhibited by both aptamers, meaning that both exosite 1 and exosite 2 are involved in recognition and cleavage completely inhibiting the activation of FV.

Similarly, aptamers generated for multiple sites on a single protein can be used to perturb the cellular network of proteins. This will be very useful in identifying inhibitors of a particular surface of cell-signaling proteins in cancer and cell-cycle regulation [15]. Two aptamers for different functional sites for a single TATA-binding protein, which is a general transcription factor and component of all three eukaryotic RNA polymerases, have been identified for yeast TATA binding protein. It was further shown that the use of two aptamers for different sites on the same protein inhibits RNA polymerase II-dependent transcription, but does so using two distinct mechanisms [15]. This principle can be further expanded to the study of protein–protein interaction on whole cells using a panel of aptamers, which recognize different surface markers, generated using cell-based SELEX [5].

Fig. 11 Effect of exosite aptamers I and II on activation of wild type FV by thrombin. Recombinant WT FV (250 pM) was activated with thrombin (10 nM) in the presence of varying concentrations of aptamer I (□ 0 μM; Δ 1 μM; ∇ 5 μM; ◇ 10 μM; E, 19 μM), aptamer II (□ 0 μM; Δ 1 μM; ∇ 5 μM; ◇ 9 μM), or both aptamers I and II (□ 0 + 0 μM; Δ 1 + 1 μM; *f* ∇; 5 + 5 μM; ◇ 5 + 8.2 μM). At the *time points* indicated, aliquots were taken from the activation mixture, and the FVa cofactor activity was measured and expressed as a percentage of the FVa cofactor activity of fully activated WT FV. The data were fit by nonlinear regression with the program GraphPad Prism [14]

▶


Aptamer I
FVa (%)
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
Time (min)
Aptamer II
FVa (%)
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
Time (min)
Aptamer I+II
FVa (%)
100
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
1
Time (min)

3.2 In vivo

Gap junction channel proteins are encoded by the innexin multigene family in *D. melanogaster*. The functional differences among various innexins are still elusive. The biological importance of heteromeric and homomeric channel formation is still unknown. Two aptamers for the cytoplasmic domain of anti-innexin 2 were identified. The selected aptamers interfere with the inter-

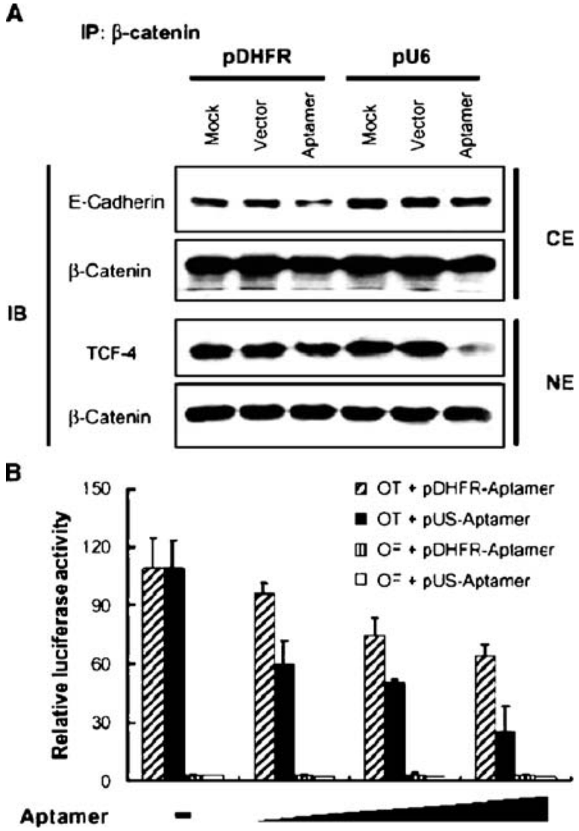


Fig. 12 Differential effects of the two RNA intramers on β-catenin–protein complexes. **A** Coimmunoprecipitation (Co-IP) assay of pDHFR- and pU6-derived RNA intramers. HCT116 cells were transfected with the indicated plasmids, and cytoplasmic and nuclear extracts were immunoprecipitated with anti-β-catenin antibody, followed by Western blotting with anti-β-catenin, anti-TCF-4, and anti-E-cadherin antibodies. **B** HCT116 cells were cotransfected with TCF-responsive wild-type (OT) and mutant (OF) luciferase reporters and an increasing amount of the pDHFR-aptamer or pU6-aptamer (0.2, 0.5, and 0.7 Ag). After overnight incubation, luciferase activities were measured. Five independent experiments were performed [18]

action of cytoplasmic tails of innexins 2 and 3 [16]. Biochemical fractionation and co-immunostaining revealed the heteromeric channel formation as crucial for morphogenesis and development in *D. melanogaster*. These in vitro results can also be translated in vivo, similar to RNAi studies, by using aptamers as inhibitors or targeting innexins 2 and 3 [16].

Another example of in–vivo protein–protein interaction studies with aptamers involves T-cell factor (TCF) family proteins which are DNA binding transcription factors that bind to a potent transcriptional activator β-catenin and also interact with transcriptional Groucho co-repressors. Disruption of this binding is a likely way of affecting gene expression. RNA aptamers were selected against TCF-1 [17]. The aptamer for TCF-1 inhibits the interaction between TCF-1 and β-catenin [18]. The aptamers were expressed in vivo under two different promoters that allowed them to be localized to nucleus (pU6) and cytoplasm (pDHFR). Real-time PCR was used to determine the relative subcellular distribution of the aptamers. The aptamers were then transfected in HCT116 cells, and the cells were fractioned in order to determine whether the formation of subcellular protein complexes was inhibited. In this case, the nucleus-expressed intramer reduced protein–protein interaction between β-catenin/TCF-4 protein complex (Fig. 12A), and the cytoplasm-expressed intramer decreased the amount of β-catenin/E-cadherin complex (Fig. 12B). It was also clear that the nucleus-expressed intramer was more effective in suppressing transcription.

4 Conclusion

In conclusion, the use of aptamers to elucidate protein–protein interaction is a growing field. With the proper design and engineering, aptamers can be applied to complex problems in vitro, such as use of anticoagulation cascades and in–vivo methods to study channel formations. Ultimately, as the number of aptamers increases for different targets, researchers will have aptamers as a standard tool to investigate protein–protein interactions. The study of systems biology will thereby be enriched, since protein–protein interactions based on aptamers will form an important part of the so-called "interactomics", which involves all the functionally relevant "interactions" in an organism.

References

1. Jones S, Thornton JM (1996) Proc Natl Acad Sci USA 93:13
2. Shoemaker BA, Panchenko AR (2007) PLoS Comput Biol 3:e42
3. Tuerk C, Gold L (1990) Science 249:505

4. Ellington AD, Szostak JD (1990) Nature 346:818
5. Shangguan D, Li Y, Tang Z, Zehui CC, Hui WC, Mallikaratchy P, Sefah K, Chaoyang JY, Tan W (2006) Proc Natl Acad Sci USA 103:11838
6. Green LS, Jellinek D, Jenison R, Stman A, Heldin CH, Janjic N (1996) Biochemistry 35:14413
7. Fang X, Sen A, Vicens M, Tan W (2003) ChemBioChem 4:829
8. Vicens M, Sen A, Vanderlaan A, Drake TJ, Tan W (2005) ChemBioChem 6:900
9. Yang JC, Jockusch S, Vicens M, Turro NJ, Tan W (2005) Proc Natl Acad Sci USA 102:17278
10. Bock LC, Griffin LC, Latham JA, Vermass EH, Toole JJ (1992) Nature 355:564
11. Huang CC, Zehui C, Huang-Tsung C, Tan W (2004) Anal Chem 76:6973
12. Fang X, Cao Z, Beck T, Tan W (2001) Anal Chem 73:5752
13. Cao Z, Tan W (2005) Chem Eur J 11:4502
14. Segers K, Dahlback B, Bock P, Tans G, Rosing J, Nicolaes AFG (2007) J Biol Chem 282:33915
15. Shi H, Fan X, Sevilimedu A, Lis JT (2007) Proc Natl Acad Sci USA 104:3742
16. Knieps M, Herrmann S, Lehmann C, Loer B, Hoch M, Famulok M (2007) Biol Chem 388:561
17. Lee SK, Park MW, Yang EG, Yu J, Jeong S (2005) Biochem Biophys Res Commun 327:294
18. Lee KH, Kwak HY, Hur J, Kim IA, Yang JS, Park MW, Yu J, Jeong S (2007) Cancer Res 67:9315

Adv Biochem Engin/Biotechnol (2008) 110: 195–214
DOI 10.1007/10_2007_090

Published online: 25 January 2008

Investigating Protein–Protein Interactions by Far-Westerns

Catherine S. Chan · Tara M. L. Winstone · Raymond J. Turner (✉)

Department of Biological Sciences, University of Calgary, 2500 University Dr. NW,
Calgary, AB T2N 1N4, Canada
turnerr@ucalgary.ca

Abstract The identification of protein interaction partners can often elucidate the function of the protein under investigation based on the "guilty by association" concept. Furthermore, the binding event between two proteins can be used as a functional assay when no such assay is available. Despite the large number of advanced techniques that are currently available for studying protein–protein interactions, far-Westerns or blot overlays are still very commonly used in the average laboratory setting due to their powerfulness. This is due to the simplicity and clarity in the results that they produce. Here, the details and mechanics of far-Westerns are discussed to help the reader choose amongst the different variations that exist depending on the question being investigated and the materials available to them. Some examples involving unique questions are also discussed in order to educate the reader on the versatility of far-Westerns. Finally, a troubleshooting section provides the reader with an understanding of the common problems that can be encountered and how these problems can be circumvented.

Keywords Blot overlay · Far-Western · Protein–protein interactions · Sandwich ELISA · Sandwich Western

Abbreviations

AP	Alkaline phosphatase
BSA	Bovine serum albumin
DMSO	Dimethyl sulfoxide
DTT	Dithiothreitol
ELISA	Enzyme-linked immunosorbent assay
GdnHCl	Guanidium hydrochloride
GST	Glutathione S-transferase
HRP	Horseradish peroxidase
PVDF	Poly(vinylidene fluoride)
SBP	Streptavidin-binding peptide
SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis

1 Introduction

Far-Westerns or blot overlays are becoming a popular in vitro approach in studying interactions. They are low cost and have low maintenance requirements of materials and equipment. The protocols are relatively easy, with a multitude of parameters that may be optimized and tailored to a variety of investigations. The principle of far-Westerns is similar to Western transfer but is probed with another protein or protein extracts prior to incubation with an antibody. They were initially used to identify interactions with the Myc protooncogene protein in the early 1990s [1, 2].

In general, the proteins of interest (prey) are separated by gel electrophoresis and then transferred to a solid membrane support. Unlike Western blots that are incubated with an antibody towards a target protein, the membrane is incubated with a second protein of interest (probe or bait proteins) in a far-Western. An interaction between the two proteins causes the immobilization of the probe (Fig. 1a). The probing protein is then detected using an antibody against it or a fusion tag that was genetically fused to it. Far-Westerns are also considered to be a modified version of sandwich ELISAs where an immobilized antibody towards one epitope of the protein is used to capture the protein on a membrane, followed by detection of the protein using an alternative antibody against a different epitope on the protein (Fig. 1b).

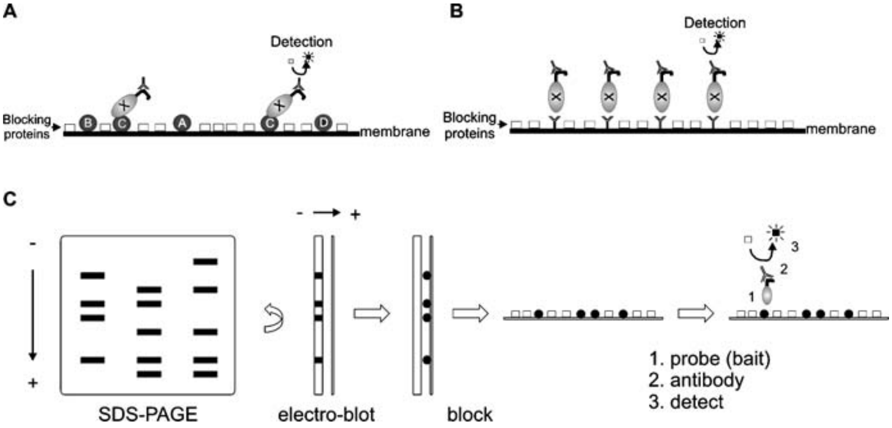


Fig. 1 Far-Westerns for detecting protein–protein interactions. **a** A simple far-Western for investigating an interaction between two proteins (C and X). In this scenario protein X contains a fusion tag which a commercially available antibody can detect. Edited and reproduced with permission from Chan et al. [21]. **b** A sandwich ELISA showing how two different antibodies towards a protein of interest can be used to capture and detect that protein from a mixture of proteins. **c** A schematic of a "typical" far-Western blot where proteins are first separated by SDS-PAGE based mostly on their size, followed by transfer of the proteins onto a solid membrane via electroblotting. During this process SDS is removed and the proteins refold on the surface of the membrane. The membrane is then incubated in a blocking solution to block any unbound sites, followed by incubation with the probe protein. The interaction is detected by incubation with an antibody against the probe protein and enzymatic detection of the antibody conjugated to AP or HRP

While far-Westerns are relatively simple to perform, there are various modifications that can be carried out depending on the available material and the question at hand. The following sections will cover the basic procedures of a typical far-Western and then discuss the requirements when choosing between the modifications.

1.1 General Theory and Procedure

Since detailed protocols vary for each individual situation and largely depend on the available material and supplies, the focus here will be on the general theory and procedure behind a typical far-Western blot. A detailed protocol has been reviewed in an article by Hall [3]. Various applications which consist of slight modifications to the described procedure below are discussed with specific examples in later sections.

A "typical" far-Western blot starts with electrophoretic separation of the protein(s) of interest, also referred to as prey proteins, by SDS-PAGE for denaturing conditions (Fig. 1c). The proteins are first treated by incubation in

Laemmli solubilization buffer [4], which reduces the disulfide bonds by DTT and linearizes the protein while adding a uniform negative charge from SDS detergent. High-temperature (>95 °C) incubation aids further linearization and penetration of SDS molecules. The proteins are then loaded onto precast polyacrylamide gels and electric current causes the proteins to separate based primarily on size (Fig. 1c). Once separated, the proteins are transferred to a solid membrane support made of nitrocellulose or PVDF, usually by electroblotting. The membrane is then incubated in a blocking solution containing a mixture of proteins known to not interact or interfere (e.g., skim milk, BSA, or gelatin) in order to "block" the unoccupied sites on the membrane. The blot is then incubated with the probing protein of interest (Fig. 1c, step 1). In this case the probe contains a fusion tag to which an antibody is available commercially (Fig. 1c, step 2). On the other hand, specific polyclonal antibodies towards the protein itself can also be generated.

The antibody will be conjugated to an enzyme, typically horseradish peroxidase (HRP) or alkaline phosphatase (AP), to which an added cleavable substrate will generate a detectable product (Fig. 1c, step 3). The method works well with the common detection procedures of chemiluminescence or colorimetry.

1.2 Questions That Can Be Investigated by Far-Westerns

There are a large number of questions that can be explored by far-Westerns to investigate protein–protein interactions. Specific examples of every question will not be covered here; however, questions that can be answered by far-Westerns are, but not limited to:

- Protein–protein interactions
  - Native
  - Denatured
  - Unfolding and stability
  - Dependence on DNA, lipid, or carbohydrate binding
- Interaction domain fingerprinting
- Differential protein interactions from phosphorylation events
- Interactome fingerprinting

1.3 Considerations Before the Experiment

As there are many variations of far-Westerns, the goal of this section is to provide the reader with help to choose the appropriate type of experiment based on the required sample material for that particular experiment. It also allows the reader to determine what steps must be taken prior to their experiment.

1.3.1 Sample Considerations—When Only One Protein Is Available at High Purity

To avoid ambiguity in the results, at least one of the proteins being investigated should be highly (>90%) pure. When only one protein is available at such purity, this should be the one used for probing (the bait), i.e., the one that the blot is incubated with (Fig. 1a, protein X). The other protein(s) of interest (the prey) are those separated by gel electrophoresis and thus need not be purified. This is because the proteins are separated based on molecular mass when separating via SDS-PAGE, and can be tentatively identified by comparing the theoretical mass to protein standards of known masses that are included in the gel. Given that SDS-PAGE gels can vary in the percentage of acrylamide to bisacrylamide, an appropriate percentage should be chosen to best resolve the molecular mass of the target protein(s) (Fig. 1a, proteins A–D). Since the possibility of multiple proteins of similar sizes also exists, especially when working with samples containing large amounts of other proteins such as cell extracts, a separate Western blot control probing for this protein should also be conducted alongside the far-Western blot.

The data presented in Fig. 2 demonstrate the importance of proper controls. The two proteins being investigated here are NarJ, the chaperone for the *Escherichia coli* nitrate reductase A, and the N-terminal fragment of NarG, the catalytic subunit of the reductase. In this example, NarJ was cloned to contain a His_6 tag and T_7 epitope, where the His_6 tag allowed for large-scale purification using Ni^{2+} affinity chromatography and the T_7 epitope allowed for detection during the far-Western using a HRP-conjugated anti-T_7 antibody. The NarG peptide was cloned with a streptavidin-binding peptide (SBP) at its C terminus and allowed for detection during the Western using HRP-conjugated streptavidin. NarG:SBP was not purified and was loaded directly onto the SDS-PAGE gel from a cell extract; lanes 2 and 3 show the Coomassie-stained gel of the extracts, with lane 3 containing extracts where the expression of NarG:SBP was induced off the promoter and lane 2 was not. Comparing the two lanes identifies that NarG:SBP was indeed overexpressed based on the new strong band around ∼19 kDa. The Western blot using HRP-conjugated streptavidin shows two bands, one corresponding to NarG:SBP and one nonspecific protein that appears to bind streptavidin (Fig. 2, lane 4). The far-Western probing with His_6-T_7:NarJ shows that NarJ binds NarG but there is also a faint band at ∼35 kDa (lane 5). Since this band did not show up in the Western but did show up in the empty vector control lane 6, it can be ruled out that NarJ is interacting with a nonspecific protein in the cell extract due to the presence of the vector, as lane 7 containing cells without any plasmid did not contain this band. Although the results shown in Fig. 2 are more extensive than is required for controls in this situation, it demonstrates their usefulness in a case where the protein is not visible as being overexpressed on a SDS-PAGE gel, when the protein of interest does not mi-

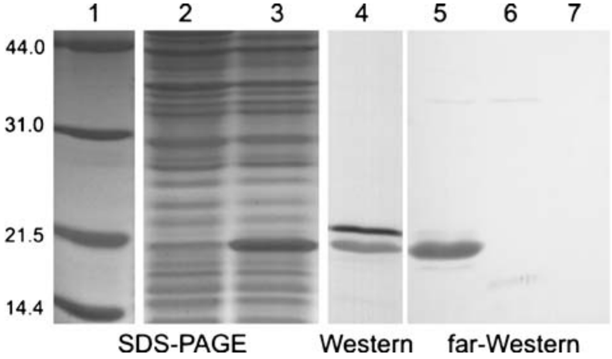


Fig. 2 Critical controls for far-Western blots when using impure samples. Samples were separated on a 12% SDS-PAGE gel and then either stained with Coomassie Brilliant Blue, transferred to nitrocellulose, and detected with streptavidin-HRP (Pierce) in the Western or incubated with 100 μg/ml NarJ and then 1 : 5000 T_7 · tag-HRP conjugate (Novagen) in the far-Western. Lane *1*, low range molecular weight markers indicated in kDa from BioRad; lane *2*, cell extracts of cells where the expression of NarG:SBP was not induced; lanes *3–5*, cell extracts of cells that were induced for the expression of NarG : SBP; lane *6*, cell extracts from cells carrying the empty vector; lane *7*, cell extracts from cells not carrying any vector or plasmid

grate at the expected size, or there is more than one band showing up on the far-Western.

The use of cell lysate containing the probing protein for incubation can be done, but higher amounts should be used to ensure that an adequate amount of probing protein is present for detecting the interaction. For this experiment it is more critical that the probing protein (the bait) contains a specific tag or epitope for recognition, to avoid ambiguity in the results due to binding of other cellular proteins.

1.3.2
Sample Considerations—When Both Proteins Are Available at High Purity

When both proteins are available in a purified form, it is possible to conduct one of the two types of nondenaturing far-Western experiments, as opposed to the denaturing far-Westerns described in Sect. 1.3.1. The two options are native far-Western or dot-blot far-Western. These two techniques are also useful when the proteins being investigated require a proper conformation or fold in the protein to interact, where the denaturing far-Western blot method described above may not be able to detect such interactions. Although it has been accepted that proteins refold on the membrane upon transfer due to the removal of the SDS, the amount of protein that refolds can vary and there is also no guarantee that all proteins refold completely and correctly. However, it should be stressed that in some cases as long as the interaction domain is refolded, then denaturing far-Western blots are a suitable option [5].

A native far-Western blot contains proteins that were first separated on a nondenaturing or native gel. Native gels are similar to SDS-PAGE gels but lack SDS and/or DTT and the samples are not heated prior to loading onto the gel. The mobility of the proteins is based on their overall net charge at the pH of the running buffer and hydrodynamic size. This method allows for investigation of interactions while maintaining the native (not refolded) form of both proteins. An example demonstrating the usefulness of this type of far-Western blot was shown in a study by Winstone et al. [6] where the N-terminal portion of the DMSO reductase catalytic subunit DmsA was shown to interact with all three folding forms of its chaperone DmsD. Prior studies showed that DmsD was found as a monomeric A form, dimeric B form, and a pH 5-induced ladder D form. Since forms B and D were converted to the A form when they were separated on a SDS-PAGE gel but not on a native gel [7], a native far-Western was the only way that the authors could determine which forms of DmsD bound the N-terminal peptide of DmsA [6].

A dot-blot far-Western contains proteins or cell extracts that are spotted directly onto the membrane without any prior electrophoretic separation. The advantage of this method is that it is quicker given that an average SDS-PAGE gel takes 1–2 h to set up and complete and 2–4 h for a native gel using mini-gels ~8–10 cm in length and ~0.75–1.5 cm in thickness. The typical protocol involves spotting the protein sample onto the dry membrane and then allowing it to dry thoroughly, followed by a wash step with buffer to wash off unbound sample prior to blocking [7]. Although one can spot manually, uniform spots can be produced by the use of equipment similar to a microfiltration device, such as the Bio-Dot® produced by BioRad or Multiscreen HTS Filter Plates by Millipore, where the membrane is sandwiched between two filter plates. As the top plate consists of an 8×12 sample chamber, they also allow for high-throughput screening of up to 96 interactions at once. Since these devices form a tight seal on the membrane that prevents excessive spreading of the protein spots, a prewetted membrane is used and the drying step is eliminated. This allows for the maintenance of aqueous conditions throughout the experiment, thus preventing any negative effects that drying of the protein may cause. Inclusion of suitable negative controls is an absolute must when using cell extracts or impure samples during a dot-blot far-Western to rule out nonspecific interactions.

An example using a dot-blot far-Western coupled with the Bio-Dot® is shown in Fig. 3. In this study, a bank of DmsD single-site mutants was screened for binding towards the DmsA N-terminal peptide and a sample titration of the W72S and C64S mutants compared to wild-type DmsD is shown. Wild-type and mutant DmsD samples were applied in serial dilutions across 12 rows of the device and allowed to bind by gravitational filtration. Following two washes by addition of buffer that was also allowed to flow through by grav-

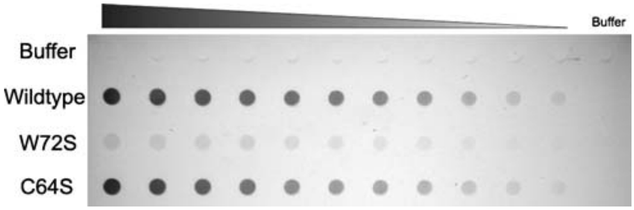


Fig. 3 Dot-blot far-Western of DmsD and DmsD mutants. DmsD protein was spotted onto nitrocellulose membrane starting from 0.25 mg/ml at 1.5-fold serial dilutions. The far-Western was then completed by incubation in 25 μg/ml DmsA : GST, followed by 1 : 5000 GST · tag monoclonal antibody (Novagen) and 1 : 3000 goat anti-mouse HRP conjugate (BioRad). Negative controls containing only buffer are included in the first row and last column

ity, the assembly was taken apart and the entire membrane was blocked in buffer containing skim milk. The far-Western was completed by incubation with DmsA peptide that was obtained by purification using a GST affinity column against a GST tag that was fused with it by cloning, and then incubation with an antibody against GST. Using the quantitative software that comes with most image capturing systems (Kodak 1D Image Analysis in this case), the pixel intensities of each spot were obtained and compared to that for wild type. In the case of W72S, the relative binding was ∼8% that of wild type and C64S was ∼95% (T. Winstone 2007, unpublished results). The results from the far-Western proved useful in quantitating the effects of the mutations in DmsD, and the use of the Bio-Dot® device allowed for high-throughput screening of a large number of mutants in repetition at the same time. Since a T_7 tag is also on each of the DmsD proteins, it was possible to ensure that binding to the membrane was equivalent for all mutants by performing a Western blot on the same titrations of DmsD proteins. This confirms that the relative intensity was indeed from the interaction with the DmsA N-terminal peptide and not due to differential amounts of each DmsD mutant.

1.4 Considerations During the Experiment

As far-Westerns are very similar to Westerns, the conditions to consider during a far-Western are discussed briefly in the following sections with major focus on the conditions that have a more direct effect on a far-Western experiment.

1.4.1 Transfer Conditions

The principles and conditions behind transferring of proteins are well covered in manuals describing Westerns [8–10]. Rather than going through the details of transfer, the focus here will be on the conditions that have effects on the subsequent detection of the interaction during a far-Western. The common method of transfer is by electroblotting. The length of transfer time is a crucial variable as large proteins require a longer time to fully transfer. This is a similar concern when transferring through thicker gels (1.5 vs 0.75 cm). However, smaller proteins and peptides may pass through the membrane to the anterior side if transferred too long; therefore, it is important to select the conditions appropriate for the proteins under investigation. The type of membrane is also important as some proteins bind specifically or selectively to one type over another [11]. While nitrocellulose is the cheaper alternative, PVDF has a higher binding capacity (>125 vs 75–90 $\mu g/cm^2$ [12]), and is less susceptible to damage during handling. PVDF may be the better choice if possible for detecting weaker interactions as the higher binding capacity allows more protein to bind and be detected.

Despite the popularity of nitrocellulose membranes, the mechanism by which proteins bind to them is still relatively unknown. Two hypothetical models have been postulated and both involve a combination of hydrophobic and electrostatic interactions, with some hydrogen bonding [13, 14]. The only difference between the two models argues that the initial attraction is via electrostatic versus hydrophobic interactions and is stabilized by the other two remaining interactions in both cases. Since evidence supporting both models exists [13], it is likely that the nature in which proteins interact with the membrane is protein-specific and should be kept in mind when choosing the appropriate type of membrane for a given experiment.

1.4.2 Blot Incubation Conditions

Regardless of the type of far-Western being performed, the blot must be blocked in a noninteracting protein solution to block the unbound sites on the membrane. As far-Westerns contain an extra incubation step with the probing protein, the choice of blocking solution is even more important to ensure that proteins in the solution do not cross-react with the probing protein. Further considerations also ensure that the proteins in the blocking solution do not interact nonspecifically with the antibody/antibodies in the next step.

The stability of the primary protein on the blot should also be considered during blocking; this may mean that the blot needs to be blocked with a higher concentration of blocking solution for a shorter period of time. In the aforementioned example of screening DmsD mutants binding to the DmsA peptide (Sect. 1.3.2 and Fig. 3), it was noticed that two mutants exhibited lower levels of DmsA peptide binding when the blots were blocked overnight compared to 1 h at room temperature (T. Winstone 2007, unpublished results). As none of the other mutants displayed this type of difference based on the length of blocking time, it is likely that the stability of these mutants was affected more significantly by the long blocking time and possibly the temperature. Although blocking at 4 °C was not tested in this case, this temperature could be used in order to maintain the stability of the proteins when blocking for longer periods of time.

The next step involves incubation with the probing protein and the key is to recognize that the interaction is described by the simple ligand-binding model in Eq. 1 using the schematics in Fig. 1a:

$$[\mathrm{C}] + [\mathrm{X}] \leftrightarrow [\mathrm{CX}]\,. \tag{1}$$

The interacting form of CX is obtained when satisfactory conditions are met to drive the equilibrium towards that form. The "tightness" of the interaction is described by the dissociation constant (K_D) and is obtained by Eq. 2:

$$K_\mathrm{D} = [\mathrm{C}] \cdot [\mathrm{X}]/[\mathrm{CX}]\,. \tag{2}$$

In general, a smaller value of K_D indicates a tighter interaction. For example, antibody–antigen interactions are in the nanomolar range whereas enzyme–substrate interactions are typically in the micromolar to millimolar range. A summary of the factors important in favoring the interaction between proteins C and X is given in Table 1.

More complex examples include interactions that are dependent on a third protein (*m* for example), where the third protein is required for the complex formation of C-*m*-X bringing together proteins C and X via an indirect interaction. The details of this complex ligand-binding example will not be discussed any further here, but careful manipulation of all three protein concentrations is typically necessary and by doing so one learns of the nature of the C-*m*-X complex.

The stability of the probing protein during its incubation should also be considered, but it should be noted that the binding kinetics at lower temperatures would be slower so longer incubation times may be required for the interaction to occur. This time should be empirically determined by pilot range-finding experiments. For the previous examples of DmsD and NarJ shown in Figs. 2 and 3, the probing protein was incubated at room temperature for 2 or 1 h, respectively, and this appeared to be the optimal length of time for these proteins.

The type and composition of buffer that the probing protein is incubated in is also important and should be determined prior to incubation. Although the probing protein may be stable in certain buffers at extreme conditions (pH, ionic strength, etc.), whether the proteins on the blot are stable to such buffers should also be kept in mind. Additives to the buffers are also useful in reducing nonspecific interactions, but may also interfere with interactions between the proteins being studied. Common additives can include a small amount of the blocking protein to drive a higher specificity of the interaction through competition and mild detergents, such as Triton X-100 or Tween-20 (or -80), to reduce background interactions.

Table 1 Conditions to consider when optimizing far-Western blots

Parameter	Pros and cons	Recommended conditions
Probe protein concentration	Higher protein concentration drives the interaction equilibrium to favor the interaction but too high forces nonspecific interactions and may contribute to background noise	Empirically determined—optimize total volume and total protein for the size of the blot
Temperature	Stabilizes proteins but affects the kinetics of the interaction	Room temperature if proteins are stable or else at 4 °C
Length of incubation	Provides time for the proteins to come into contact and the interaction to occur. Too lengthy an incubation can result in protein degradation and/or cause nonspecific interactions that contribute to background noise	1–3 hours at room temperature and approx. double at 4 °C
pH	Maintains stability of protein and keeps surface charge "native" in order to maintain physiological conditions for the interaction	Empirically determined but near 7.0 for most cases. Type of buffer is specific to each protein
Ionic strength	Maintains near physiological conditions for the proteins. Also helps reduce nonspecific interactions due to electrostatic interactions	~100–200 mM NaCl
Additives	Chaotropic agents can disrupt weak protein–protein interactions and be used to assess the affinity of the interaction	Empirically determined
	Detergents at low levels reduce background interactions	~0.02–0.05%
	Small amount of blocking protein drives the interaction specificity and reduces background interactions	~10-fold less than the concentration used for blocking

1.4.3 Detection Considerations

The interaction between the proteins of interest can be through a variety of methods and the choice depends on a few factors: (1) whether the method interferes with the C–X interaction site, (2) whether the chemistry affects the stability of the C–X interaction, (3) sensitivity of detection, (4) availability of materials, and (5) cost. The first three factors should be ranked higher in importance as they directly affect the ability to detect the interaction at all.

An increasingly useful detection approach involves using an antibody against a fusion tag that is on the protein of interest, such as the T7-epitope on the probe proteins in Figs. 2 and 5 or the GST tag in Fig. 3. This method is advantageous as antibodies against these fusion tags are commercially available in a higher specificity monoclonal form at a relatively low cost. Some are even synthesized to be conjugated with the enzymes AP or HRP, such as the T7-tag HRP conjugate (Novagen) used in Fig. 2, eliminating the need to probe with a secondary antibody.

The other type of detection involves using custom polyclonal antibodies against the probing protein, typically raised in mice or rabbits. These are useful in situations where fusion tags affect the function of the protein or interfere with the interaction site regardless of which termini they are placed at. As polyclonal antibodies recognize multiple epitopes on the protein, problems associated with occlusion of the epitope due to display and interaction is less likely to occur. However, the cost associated with generating these antibodies is much higher and is not recommended in large-scale studies, such as those presented in Fig. 5, which will be discussed later. As a secondary antibody (anti-mouse or anti-rabbit) is required to detect these antibodies, this adds more time and cost.

With either method in which an antibody is used to detect the probing protein, modification of the antibody concentrations (from what is recommended for a Western blot) may be necessary. Typically a higher concentration of antibody has been used for far-Western detection than in Western detection when comparing equivalent antibodies. This is likely because the amount of prey protein on the membrane binds less of the probe/bait protein which is then detected by the antibody.

The other methods involve using a probing protein that is radioactively labeled or biotinylated. Detecting radioactively labeled proteins is the most direct method and requires no subsequent incubations, but is more hazardous and is also subject to interference of interaction sites by the labels. The direct detection of the probe protein often leads researchers to argue that this is actually a Western blot. Detection of biotinylated proteins relies on AP or HRP conjugated streptavidin and requires that a biotinylation signal is genetically fused to the gene of the protein of interest.

Other than using radioactively labeled proteins, all other methods require the addition of a substrate for the enzymatic detection of AP or HRP. The choice of substrate is sensitivity-based with chemiluminescence being more sensitive than colorimetry for detecting weaker interactions. Various modified kits of chemiluminescence and colorimetry are also available for enhancing the detection of weaker interactions, such as the SuperSignal chemiluminescent substrate by Pierce or the Amplified Opti-4CN substrate for colorimetric detection by BioRad.

2 Specific Application Examples

The previous section included several examples of far-Westerns in studying protein–protein interactions. While far-Westerns are great for proving interactions, they can also be useful in determining the physiochemical circumstances in which the proteins interact. They are also useful in fishing out unknown interactions through proteomics-like experiments. This next section will go through some examples using far-Westerns to study more complicated questions.

2.1 Probing Specific Interactions

The following sections target specific interactions between two or more proteins; the prerequisite is that the target protein (prey) identities are known prior to experimentation.

2.1.1 Determining Whether Proteins Bind in Denaturing Conditions

Studying protein–protein interactions in the presence of denaturants can be useful in determining the types of forces that are involved in the interaction between two proteins. In a study by Winstone et al. [6], the binding of DmsD towards the DmsA N-terminal peptide was investigated in the presence of the denaturants urea and GdnHCl using a dot-blot far-Western. In this approach pure DmsA:GST (DmsA N terminus fused to GST) protein was spotted onto a nitrocellulose membrane and then incubated in solutions containing purified DmsD in differing concentrations of denaturant. The level of DmsD binding was quantitated by imaging software and the relative binding was obtained by comparison to DmsD binding in the absence of denaturant. These titrations showed that DmsD can bind DmsA at 100% levels up to 2 M urea, whereas it only binds at 100% up to 0.5 M GdnHCl (Fig. 4). By comparing the far-Western data to fluorescence data that were also collected by titra-

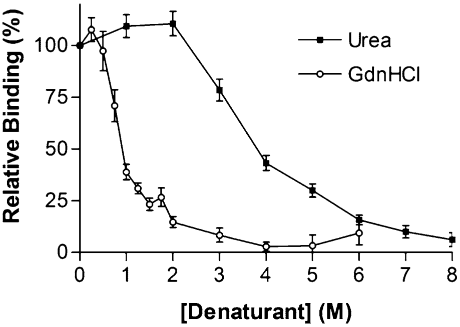


Fig. 4 Binding of DmsD in urea or GdnHCl towards DmsA N-terminal peptide. Relative intensities from far-Western blots described in Winstone et al. [6] are shown to demonstrate how denaturants affect the binding of the two. Modified from Winstone et al.

tions of denaturants, the authors were able to correlate the unfolding of DmsD with the relative binding towards DmsA. Similar experiments in GdnHCl also showed correlation between the conformation changes of DmsD and its relative binding, suggesting that the binding involves electrostatic interactions between DmsD and the DmsA peptide [6].

2.1.2 Large-Scale Probing of Simple Interactions

Far-Westerns are also useful in studying the interaction specificity of a family of related homologous proteins. Using this approach, the researcher can determine whether a protein of interest binds proteins of similar functions and possibly separate the proteins into specific families based on their specificities. An example using far-Western blots to investigate the binding of a family of proteins is demonstrated in Fig. 5. In order to investigate the binding of these proteins towards a set of substrate proteins, the substrate proteins (A to I) were cloned with a SBP and overexpressed. The cell extracts were separated via SDS-PAGE in order to eliminate the need to purify all nine proteins. To rule out false negative results due to problems with protein expression, the amount of each protein was approximately equalized on the gel by Coomassie staining prior to the far-Western (Fig. 5, top panel). The probing proteins X, Y, and Z were cloned with a His_6-T_7 tag similar to that used in Fig. 2 and their purified forms were used to probe for their binding to proteins A to I on separate blots (Fig. 5, bottom three panels). The data shown here demonstrate that protein X binds proteins B, G, and H, which are all homologous proteins. Sim-

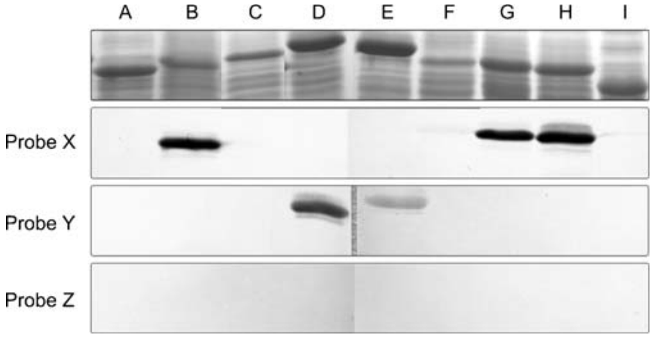


Fig. 5 Far-Western blots investigating binding partners of protein families. Substrate proteins (*A* to *I*) were cloned with a C-terminal SBP tag and then overexpressed. Cell extracts were separated via 15% SDS-PAGE and the amount of each substrate was approximately equalized based on Coomassie staining (*top panel*). Subsequent far-Western blots (*bottom three panels*) were done using purified forms of probe proteins X, Y, and Z at 100 µg/ml for 2 h at room temperature. The probes were then detected against their T_7 epitopes as described in Fig. 2

ilarly protein Y binds proteins D and E, also homologous proteins. Protein Z did not appear to bind any of the nine proteins.

Although the example described here is relatively efficient at probing large-scale interactions, the previously described dot-blot approach combined with the 96-well apparatus is more efficient. However, the example here was able to screen a large number (10 probes against 17 substrates in total; data not shown) of interactions without purifying any of the 17 substrates, and is therefore still very efficient for screening a large number of interactions at the same time.

2.1.3 In-Gel Far-Westerns

In-gel detection of antibody interactions can be done by pretreating the electrophoresed gels with 50% isopropanol and then distilled water, followed by detection using a high-sensitivity SuperSignal chemiluminescent substrate (Pierce) [15]. The crucial step of dehydration and rehydration removes the SDS and allows proteins to refold in the gel and bind the antibody but requires a high-sensitivity substrate for detection. Since then, Pierce has developed two ProFound far-Western kits optimized for in-gel far-Western interactions based on the streptavidin–biotin and GST–anti-GST interactions, which are attractive options as they eliminate the transfer process altogether. Although these kits have been proven in principle by Pierce, it does not appear that they have been used for studying protein–protein interactions yet. Furthermore, it is likely that the interaction of the proteins will be limited to the ability of the probing protein to diffuse into the gel and interact with its target. Since the proteins will need to diffuse into the gel for the interaction rather than on the surface of a membrane, larger amounts of protein will also be required, which may not be feasible for proteins of low abundance and accumulation or those that are difficult to purify.

2.2 Probing Blind Interactions—The Proteomic Approach

Far-Westerns are also useful in identifying previously unknown interacting partners. Since "interactomics", whereby all the prey interactions of a given bait protein are studied, is gaining popularity as a question worth investigating, far-Westerns have proven useful in finding the protein partners or at least showing that such an interactome exists. For this approach, a SDS-PAGE separation step would be required to separate the sample containing all the cellular proteins prior to transfer. Figure 6 is an example of a far-Western fingerprinting experiment where membrane fractions from wild-type *E. coli* or a mutant lacking the entire *tat* operon (Δ*tat*) were probed by the NarJ chaperone of nitrate reductase A [16]. From this far-Western blot,

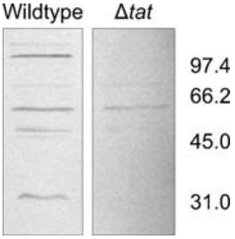


Fig. 6 Interactome fingerprinting of NarJ. Membrane fractions from wild-type or mutant (Δ*tat*) *E. coli* were separated on a 12% SDS-PAGE gel then transferred to a nitrocellulose membrane. The membrane was blocked overnight in 10% skim milk and then incubated with 40 μg/ml purified His_6-T_7 : NarJ for 2.5 h at room temperature. The binding of NarJ was then detected as described in Fig. 2. Modified from Chan et al. [16]

it is evident that the interactome of NarJ is different in the mutant compared to wild type.

Other examples of interactomics investigations can also include separation of cell extracts using two-dimensional PAGE prior to blotting, whereby the proteins are separated first based on their net charge followed by separation based on size. This type of separation further eliminates uncertainties due to proteins of similar size that can occur during SDS-PAGE. An example of a whole-genome interactomics study utilized a cDNA library to express proteins on a high-density filter membrane as a protein filter array [17]. The membranes were produced in multiples and then screened against various probe proteins of interest and when a positive interaction was observed, the identity of the protein was revealed based on its cDNA sequence, which was used to produce the array.

By combining far-Western fingerprinting experiments with peptide mapping, the identities of the differently interacting proteins can also be obtained. Using this approach, Reddy and Kumar were able to identify two *Mycobacterium avium* complex proteins involved in interacting with cell surface proteins of human epithelial cells [11]. This example is unique because the probe in this experiment was a total mixture of human epithelial cell surface proteins that were biotinylated. While this example provided clean results, it should be cautioned that when using a multi-protein mixture as a probe for far-Westerns of any variety, there is a possibility of producing ambiguous results due to the multitude of interacting combinations that can occur.

Access to some sort of protein identification instrument will also be useful for identifying the interacting proteins, unless only the fingerprint of interactions is being analyzed. In this case, a duplicate gel would be run alongside that is stained and then once the far-Western blot has identified binding partners, the corresponding band would be excised from the gel and then subjected to N-terminal sequencing or identification by mass spectrometry.

3 Problems and Caveats of the Far-Western Method

Previous sections have demonstrated the successfulness of using far-Westerns for studying protein–protein interactions. However, it should be noted that this technique is not perfect in every scenario and is subject to various technical problems, as expected with any experiment.

3.1 Problems Due to the Display of the Protein

Assuming that the protein is properly displayed on the membrane regardless of the fusion, whenever a protein is recombinantly fused to an artificial tag there is always the possibility that the tag may interfere with the protein's function or binding domain(s). A tag that has detrimental effects on the function of a protein is considered useless to many. Even if the addition of a tag has no apparent effect on protein function, it may still affect the protein's native fold and has the potential of occluding interaction sites. For these reasons, the type and location of the tag is a crucial factor when studying protein–protein interactions by any means. Many recent reviews are available for comparing the advantages and disadvantages of various fusion tags [18, 19]. Although fusion tags can be cleaved by addition of specific cleavage recognition sequences, proteolytic cleavage can often occur in other site(s) of the protein [20] or be incomplete. The following sections will discuss some examples where these two factors were shown to affect the outcome of interaction studies using far-Westerns.

3.1.1 Terminal Display Problems

Almost all fusion tags can be placed on the N or C terminus of the protein of interest [19]. Placing fusions within the protein sequence as an insertion is not usually recommended as this can severely affect the fold of the protein. However, insertion tags are useful in some cases where the domains are clearly defined and separated by unstructured regions or in loops that connect transmembrane helices of integral membrane proteins. The choice of which terminus is suitable for a fusion tag is specific to each protein and is best determined by testing both positions if possible. The yields of the protein may differ drastically when the tag is placed at one terminus versus the other [18], but this form of the protein may not be active. This was the case for the DmsA N-terminal peptide used in Figs. 3 and 4, where placing the GST fusion at the N terminus (GST:DmsA) produced far more protein than at the C terminus (DmsA : GST) yet the GST : DmsA chimera was unable to bind its chaperone DmsD [7]. Therefore, when it comes to choosing the appropriate terminus for a fusion tag, maintaining the function of the protein should always be a priority.

3.1.2 Problems Due to the Type of Fusion Tag

As discussed previously, the type of fusion to be used depends largely on the purpose and their noninterfering character. A review by Terpe [19] discusses and compares most of the currently used fusion tags in detail and provides advice on using these tags for affinity purification. As with terminus choices, the type of fusion should not interfere with binding either. When using fusion tags there is always the possibility that the tag may interfere with the interaction domain and prevent binding. Due to this possibility, one should always start with a tag that has previously been shown to not interfere with binding in a far-Western experiment. Then during further experimentation conditions (such as investigating binding of protein families described in Fig. 5), this interaction pair should always be included as a positive control. When an experiment fails to detect a known interaction that was demonstrated by other techniques, the linker region separating the protein of interest and the fusion should also be considered. Some interactions may require a large degree of flexibility for the conformational changes that may need to occur during binding. With the type of fusion and linker in mind one should also realize that the assumption that similar proteins will work using the same fusion is a dangerous supposition and does not always hold true.

3.2 Inability to Detect an Interaction Under Any Condition

All of the previous sections have described the parameters in optimizing a far-Western experiment for studying protein–protein interactions. However, there are still situations where an interaction is not detected by this method no matter how much optimization and testing was done. It should be mentioned that Sect. 3.1 focused on problems with the fusion and that a key control should also involve reverse probing, such as using protein C as the probe in Fig. 1. The example of protein Z in Fig. 5 showed that it did not bind any of the nine substrate proteins, yet experiments using BIAcore surface plasmon resonance showed that protein Z binds protein F (results not shown). To rule out problems with the fusion, six other fusion tags of varying sizes and linkers were tested and were still unable to detect any interaction between the two (not shown). This example demonstrates that far-Westerns may not always detect every interaction and that alternative methods should be considered when probing for protein–protein interactions.

With the above observations in mind, it should be noted that the far-Western approach is a positive detection method and that not observing an interaction is not indicative of a negative result. It should be remembered that the denatured version relies on refolding, and both denatured and native versions rely on the stability of the prey protein on the membrane. Therefore, it cannot be assumed that the proteins are properly (re)folded or active under all experimental conditions.

3.3 Background Problems

Issues due to nonspecific interactions can contribute to the background signal during a far-Western experiment. Nonspecific interactions can arise from improper blocking, too high a concentration of protein, and too lengthy an incubation. Starting with a highly pure sample for both proteins of interest can prevent nonspecific interactions by limiting the number of contaminating proteins that could interact. The parameters and general guidelines listed in Table 1 are worth considering when troubleshooting background problems.

Background issues can also arise from the use of denaturants, such as the example described in Fig. 4. In order to obtain denaturation profiles of DmsD binding to the DmsA N-terminal peptide, blots were exposed to DmsD in the presence of the denaturants. Initial experiments resulted in blots that had a high background for the higher denaturant concentrations, which was likely due to removal of some of the blocking proteins while in the denaturant solutions. The background problem was eliminated by addition of a second blocking step following the incubation with DmsD in the presence of denaturants. This completely removed the background and allowed the level of protein interaction to be assayed and quantitated (Fig. 4) and subsequently correlated with other biophysical results.

4 Summary and Conclusions

The use of far-Westerns to study protein–protein interactions is a relatively simple and amenable technique. Many variations of far-Westerns have been developed to study interactions under different conditions ranging from a completely blind proteomics/interactomics approach to the ability to map the effect of specific residues and denaturation profiles on the protein–protein interaction. With the appropriate conditions prescribed in the previous sections, there is usually a good experimental solution for a variety of questions being investigated. While confirmatory experiments should also be done using alternative approaches, far-Westerns have proven to be a simple and efficient method to study protein–protein interactions.

Acknowledgements We thank Jenika M. Howell for her previous work on far-Westerns and protein–protein interactions. We also thank the Canadian Institute for Health Research for funding to RJT towards the research that was used for examples in this review. CSC and TMLW are funded by graduate studentships from the Natural Sciences and Engineering Research Council of Canada.

References

1. Blackwood EM, Eisenman RN (1991) Science 251:1211
2. Kaelin WG, Krek W, Sellers WR, DeCaprio JA, Ajchenbaum F, Fuchs CS, Chittenden T, Li Y, Farnham PJ, Blanar MA, Livingston DM, Flemington EK (1992) Cell 70:351
3. Hall RA (2004) Methods Mol Biol 261:167
4. Laemmli UK (1970) Nature 227:680
5. Burgess RR, Arthur TM, Pietz BC, Thorner J, Emr SD, Abelson JN (2000) Methods Enzymol 328:141
6. Winstone TL, Workentine ML, Sarfo KJ, Binding AJ, Haslam BD, Turner RJ (2006) Arch Biochem Biophys 455:89
7. Sarfo KJ, Winstone TL, Papish AL, Howell JM, Kadir H, Vogel HJ, Turner RJ (2004) Biochem Biophys Res Commun 315:397
8. Sambrook J, Fritsch EF, Maniatis T (1989) In: Nolan C (ed) Molecular cloning: a laboratory manual, vol 3. Cold Spring Harbor Laboratory Press, New York, p 18.60
9. Millipore (2004) Protein blotting handbook. Millipore, Billerica, p 26
10. Pierce (2005) Western blotting handbook and troubleshooting guide. Pierce Biotechnology, Rockford
11. Reddy VM, Kumar B (2000) J Infect Dis 181:1189
12. Whatman (2005) Blotting membranes selection guide. Whatman International, Maidstone
13. Jones KD (1999) IVD Technol March/April http://www.devicelink.com/ivdt/archive/99/03/009.html
14. Oehler S, Alex R, Barker A (1999) Anal Biochem 268:330
15. Desai S, Dworecki B, Cichon E (2001) Anal Biochem 297:94
16. Chan CS, Howell JM, Workentine ML, Turner RJ (2006) Biochem Biophys Res Commun 343:244
17. Mahlknecht U, Ottmann OG, Hoelzer D (2001) J Biotechnol 88:89
18. Waugh DS (2005) Trends Biotechnol 23:316
19. Terpe K (2003) Appl Microbiol Biotechnol 60:523
20. Jenny RJ, Mann KG, Lundblad RL (2003) Protein Expr Purif 31:1
21. Chan CS, Howell JM, Turner RJ (2006) BIOforum Europe 9:36

Adv Biochem Engin/Biotechnol (2008) 110: 215–245
DOI 10.1007/10_2007_084

Published online: 6 October 2007

Using Product Kernels to Predict Protein Interactions

Shawn Martin[1] (✉) · W. Michael Brown[1] · Jean-Loup Faulon[2]

[1]Computational Biology, Sandia National Laboratories, PO Box 5800,
Albuquerque, NM 87185-1316, USA
smartin@sandia.gov

[2]Computational Bioscience, Sandia National Laboratories, PO Box 5800,
Albuquerque, NM 87185-1413, USA

Abstract There is a wide variety of experimental methods for the identification of protein interactions. This variety has in turn spurred the development of numerous different computational approaches for modeling and predicting protein interactions. These methods range from detailed structure-based methods capable of operating on only a single pair of proteins at a time to approximate statistical methods capable of making predictions on multiple proteomes simultaneously. In this chapter, we provide a brief discussion of the relative merits of different experimental and computational methods available for identifying protein interactions. Then we focus on the application of our particular (computational) method using Support Vector Machine product kernels. We describe our method in detail and discuss the application of the method for predicting protein–protein interactions, β-strand interactions, and protein–chemical interactions.

Keywords β-strand interactions · Product kernels · Protein–chemical interactions ·
Protein–protein interactions · Support Vector Machines

1 Introduction

Protein interactions occur in many, if not most, cellular processes. This fact has motivated the development of a multitude of experimental methods for the identification of protein interactions. Experimental methods can be roughly categorized according to throughput [1]. Detailed, low-throughput methods yield accurate information about interaction, structure, mechanics, kinetics, and dynamics. Low-throughput methods include X-ray crystallography and NMR spectroscopy, fluorescence resonance energy transfer [2], surface Plasmon resonance [3], atomic force microscopy [4], and electron microscopy [5]. High-throughput methods, on the other hand, yield less accurate statistical information about interactions, but allow for the experiments to be performed on a large proteome-wide scale. High-throughput methods include Yeast two hybrid [6, 7], affinity purification and mass spectrometry [8], DNA microarray gene coexpression [9], protein microarrays [10], synthetic lethality [11], and phage display [12].

For each of the experimental methods available there are corresponding computational approaches designed to take advantage of the unique information produced from a given experiment [13]. Structure-based methods are low-throughput but produce detailed experimental data to provide highly accurate models and predictions on a small scale [14]. Structure-based methods include docking [15], fold recognition by threading [16], and certain domain/motif-based methods [17, 18]. Sequence-based methods use high-throughput data to make less accurate statistical predictions on entire proteomes. Sequence-based methods include conservation of gene neighborhoods and gene order [19, 20], phylogenetic profiling [21, 22], gene fusion [23], co-evolution [24, 25], association methods [26–28], and Bayesian models [29, 30]. The greater variety of sequence-based methods compared to structure-based methods reflects the greater amount of sequence-based data available.

Our method (to be described in detail in this chapter) is a statistical sequence-based approach that has been categorized as an association method [1]. The prototypical association method was first suggested by Sprinzak and Margalit [28]. In an association method, characteristic sequences or motifs are identified that separate interacting from non-interacting protein pairs. Once the appropriate motifs are isolated, a machine learning classification method is applied in order to make predictions on new protein pairs. In Sprinzak and Margalit's approach, InterPro domains [31] are used as motifs and a log-odds ratio is used as a classifier. In our approach, the motifs are automatically identified from subsequences of protein amino acid sequences using a Support Vector Machine classification method.

In this chapter, we describe our method in detail, including some background on Support Vector Machines and kernel methods (Sect. 2), string and graph kernels (Sects. 3 and 4), tensor product kernels (Sect. 5), and applications of our approach to protein–protein interaction prediction (Sect. 6), β-strand interaction prediction and ordering (Sect. 7), and protein–chemical interaction prediction (Sect. 8).

2 Support Vector Machines

Our method for predicting protein interaction is based on the use of Support Vector Machines (SVMs). SVMs belong to a class of algorithms known as kernel methods [32, 33]. Kernel methods are machine learning algorithms that can accommodate nearly any type of data set as input. The reason that kernel methods are so general is due to their use of kernel functions. Kernel functions provide a formal separation between data and algorithm. Once the user provides data and a kernel function for that data, any of a variety of algorithms can be easily applied, including SVM classification [33, 34], regression [35], clustering [36], and dimensionality reduction [37, 38]. In this section we provide an overview of the kernel-based approach, with an emphasis on the prototypical kernel method, the SVM.

At its core, a SVM is a linear binary classifier. Suppose we have a dataset $\{\boldsymbol{x}_i\} \subseteq R^n$, and that each point $\boldsymbol{x}_i$ in our dataset has a corresponding class label $y_i \in \{\pm 1\}$. Our goal is to separate the points in our dataset according to their class label. Since there are two classes, this is known as binary classification. An SVM attempts this classification by using a linear hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b$, ($\boldsymbol{w} \neq 0$), as shown in Fig. 1.

Assuming that our dataset is in fact linearly separable, there will in general be many possible hyperplanes that can achieve the separation. An SVM uses an optimal separating hyperplane known as the maximal margin hyperplane. The hyperplane margin is twice the distance from the separating hyperplane to the nearest point in one (or the other) of the two classes. In Fig. 1 this is the distance between the two dotted lines. As might be inferred from the name, the maximal margin hyperplane is found by solving an optimization problem. Without going into detail [39, 40], the SVM hyperplane is found by solving the quadratic programming problem

$$\begin{aligned} \max_\alpha \quad & \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j - \sum_i \alpha_i \\ \text{s.t.} \quad & \sum_i y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \,, \end{aligned} \tag{1}$$

where $\boldsymbol{w} = \sum_i y_i \alpha_i \boldsymbol{x}_i$ is the normal to the SVM hyperplane. Using $\boldsymbol{w}$ we form the SVM decision function $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$, where b is obtained im-

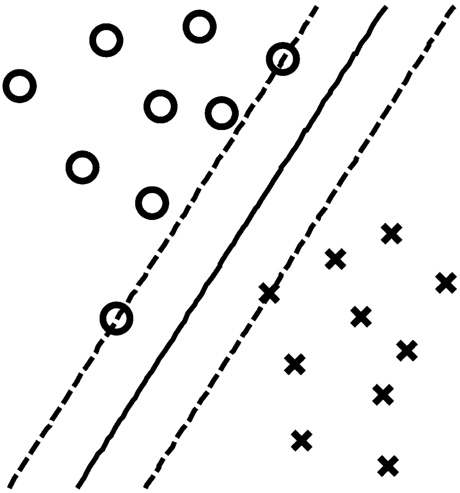


Fig. 1 A linear SVM. The SVM decision boundary is defined by the support vectors, which are the examples falling on the *dotted lines*. The distance between the *dotted lines* is known as the margin and is maximized to obtain the SVM decision boundary, shown as the *solid line* separating the circles (class 1) from the x marks (class –1)

plicitly [39,40]. We note that $\alpha_i \neq 0$ only when $\boldsymbol{x}_i$ is a support vector (see Fig. 1), and that the formulation given in Eq. 1 is actually the soft margin generalization of the SVM quadratic programming problem. The soft margin generalization accounts for class label errors by incorporating a bound C on the variables α_i [39,40]. If there are no errors in the class labels, we may assume that $C = \infty$.

Solving the quadratic program in Eq. 1 is known as *training* the SVM. Once the SVM has been trained, we can use the SVM decision function $f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ to make predictions on new samples not in our original dataset. If $\boldsymbol{x}$ is a new sample then $f(\boldsymbol{x})$ is the predicted class label. The magnitude of $\boldsymbol{w}^T\boldsymbol{x} + b$ can also give us an indication of the strength of our prediction.

The SVM problem given in Eq. 1 only applies to datasets $\{\boldsymbol{x}_i\} \subseteq R^n$. Often, however, we want to use an SVM on a dataset that is not a subset of R^n. This occurs in the case of biology and chemistry problems, when we are likely to use amino acid sequences or chemical structures to describe our data. Fortunately, there is a ready solution to this problem, formalized in the use of kernel functions.

Suppose our data $\{\boldsymbol{x}_i\} \subseteq S$, where S might be the set of all finite length protein sequences or all finite diameter chemical graphs. We can then define a kernel function as a map $k : S \times S \to R$ such that

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}_j)\,, \tag{2}$$

where $\Phi: S \rightarrow F$ is a map from our original data space S into a space F with a defined dot product (such as R^N). Technically, F can be a potentially infinite dimensional separable Hilbert space, but for our purposes F is always R^N for some sufficiently large N. Thus our notation $\Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x}_j)$ is just the dot product of $\Phi(\boldsymbol{x}_i)$ with $\Phi(\boldsymbol{x}_j)$ in R^N.

Once we have defined a kernel function, we simply replace the dot product $\boldsymbol{x}_i^T\boldsymbol{x}_j$ in Eq. 1 with the kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ to obtain the full SVM quadratic programming problem

$$\begin{aligned} \max_\alpha \quad & \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_i \alpha_i \\ \text{s.t.} \quad & \sum_i y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \,. \end{aligned} \tag{3}$$

A similar procedure can be used for any method that is written in terms of dot products. Methods of this type are known as kernel methods.

At first, the use of kernel functions may seem overly formal, considering the fact that we are really just replacing our dataset $\{\boldsymbol{x}_i\}$ with the dataset $\{\Phi(\boldsymbol{x}_i)\}$ before we perform our calculation. However, this formalism encourages a very useful separation between data and method. A computer scientist interested in developing algorithms or methods can take as a starting point a kernel matrix (with entries given by $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$), while a scientist interested in a particular problem needs only provide a kernel matrix in order apply the methods developed by the computer scientist. This approach is even more appealing when we consider that a kernel matrix is in fact a matrix of pairwise similarities. The scientist interested in applying a kernel method needs only to answer one question: what is a good measure of similarity between the objects in my study? Once this question has been answered, any number of machine learning algorithms can be brought to bear for exploring the dataset and making predictions.

3
String Kernels

Our method for predicting protein interactions relies on SVMs. According to the previous discussion (Sect. 2), this approach has two implications: first, we must have at hand databases of known interactions so that we may train SVMs; second, we must supply kernel functions that provide methods for computing the similarity between our objects of study (e.g., proteins). The first constraint is satisfied by high-throughput experimental methods [1] and databases where the results of such experiments are archived. Examples include the Database of Interacting Proteins (DIP) [41], the Biomolecular Interaction Network Database (BIND) [42], and the Munich Information Center for Protein Sequences (MIPS) [43]. The second constraint (supplying appropriate kernel functions) is the subject of the next two sections.

Our method can at present be used to predict interactions between proteins, β-strands, and chemicals. These predictions are made by first computing similarities between pairs of proteins, pairs of β-strands, and pairs of chemicals. Such computations are carried out using string kernels [44, 45] and graph kernels [46–48].

In the case of proteins, we use primary sequence. In mathematical terms, a protein is a finite length string over an alphabet of 20 letters corresponding to the 20 possible amino acid residues. To calculate the similarity between two proteins we must calculate the similarity between two strings. Hence we use a string kernel. In the case of chemicals, we represent structure using chemical graphs where nodes correspond to atoms and edges correspond to bond types. Thus to compute the similarity between two chemicals we use graph kernels.

To define a string kernel, we first define a map Φ_s^l: {finite length amino acid strings} $\to Z^{N_l}$, where N_l is the number of possible amino acid sequences of length l. If we denote by z_j a basis of Z^{N_l} where each basis vector z_j corresponds to an amino acid sequence of length l then Φ_s^l is given by

$$\Phi_s^l(P_i) = \sum_j \sigma_j z_j \,, \tag{4}$$

where P_i is a finite length amino acid string and σ_j is the number of times that the amino acid string corresponding to z_j occurs in the string P_i. We then define the string kernel k_s^l (P_i, P_j) between two proteins P_i and P_j by

$$k_s^l(P_i, P_j) = \Phi_s^l(P_i)^T \Phi_s^l(P_j) \,. \tag{5}$$

As an example, suppose we have amino acid strings LVMLVM and LVMTTM. We want to calculate Φ_s^3 (LVMLVM), Φ_s^3 (LVMMTT), and k_s^3 (LVMLVM, LVMTTM). There are four substrings of length 3 (also known as trimers) in LVMLVM, namely LVM, VML, MLV, and LVM. There are also four substring of length 3 in LVMTTM, namely LVM, VMT, MTT, and TTM. Suppose that z_1 corresponds to LVM, z_2 corresponds to VML, z_3 corresponds to MLV, z_4 corresponds to VMT, z_5 corresponds to MTT, and z_6 corresponds to TTM. We now see that Φ_s^3 (LVMLVM) = $(2, 1, 1, 0, 0, 0)^T$ and Φ_s^3 (LVMMTT) = $(1, 0, 0, 1, 1, 1)^T$ so that k_s^3 (LVMLVM, LVMMTT) = 2. A visual demonstration arriving at Φ_s^3 (LVMLVM) is shown in Fig. 2.

The string kernel just described was introduced by Leslie et al. [44, 45]. We use a slight variant on this kernel based on the idea that our similarities should be identical regardless of the direction that we traverse the amino acid sequence. To accommodate this symmetry, we convert all length l substrings into height $h = (l - 1)/2$ amino acid *signatures* before applying the previous

$$\Phi_s^3(\mathrm{LVMLVM}) = \sum\nolimits_j \sigma_j \mathbf{z}_j = 2 \underset{\mathrm{LVM}}{\underset{\updownarrow}{\begin{pmatrix}1\\0\\0\\0\\0\\\vdots\end{pmatrix}}} + 1 \underset{\mathrm{VML}}{\underset{\updownarrow}{\begin{pmatrix}0\\1\\0\\0\\0\\\vdots\end{pmatrix}}} + 1 \underset{\mathrm{MLV}}{\underset{\updownarrow}{\begin{pmatrix}0\\0\\1\\0\\0\\\vdots\end{pmatrix}}} + 0 \underset{\mathrm{VMT}}{\underset{\updownarrow}{\begin{pmatrix}0\\0\\0\\1\\0\\\vdots\end{pmatrix}}} + \cdots$$

Fig. 2 Representing an amino acid string as a vector. Here we show how a finite length amino acid string can be represented as a vector. The string VLMVLM is mapped to a vector Φ_s^l (VLMVLM) by counting the number of occurrences of different (trimer) substrings of length 3

definition. A height $h = (l - 1)/2$ amino acid signature is obtained from an odd length l substring P by first forming three disjoint substrings P_1, P_2, P_3 from P. P_2 is given by the middle character of P (hence we require that P is of odd length), P_1 is given by the left-hand side of P starting before P_2 and traversing left, and P_3 is given by the right-hand side of P starting after P_2 and traversing right. The signature corresponding to the substring P is given by $P_2P_1P_3$ when $P_1 < P_3$ in alphabetical order, or $P_2P_3P_1$ if $P_3 < P_1$.

After converting length l substrings to height $h = (l - 1)/2$ signatures, we obtain maps Φ_s^l and kernels k_s^l that are invariant under string traversal direction, although we must now assume that l is an odd integer. To revisit our example, suppose we want to calculate k_s^3 (LVMLVM, MTTMVL) using height 1 signatures. (Note that LVMTTM has been replaced by the reverse ordered MTTMVL.) For the string LVMLVM our four substrings LVM, VML, MLV, and LVM are converted into signatures VLM, MLV, LMV, and VLM. For the string MTTMVL our four substrings MTT, TTM, TMV, and MVL are converted to signatures TMT, TMT, MTV, and VLM. Letting z_1 correspond to VLM, z_2 correspond to MLV, z_3 correspond to LMV, z_4 correspond to TMT, and z_5 correspond to MTV we have Φ_s^3(LVMLVM) = $(2, 1, 1, 0, 0)^T$ and Φ_s^3 (MTTMVL) = $(1, 0, 0, 2, 1, 0)^T$ so that k_s^3 (LVMLVM, MTTMVL) = 2. This is of course the same result that was obtained previously using length 3 substrings.

4 Graph Kernels

In order to make predictions about protein–chemical interactions, we need both string kernels and graph kernels. The string kernels can be used to measure similarity between amino acid sequences and the graph kernels can be used to measure similarity between chemical structures. We use graph kernels because we represent chemical structures using undirected vertex- and edge-labeled graphs. Each vertex corresponds to an atom (e.g., C, N, O, etc.), and each edge corresponds to a bond type (e.g., single, double, aromatic, etc.). Such graphs are known as molecular graphs.

Our graph kernel measures the similarity between two molecular graphs. This is done in a manner analogous to the previous string kernel. In the case of the string kernel, we represented an amino acid string as a vector counting the number of substrings in the original string. In the case of the graph kernel, we represent a molecular graph as a vector counting the number of subgraphs in our original graph. This representation is known as molecular signature [49–51].

The molecular signature representation of a molecular graph is a vector whose components correspond to atomic signatures. Each component of the signature vector counts the number of occurrences of a particular atomic signature in the molecule. An atomic signature is a canonical representation of the subgraph surrounding a particular atom. This subgraph includes all atoms up to a predefined distance from the given atom. As in the case of amino acid substrings, this distance is called the signature height.

Formally, we define a map Φ_g^h: {molecular graphs} $\to Z^{N_h}$, where N_h is the number of possible atomic signatures of height h. Borrowing the notation from Sect. 2, we again denote by $\boldsymbol{z}_j$ a basis of Z^{N_h} where each basis vector $\boldsymbol{z}_j$ corresponds to a height h subgraph of a molecular graph. If M_i denotes a molecular graph then Φ_g^h is given by

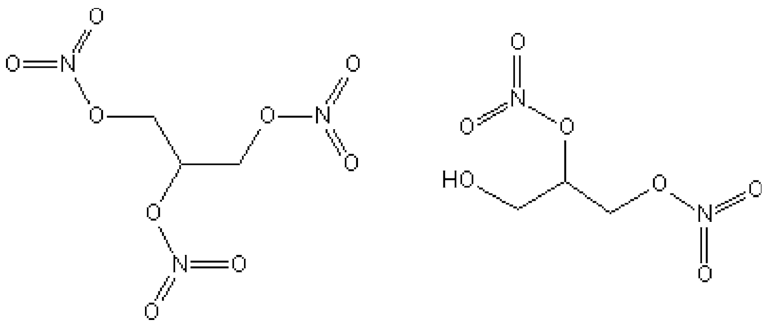
$$\Phi_g^h(M_i) = \sum_j \sigma_j \boldsymbol{z}_j \,, \tag{6}$$

where σ_j is the number of times that the molecular subgraph corresponding to $\boldsymbol{z}_j$ occurs in M_i. Now we define a graph kernel just like we defined the string kernel. Namely, the graph kernel k_g^h (M_i, M_j) between two molecules M_i and M_j is given by

$$k_g^h(M_i, M_j) = \Phi_g^h(M_i)^T \Phi_g^h(M_j) \,. \tag{7}$$

To provide an example, consider the two molecules shown in Fig. 3. Both molecules, nitroglycerine and 1,2-dinitroglycerine, are represented as undirected edge- and vertex-labeled molecular graphs (carbons and hydrogens are implicit). To obtain height 1 signatures from these graphs, we first visit each node in each graph and record the subgraph formed by that node and its neighbors. This is known as a height 1 atomic signature. If we wanted to compute height 2 signature we would have to visit the neighbors of the neighbors.

The atomic signatures are recorded as strings $A_1(b_2A_2b_3A_3\ldots)$, where A_1 is the vertex type of the root node (the node we are visiting), A_2 is a neighbor



Nitroglycerine	**1,2-Dinitroglycerine**
$3\,\mathrm{O(NC)} \leftrightarrow \mathbf{z}_1$	$2\,\mathrm{O(NC)} \leftrightarrow \mathbf{z}_1$
$6\,\mathrm{O(=N)} \leftrightarrow \mathbf{z}_2$	$1\,\mathrm{O(HC)} \leftrightarrow \mathbf{z}_7$
$3\,\mathrm{N(O=O=O)} \leftrightarrow \mathbf{z}_3$	$4\,\mathrm{O(=N)} \leftrightarrow \mathbf{z}_2$
$5\,\mathrm{H(C)} \leftrightarrow \mathbf{z}_4$	$2\,\mathrm{N(O=O=O)} \leftrightarrow \mathbf{z}_3$
$2\,\mathrm{C(OHHC)} \leftrightarrow \mathbf{z}_5$	$1\,\mathrm{H(O)} \leftrightarrow \mathbf{z}_8$
$1\,\mathrm{C(OHCC)} \leftrightarrow \mathbf{z}_6$	$5\,\mathrm{H(C)} \leftrightarrow \mathbf{z}_4$
	$2\,\mathrm{C(OHHC)} \leftrightarrow \mathbf{z}_5$
$\updownarrow$	$1\,\mathrm{C(OHCC)} \leftrightarrow \mathbf{z}_6$
	$\updownarrow$
$(3,6,3,5,2,1,0,0)^T$	$(2,4,2,5,2,1,1,1)^T$

Fig. 3 Height one signature representations of nitroglycerine (*left*) and 1,2-dinitroglycerine (*right*). Proceeding from top to bottom we show (*top row*) the molecular graph representations of the two molecules, (*middle row*) the number of occurrences of height 1 atomic signatures, and (*bottom row*) the signature vector representation of the two molecules. The atomic signature occurrences give the number of times a given chemical fragment occurs in the molecules. In the case of nitroglycerine, we have three oxygen bonded to a nitrogen and a carbon (shown as 3 O(NC)), 6 oxygen double bonded to a nitrogen (shown as 6 O(= N)), 3 nitrogen bonded to an oxygen and double bonded to two other oxygen atoms (shown as 3 N(O = O = O)), and so on

of A_1 with bond type b_2, A_3 is a neighbor of A_1 with bond type b_3, etc. If b_2, b_3, ... are single bonds, then they are omitted. This representation is canonical if we alphabetize the list (b_2A_2, b_3A_3, ...) of bonds and atoms [51]. In Fig. 3, we have written oxygen bonded to nitrogen and carbon as O(NC), and oxygen double bonded to nitrogen as O(= N). Note that the atomic signature representation is in fact a generalization of the amino acid substring representation that we used in Sect. 3.

After visiting each node of each molecular graph, we obtain a list of atomic signature string representations. This list is identified with a set z_1, z_2, ... of basis vectors. In Fig. 3 we have identified O(NC) with z_1 and O(= N) with z_2. Using these basis vectors, we record the number of times a given atomic signature subgraph occurs in a molecular graph to obtain our molecular signature vector representation. Since O(NC) occurs three times in nitroglycerine and O(= N) occurs six times, the first two entries of our signature vector for nitroglycerine are 3 and 6. The full vector is $(3,6,3,5,2,1,0,0,0)^T$. This analysis is also performed on 1,2-dinitroglycerine to get the signature vector $(2,4,2,5,2,1,1,1)^T$.

Once we have molecular signature vectors for the various molecular graphs in our dataset, it is a simple matter to compute kernel similarities by taking dot products of signature vectors. Using the signature vectors for nitroglycerine and 1,2-dinitroglycerine we compute a similarity of $(3,6,3,5,2,1,0,0)^T \cdot (2,4,2,5,2,1,1,1)^T = 66$.

Finally, we define a graph-based kernel for use in comparing chemical reactions. This is a straightforward extension of the signature-based kernel just described and is applicable to predicting enzymes that catalyze reactions. We first define a reaction signature for an enzymatic reaction. We assume that all enzymatic reactions take the general form R: $s_1S_1 + s_2S_2 + \ldots + s_nS_n \rightarrow p_1P_1 + p_2P_2 + \ldots + p_mP_m$, where s_i and p_j are the stoichiometric coefficients of substrates S_i and products P_j. The height h signature of reaction R is then defined by

$$\Phi_g^h(R) = \sum_j p_j \Phi_g^h(P_j) - \sum_i s_i \Phi_g^h(S_i)\,, \tag{8}$$

where $\Phi_g^h(P_j)$ and $\Phi_g^h(S_j)$ are the height h molecular signatures of substrate S_i and product P_j computed using Eq. 6.

5 Product Kernels

Using the string kernels from Sect. 3, we can compare two proteins based on their similarity. However, we would still like to compare two *pairs* of proteins, say an interacting protein pair with another interacting pair, or with a non-interacting pair. For this purpose, we introduce a product kernel. Product kernels were first introduced for predicting protein–protein interactions [26, 27], but they are generally applicable to any dataset with pairs of objects, including drug interaction datasets containing protein-chemical pairs.

To define a product kernel, we first recall the definition of a tensor product [52]. In its simplest incarnation, the tensor product between $\boldsymbol{x} = (x_1,\ldots,x_n)^T \in R^n$ and $\boldsymbol{y} = (y_1,\ldots,y_m)^T \in R^m$ is $\boldsymbol{x} \otimes \boldsymbol{y} = (x_1y_1, x_1y_2,\ldots, x_1y_m, x_2y_1, \ldots, x_ny_m)^T \in R^{nm}$. The entries in $\boldsymbol{x} \otimes \boldsymbol{y}$ are the same as the entries in $\boldsymbol{xy}^T$.

By using the tensor product, we can represent a protein pair (P_i, P_j) as a vector. We define $\Phi_{s\otimes s}^{l_1\otimes l_2}$: {pairs of amino acid sequences} $\rightarrow Z^{N_{l_1}N_{l_2}}$ by

$$\Phi_{s\otimes s}^{l_1\otimes l_2}(P_i, P_j) = \Phi_s^{l_1}(P_i) \otimes \Phi_s^{l_2}(P_j)\,, \tag{9}$$

where we use the previous definition of Φ_s^l in Eq. 4. Now the product kernel between two protein pairs (P_{i_1}, P_{j_1}) and (P_{i_2}, P_{j_2}) can be defined as

$$k_{s\otimes s}^{l_1\otimes l_2}\left(\left(P_{i_1}, P_{j_1}\right), \left(P_{i_2}, P_{j_2}\right)\right) = \Phi_{s\otimes s}^{l_1\otimes l_2}\left(P_{i_1}, P_{j_2}\right)^T \Phi_{s\otimes s}^{l_1\otimes l_2}\left(P_{i_2}, P_{j_2}\right). \tag{10}$$

Although we have used l_1 and l_2 in our definitions, we have never in practice encountered a situation requiring $l_1 \neq l_2$. In addition, since the notation $s \otimes s$ is redundant, we use Φ_s^l to denote both $\Phi_s^l(P_i)$ and $\Phi_{s\otimes s}^{l\otimes l}(P_i, P_j)$, using the context (P_i) or (P_i, P_j) to differentiate Φ_s^l from $\Phi_{s\otimes s}^{l\otimes l}$. Similarly, we use k_s^l to denote both k_s^l (P_i,P_j) and $k_{s\otimes s}^{l\otimes l}((P_{i_1}, P_{j_1}), (P_{i_2}, P_{j_2}))$.

In practice, it is cumbersome to compute $\Phi_s^l(P_i, P_j)$ and even more cumbersome to compute $k_s^l((P_{i_1}, P_{j_1}), (P_{i_2}, P_{j_2}))$. Fortunately, a straightforward observation can help remedy this situation. If we write $\boldsymbol{p}_{i_1} = \Phi_s^l(P_{i_1})$, $\boldsymbol{p}_{j_1} = \Phi_s^l(P_{j_1})$, $\boldsymbol{p}_{i_2} = \Phi_s^l(P_{i_2})$, and $\boldsymbol{p}_{j_2} = \Phi_s^l(P_{j_2})$ then we can see that

$$\begin{aligned}
k_s^l\left(\left(P_{i_1}, P_{j_1}\right), \left(P_{i_2}, P_{j_2}\right)\right) &= \left(\Phi_s^l(P_{i_1}) \otimes \Phi_s^l(P_{j_1})\right)^T \left(\Phi_s^l(P_{i_2}) \otimes \Phi_s^l(P_{j_2})\right) \\
&= \operatorname{trace}\left(\left(\boldsymbol{p}_{i_1}\boldsymbol{p}_{j_1}^T\right)\left(\boldsymbol{p}_{i_2}\boldsymbol{p}_{j_2}^T\right)^T\right) \\
&= \operatorname{trace}\left(\boldsymbol{p}_{i_1}\boldsymbol{p}_{j_1}^T\boldsymbol{p}_{j_2}\boldsymbol{p}_{i_2}^T\right) \\
&= \boldsymbol{p}_{j_1}^T\boldsymbol{p}_{j_2}\operatorname{trace}\left(\boldsymbol{p}_{i_1}\boldsymbol{p}_{i_2}^T\right) \\
&= \boldsymbol{p}_{j_1}^T\boldsymbol{p}_{j_2}\boldsymbol{p}_{i_1}^T\boldsymbol{p}_{i_2} \\
&= k_s^l\left(P_{i_1}, P_{i_2}\right) k_s^l\left(P_{j_1}, P_{j_2}\right)\,,
\end{aligned} \tag{11}$$

where trace (X) is the sum of the diagonal elements in the square matrix X. This result shows that we can compute $k_s^l((P_{i_1}, P_{j_1}), (P_{i_2}, P_{j_2}))$ by instead computing the much simpler values $k_s^l(P_{i_1}, P_{i_2})$ and $k_s^l(P_{j_1}, P_{j_2})$.

Before continuing, let's take a moment to work out an example. For simplicity, let's compare the pair (LVMLVM, MTTMVL) of amino acid strings previously considered in Section 3 with an additional pair (VLMVLM, TTMVLM) of amino acid strings. For the first pair of amino acid strings, we have height 1 signature substrings VLM $\leftrightarrow z_1$, MLV $\leftrightarrow z_2$, LMV $\leftrightarrow z_3$, TMT $\leftrightarrow z_4$, and MTV $\leftrightarrow z_5$. The second pair (VLMVLM, TTMVLM) contributes no additional signatures so that we may write (LVMLVM, MTTMVL) as $((2,1,1,0,0)^T$, $(1,0,0,2,1)^T)$ and (VLMVLM, TTMVLM) as $((1,1,2,0,0)^T$, $(1,0,1,1,1)^T)$. According to definition by Eq. 9, we have

$$\begin{aligned}\Phi_s^3(\text{LVMLVM, MTTMVL}) &= (2,1,1,0,0)^T \otimes (1,0,0,2,1)^T \\ &= (2{,}1{,}1{,}0{,}0)^T(1{,}0{,}0{,}2{,}1) \\ &= \begin{pmatrix} 2 & 0 & 0 & 4 & 2 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},\end{aligned} \tag{12}$$

and

$$\begin{aligned}\Phi_s^3(\text{VLMVLM, TTMVLM}) &= (1{,}1{,}2{,}0{,}0)^T \otimes (1,0,1,1,1)^T \\ &= (1{,}1{,}2{,}0{,}0)^T(1{,}0{,}1{,}1{,}1) \\ &= \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 2 & 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.\end{aligned} \tag{13}$$

Now when we consider the matrices in Eqs. 12 and 13 as vectors, we get k_s^3 ((LVMLVM, MTTMVL), (VLMVLM, TTMVLM)) = 20. Of course, as described in Eq. 11, we could also compute k_s^3 ((LVMLVM, MTTMVL), (VLMVLM, TTMVLM)) = k_s^3 (LVMLVM, VLMVLM) $\times k_s^3$ (MTTMVL, TTMVLM) = $5 \times 4 = 20$.

As mentioned previously, product kernels are not limited to computing protein pair similarity. They are just as easily adapted to protein-chemical or protein-reaction pair similarity calculation. In the case of two protein-chemical pairs (P_i, C_i) and (P_j, C_j) we define

$$\Phi_{s\otimes g}^{l\otimes h}(P_i, C_i) = \Phi_s^l(P_i) \otimes \Phi_g^h(C_i)\,, \tag{14}$$

so that we can define

$$k_{s\otimes g}^{l\otimes h}\left((P_i, C_i), (P_j, C_j)\right) = \Phi_{s\otimes g}^{l\otimes h}(P_i, C_i)^T \Phi_{s\otimes g}^{l\otimes h}(P_j, C_j)\,. \tag{15}$$

Using the relation in Eq. 11, we also have

$$k_{s\otimes g}^{l\otimes h}\left((P_i, C_i), (P_j, C_j)\right) = k_s^l\left(P_i, P_j\right) k_g^h(C_i, C_j)\,. \tag{16}$$

These definitions also work for protein-reaction pair similarity calculations, simply by substituting R_i and R_j for C_i and C_j.

Finally, various options for normalized, symmetric, and nonlinear versions of the product kernel are available for tuning classifier performance [27]. A symmetric product kernel for protein pairs can be obtained by redefining $\Phi_s^l(P_i, P_j)$ as

$$\Phi_s^l(P_i, P_j) = \Phi_s^l(P_i) \otimes \Phi_s^l(P_j) + \Phi_s^l(P_j) \otimes \Phi_s^l(P_i)\,. \tag{17}$$

A normalized product kernel is often desirable and is also easily obtained. If k is a generic kernel then $k(\boldsymbol{x},\boldsymbol{y})/\sqrt{k(\boldsymbol{x},\boldsymbol{x})k(\boldsymbol{y},\boldsymbol{y})}$ is the normalized version of k. Occasionally, it is useful to compose a generic kernel with a Gaussian to obtain a nonlinear kernel. This can improve performance and is done by computing $\exp(-\gamma(k(\boldsymbol{x},\boldsymbol{x}) - 2k(\boldsymbol{x},\boldsymbol{y}) + k(\boldsymbol{y},\boldsymbol{y})))$, where γ is some positive constant.

6 Predicting Protein–Protein Interactions

Our first application of a product kernel was to predict protein–protein interactions [27]. For this application, we used the product of two string kernels of height 1 as described in Sect. 3. We tested our algorithm on publicly available *S. cervisiae* [41, 53] and *H. pylori* [54] data. For each of these datasets, we drew at random an equal number of non-interacting protein pairs (negatives) for training the SVMs.

To assess the ability of our method to predict protein–protein interactions, we used 10-fold cross-validation. Ten-fold cross-validation is a process whereby the dataset under investigation is divided at random into ten equally sized subsets (known as *folds*). Each subset is held as a test set while the remaining nine subsets are used as training sets. Predictions are made on each test set and the results are used to compute accuracy, precision, and sensitivity. Accuracy is (TP + TN)/(TP + FP + TN + FN), precision is TP/(TP + FP), and sensitivity is TP/(TP + FN), where TP, TN, FP, and FN are counts of true-positives, true-negatives, false-positives, and false-negatives.

Accuracy gives the overall performance of the classifier, precision gives the percentage of positive predictions that are actually positive, and sensitivity gives the percentage of actual positives that are predicted. By looking at precision and sensitivity, we can determine if a classifier will identify positives correctly. This is important in the case of protein–protein interaction prediction, since there are likely to be many more non-interactions (negatives) than interactions (positives).

If a classifier has a high precision and a low sensitivity, then it is likely to be correct when it makes a positive prediction, although it will make many false-negative predictions. Conversely, a classifier with a low precision and a high sensitivity is likely to identify most true positives, even though many of its predictions will be false. In some sense, the first classifier is too conservative while the second is too optimistic.

Using yeast SH3 domain data [53], data for the full yeast proteome [41], and *H. pylori* data [54], we computed the 10-fold cross-validation statistics shown in Table 1. The same datasets and statistics were also used in other methods, allowing us to compare our method to the position specific scoring matrix (PSSM) method [53], the correlated sequence signature (CSS) method [28], and the SVM method developed by Bock and Gough [55]. To

Table 1 Comparison of methods. Here we compare the product kernel method and its alternatives using yeast and *H. pylori* datasets. Accuracy, precision, and sensitivity are computed as percentages using ten-fold cross-validation. Product refers to our method; PSSM refers to the position sensitive scoring matrix method [53]; CSS refers to the correlated sequence signature method [28]; and Bock & Gough refers to the reported results in [55] on *H. pylori*

	Accuracy	Precision	Sensitivity
Yeast SH3			
Product	80.7%	71.4%	75.2%
PSSM	75.4%	68.8%	81.3%
Full yeast			
Product	69.0%	71.5%	63.2%
CSS	68.8%	79.8%	50.0%
H. pylori			
Product	83.4%	85.7%	79.9%
Bock & Gough	75.8%	80.2%	69.8%

avoid confusion in the below discussion, we denote the SVM method by BG (for Bock and Gough) and we note that the signatures in the CSS method are unrelated to the signatures used in our method.

The results from Table 1 show that our method performs well compared to PSSMs, the CSS method, and the method of BG. Perhaps more interesting is a qualitative comparison of the different methods (previously described in [27]). Such a comparison gives insight into the underlying biological reasons that a product kernel type method might work.

Our method is most similar to the methods of BG and the method of CSS. Our method and the BG method both use SVMs, sequence information, and experimental data to predict protein–protein information. However, while Bock and Gough transform sequence information into physicochemical information (charge, hydrophobicity, and surface tension), the string kernel does not require us to perform such a transformation. Furthermore, Bock and Gough encode and compare protein pairs by concatenating normalized versions of the amino acid sequences of each protein, and hence use a global representation of a protein pair. Although local information can be encoded implicitly in BG (by using a nonlinear kernel, such as a polynomial kernel), our method uses explicit pair-oriented, local sequence information (product kernels). In other words, we are looking for amino acid subsequence pairs which occur together when two proteins interact.

Our use of subsequence pairs is similar to the correlated InterPro entry pairs used in CSS [28]. However, instead of using InterPro entries [31], we use an automatic method for generating potential domains which depends only on sequence information. Finally, our method has the advantage of using a principled method (SVMs) to obtain our final classifier.

To further establish the relationship between our technique and the CSS method, we explored the potential of our method for predicting domains. For this exercise, we returned to the full yeast dataset, where we selected the top ten pairs predicted by our method to interact (see Table 2). Using the 14 proteins present in these pairs, we constructed domain-sized amino acid subsequences by sliding a window across each of the protein sequences. Our window was of size 50, and we moved the window in increments of ten amino acid residues. Using this method we obtained 1681 subsequences, each 50 amino acids long.

Table 2 Top ten yeast protein pairs predicted to bind using the product kernel method

Swiss-Prot ID	Swiss-Prot ID	Prediction
P27895	P27895	2.36
P27895	P36022	1.95
P22579	Q00916	1.61
P00546	Q02821	1.56
P40064	Q00916	1.53
P50875	P19659	1.48
P19659	P09547	1.46
Q06142	Q02821	1.44
Q06245	P19524	1.41
Q00916	P08964	1.41

From the model obtained using the full yeast dataset, we predicted which pairs of these subsequences would interact. By examining the positions of these interacting subsequences within the full protein sequences, we could make domain predictions as shown in Fig. 4.

In Fig. 4 we examine the domain predictions for P09547 and P50875. In particular, Fig. 4a shows that the region between 300 and 400 is more likely to bind with the other regions (windows) among the 14 proteins examined. We hypothesize that this is a domain. In Fig. 4b it is shown that this domain binds with itself and therefore that P09547 binds with itself. This is only a prediction, but when we examine a known interaction, P09547 with P50875, we see a similar result. In fact, we see in Fig. 4c that P09547 binds with P50875 and that our previously hypothesized domain binds to regions 100–150, 200–250, 400–450, and 500–550. We again hypothesize that these are domains, this time in P50875.

To see that these predictions match known information, we looked up the domain information for P09547 and P50875 in the Swiss-Prot database. There were two domains mentioned for P09547, an Asn/Thr-rich region from 5 to 65, and a Gln-rich region from 337–385. Our domain correlates well with the Gln-rich region. For P50875 Swiss-Prot gives five domains: a Poly-Gln domain from

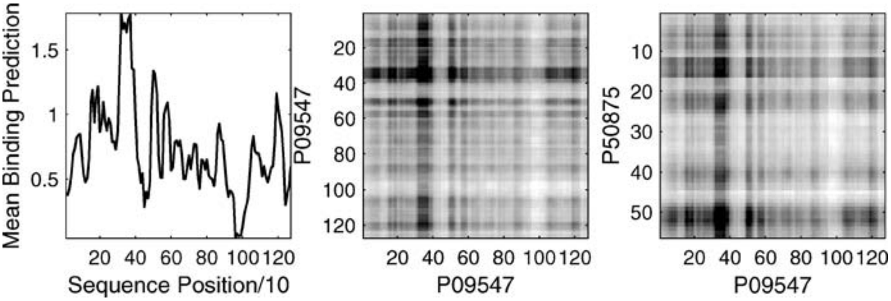


Fig. 4 Domain predictions for P09547. The x-axis of each plot gives the position of a 50-residue window moved ten residues at a time across the full sequence of P09547. The *leftmost plot* (**a**) shows the mean binding activity of the windows along with x-axis with the other 1681 windows considered in the domain prediction example; the *middle plot* (**b**) shows an intensity plot of the binding activities of all pairs of windows in P09547; and the *rightmost plot* (**c**) shows an intensity plot of the binding activities of all pairs of windows in P09547 and P50875 (a known binder). In **b** and **c** *dark* denotes activity and *light* denotes inactivity

157–162, a Poly-Ser domain from 235–240, another Poly-Ser domain from 422–425, a Poly-Ala domain from 454–463, and a Poly-Asn domain from 552–559. Although not perfect, these domains also correlate with our predictions.

While predicting domains is not the focus of our method, these results give a plausible explanation for the success of the product kernel in the prediction of protein–protein interactions. The fact that a model obtained on the full yeast dataset was able to identify domains indicates that the product kernel is able to isolate domain pieces (in terms of length 3 subsequences), which occur often in pairs of interacting proteins. This indicates that the success of the product kernel is due to the fact that it exploits naturally occurring patterns of domain interaction (in terms of sequence) when it makes protein–protein interaction predictions.

7 Predicting β-Strand Interactions

We have also applied the string product kernel to the prediction of β-strand packing interactions [56]. Protein β-sheet topology is determined by these long range β-strand packing interactions. Since these interactions are not necessarily consecutive in sequence, they are difficult to determine using structure-based ab initio simulation type methods [57]. However, sequence order is unimportant when using the string product kernel, making our method ideal for providing initial predictions for β-strand interactions, as well as β-strand ordering within a β-sheet. In addition, our approach complements existing machine learning approaches forβ-sheet topology prediction, including prediction of β-turns [58], the determination of strand register in β-sheets [59–61], and the prediction of edge strands within β-sheets [62, 63].

Our method for predicting β-strand packing interactions is very similar to the method we used for predicting protein–protein interactions in Sect. 6. As a pre-requisite, the approach requires the identification of β-strand sequences within a given protein. These sequences are available from databases such as the Protein Data Bank [64] when structure is available, and can also be predicted using other methods [65, 66]. Given β-strand sequences, we first train a string product kernel SVM to obtain a classifier for the prediction of β-strand interactions. As in the case of our protein–protein interaction model, we used randomly drawn β-strand pairs as negatives.

After obtaining a β-strand interaction SVM model, we made predictions on new β-strand pairs as well as β-strand orderings within β-sheets. The method used for ordering β-strands is outlined in Fig. 5. To order the β-strands, we first isolated β-strand sequences for a given protein, in our case using information already available in the PDB. Next, all possible strand pairs within the sheet were classified using the SVM model. The sign of the SVM prediction provided the class (adjacent or non-adjacent) and the magnitude gave the "strength" of the prediction. We then enumerated all possible orderings of the β-strands within the β-sheet. For each of these orderings we arranged the SVM predictions into a packing interaction matrix. This matrix is a symmetric matrix with one row (and one column) for each β-strand, where the row (and column) order is given by the proposed β-strand ordering. From the packing interaction matrix we derived two scores: a packing likelihood score, which was the average of the super-diagonal elements of the matrix (the elements directly above the diagonal); and a non-packing likelihood score, which was the average of the upper diagonal elements, not including the diagonal or the super-diagonal. The total score for a β-sheet is given by the packing-likelihood minus the non-packing likelihood of opposite sign. As an example, consider the β-sheet ordering of "3, 1, 2". This ordering would be probable if the elements on the super-diagonal ("3, 1" and "1, 2") of the packing interaction matrix were highly positive, and the upper diagonal elements ("3, 2") of the matrix were highly negative. Edge strands are "3" and "2".

Our score can also be used as a confidence metric. This is true because the output of a SVM for a given strand pair can be interpreted as a confidence. In particular, the output of a SVM is the distance of the input from a separating hyper-plane. Therefore, classifications of β-strand pairs that have very small distances from the hyper-plane will be less accurate than those with very large distances. β-strand pairs near the hyper-plane have similarity to both pairing and non-pairing strands. Because the β-sheet score presented here is a summation of the mean magnitudes of the distances of pairing-strands from the hyper-plane (packing likelihood) and distances of non-pairing strands from the hyper-plane (non-packing likelihood), it not only represents a score for a given β-sheet ordering, but also a confidence in that score.

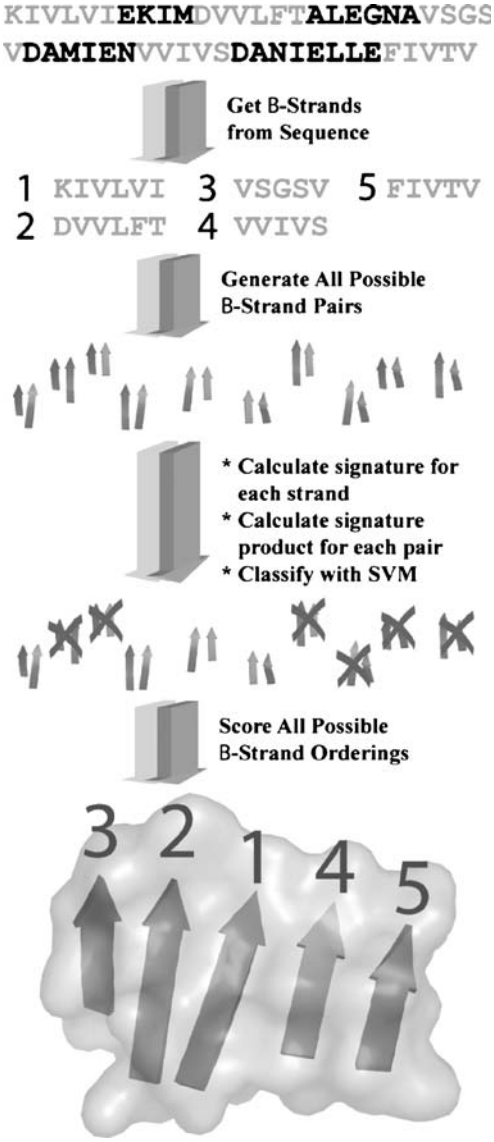

KIVLVIEKIMDVVLFTALEGNAVSGS
VDAMIENVVIVSDANIELLEFIVTV
Get B-Strands from Sequence
1 KIVLVI
2 DVVLFT
3 VSGSV
4 VVIVS
5 FIVTV
Generate All Possible B-Strand Pairs
* Calculate signature for each strand
* Calculate signature product for each pair
* Classify with SVM
Score All Possible B-Strand Orderings
3
2
1
4
5

Fig. 5 Here we show how to use string product kernels to predict the packing interactions within a β-sheet. We start by identifying β-strands from an amino acid sequence. All possible pairs of strands are ranked by likelihood of interaction using an SVM. Finally, the pairs are ordered by maximizing aggregate interaction likelihood

We applied our β-strand interaction and ordering prediction method to β-strands extracted from a 2004 release of the RCSB Protein Data Bank (PDB) [64]. Any protein with over 95% homology to another protein in the dataset was removed, giving 6,682 proteins. For cross-validation, a random ordering of the proteins was divided into 10 test sets each consisting of approximately 10% of the proteins (668). All β-strand sequences, as assigned within the PDB records, were extracted and β-strand pairs were generated for every possible combination of β-strands within any given β-sheet. From these, all duplicate pairs, generated from multiple subunits, were removed. In addition, pairs containing less than 4 residues and greater than 100 residues were removed. Finally, a random selection of non-adjacent strands was removed to balance the number of adjacent and non-adjacent strands for SVM training. The resulting set was composed of 27 196 adjacent strands and 27 196 non-adjacent strands to be used for training, with each cross-validation fold consisting of approximately 90% of these pairs.

For the non-homologous protein dataset, all β-sheets were isolated for validation of strand-ordering accuracy. Any β-sheet containing less than three strands, strands with less than four residues, strands with greater than 100 residues, or strands with unnatural amino acid residues was removed. Duplicate β-sheets were also removed. It was verified that no strands in the test sets were present in the training sets. The cross-validation accuracy of β-strand pairing prediction was performed on all possible pairs in each sheet of the test sets. β-strand ordering based on the test-set predictions was also performed using the same ten-fold cross-validation.

We first assessed the accuracy of our method for predicting which β-strand pairs pack adjacently within a protein. We trained our SVM on ten folds using approximately 24 400 adjacent β-strands and 24 400 non-adjacent β-strands for training for each fold (note that these counts represent approximately 90% of the total dataset). The calculations were done in two steps. We first precomputed the string kernels (not the product kernels) for use by the SVM. We next trained our SVM using the product kernel. The resulting models misclassified an average of 26.6% of the training set. Classification of the test sets (with β-strand pairs extracted from 10% of the PDB in each case) resulted in a ten-fold cross-validation accuracy of 74.0%.

We next tested the ability of our method to predict the ordering of β-strands within a β-sheet. We benchmarked our method by using the strand orderings for all of the β-sheets in the PDB that met our criteria. Choosing the correct β-sheet as the ordering which resulted in the highest score, we achieved an overall ten-fold cross-validation accuracy of only 49.3%. The accuracy for three-stranded sheets was 63.4%, for four-stranded sheets was 56.58%, and for nine-stranded sheets was 11.11%. The decrease in accuracy with sheet size is due simply to the increase in the number of classifications required to compute a score and the increase in the number of possible orderings for any given sheet. For example, a n-stranded sheet requires $1/2n(n-1)$ strand-pairing classifications and has $n!/2$ possible orderings. For a sheet with nine strands, 36 strand-pair classifications are required to calculate the score and the correct sheet must be selected from 181 440 possible orderings.

An overall accuracy of 49.3% is impressive considering the difficulty of the problem and the fact that the baseline accuracy using randomly generated strands was 13.6% (note that this calculation takes into account the high percentage of three-and four-stranded β-sheets). However, the result is suboptimal in terms of the end-user objective of generating orderings for protein structure predictions. To overcome this problem, we can use the confidence measure previously discussed. Due to the fact that the β-sheet score is a measure of the mean distance of the pairing and non-pairing strands from the hyperplane, we can use this score not only to predict the correct ordering, but also as a measure of confidence in that prediction.

It turns out that this confidence metric correlates surprisingly well with prediction accuracy, as shown in Fig. 6 and Table 3. In Table 3, we divide the dataset into four equally sized subsets based on the β-sheet score quartiles (0.11 for the lower quartile, 0.21 for the median quartile, and 0.8 for the upper quartile), and recalculate the accuracies. For the 25% of the predictions with the highest confidence, a 95.7% ordering accuracy was achieved, while for the bottom 25%, the accuracy was 19.92%. A breakdown of the accuracies by the size of the β-sheet is given in Fig. 7.

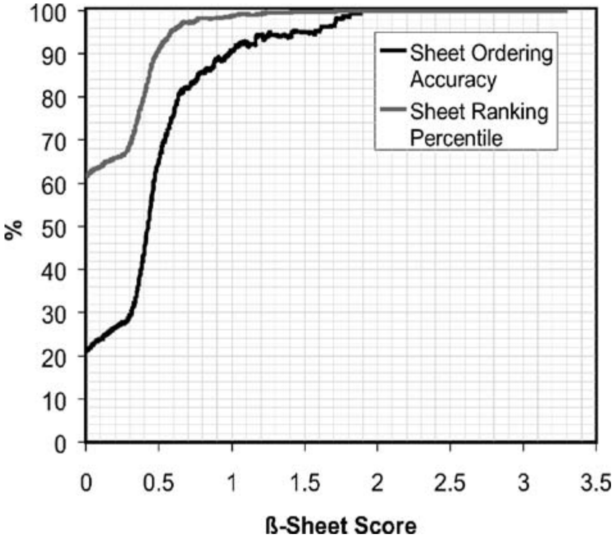


Fig. 6 Moving average plot of the ten-fold cross validation β-sheet ordering accuracy and β-sheet rank percentile for the non-homologous PDB dataset as a function of the β-Sheet Score. The window for the moving average is 0.5. As the magnitude of the β-sheet ordering score increases, so does the confidence that the ordering is correct. For example, if the best score for all possible orderings of a β-sheet is 1.5, there is an expected 95% chance that this is the correct ordering and on average, 99.7% of incorrect orderings will be scored lower (based on the ten-fold cross-validation accuracy of β-sheets with scores between 1.25 and 1.75)

Table 3 Ten-fold cross-validation β-sheet ordering accuracy, sheet rank percentile, and edge strand accuracy for the non-homologous PDB dataset divided by the β-sheet score quartiles. For the 25% of the dataset which was scored with the highest confidence (Score > 0.8), the sheet ordering accuracy is 95.65% for β-sheets of all sizes. On average, 99.53% of β-sheets were scored below the correct ordering and edge strands were predicted with 98.23% accuracy

Score	% of Dataset	Ordering Accuracy	Sheet rank Percentile	Edge strand Accuracy
≥ 0.8	25%	95.65%	99.53%	98.23%
0.21–0.8	25%	60.73%	89.09%	81.78%
0.11–0.21	25%	20.84%	66.14%	62.59%
< 0.11	25%	19.92%	55.23%	59.70%
All	100%	49.30%	77.36%	75.55%

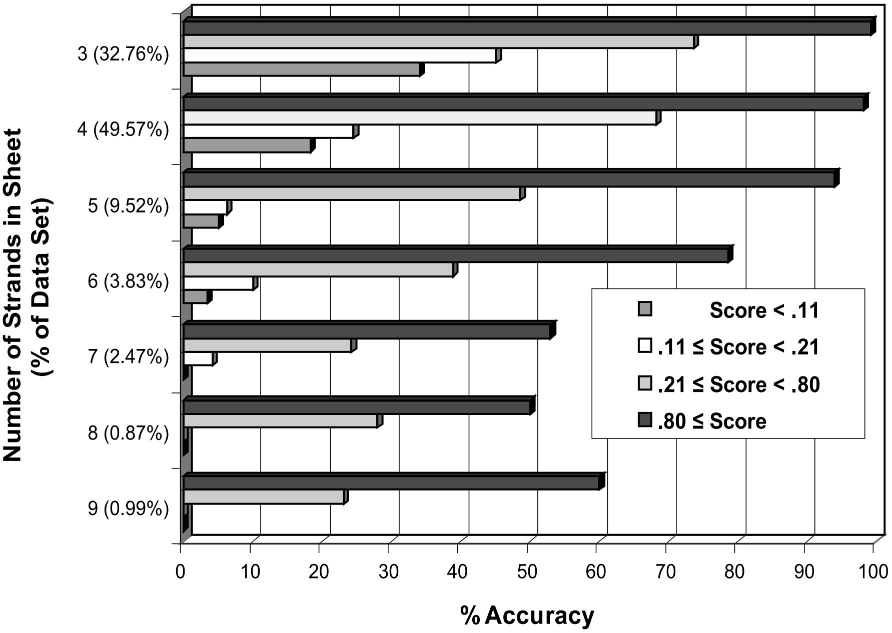


Fig. 7 Ten-fold cross-validation prediction accuracy for β-strand ordering as a function of the number of strands within a sheet for the non-homologous PDB dataset divided by the β-sheet score quartiles. The number in parenthesis represent the percentage of β-sheets within the dataset containing that number of strands

The results show that for about one in four of the β-sheets encountered in the PDB, the β-sheet score is sufficient to have high confidence in the predicted ordering. What if, however, we are interested in a β-sheet with a lower confidence score? Under these circumstances, it may not be appropriate to select only one β-sheet ordering as correct, but rather remove those β-sheet orderings which are highly unlikely. For these cases, a β-sheet ranking percentile is appropriate. We calculate the ranking percentile as the average percentage of β-sheet orderings which score below the correct one. Using this approach, we can remove on average from 55% to 90% of the alternate orderings (Table 3, Fig. 6).

For the prediction of edge strands, an overall accuracy of 75.6% is obtained. As with the β-sheet ordering accuracy, the confidence correlates with the β-sheet score. For the top 25% of the database, the edge strand prediction accuracy is 98.2% and for the bottom 25% it is 59.7% (Table 3). As is to be expected, there is a decrease in prediction accuracy with an increase in the number β-strands within a given sheet (data not shown).

In summary, these results show that the string product kernel can be used not only for protein–protein interaction prediction but also for β-strand packing interaction and ordering prediction. Using the string product SVM, we can predict whether or not two β-strands will pack adjacently within a protein. We used the entire PDB database to validate our method and achieved an overall accuracy of 74.0%. When given the strands within a β-sheet, the model predicted the ordering with an overall accuracy of 49.3%. When we used a simple prediction confidence metric, we were able to determine a priori when the accuracy of a prediction should be high enough to trust as a correct ordering. For test cases where the confidence is low or where the number of β-strands in the sheet is high, the model is not sufficient for predicting β-strand ordering as a starting point for ab initio protein structure prediction methods. Rather, it can be used to throw out potential folds that are predicted to be highly unlikely. On average, 77.36% of the possible β-strand orderings were predicted with a lower score than the correct one.

8 Predicting Protein–Chemical Interactions

Our most recent application of the product kernel is to the prediction of protein–chemical interactions [67], useful in the field of drug discovery. In the case of protein–chemical interactions, we use the string-graph product kernel $k_{s\otimes g}^{l\otimes h}$ as described in Sect. 5. For this application, we used both the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [68] and the DrugBank database [69]. The DrugBank database was also used in a recent study that used kernels capable of processing strings, graphs, and mass spectrometry data for predicting protein-chemical interactions [70].

Using the KEGG database, we compiled a dataset linking drugs with protein targets. This dataset contained 873 drug-target pairs taken from 121 targets and 551 drugs. We experimented with different combinations of string height l and graph height h, using five-fold cross-validation to assess accuracy, as shown in Fig. 8. Five-fold cross-validation is similar to ten-fold cross-validation, described in Sect. 6, except that the dataset is divided into five equal subsets instead of ten.

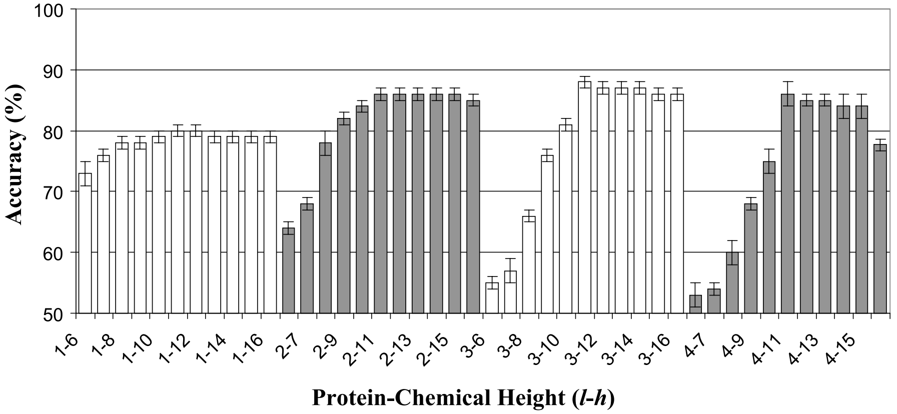


Fig. 8 Five-fold cross-validation accuracies for the KEGG drug-target dataset. The x-axis shows different combinations of string-graph product kernel height l–h

Our results on the KEGG dataset were good, with accuracies near 90% for certain l–h combinations. Encouraged by this success, we used the best SVM to make independent predictions on the DrugBank database. As of December 2006, Drugbank was composed of 1133 drugs and 509 targets, with 1849 drug-protein pairs. Out of the 1133 drugs, 124 had a name and a structure stored in KEGG. Consequently, out of the 1849 pairs, only 298 pairs could potentially be predicted from the KEGG training set. There were only 32 pairs in common between KEGG and DrugBank. Despite this small number, 189 additional interactions not in KEGG predicted by the signature product kernel were found in DrugBank, including 67 interactions between drugs and targets not present in KEGG.

The results obtained from our predictions on the DrugBank dataset are shown in Table 4. Here, we report accuracy, precision, and sensitivity, as described previously in Sect. 6. We divided the dataset into five classes, depending on KEGG and DrugBank intersection. These classes are described in the legend of Table 4 but are roughly arranged in order from least intersection (I) to greatest intersection (V). As might be expected, performance increases from least to greatest intersection.

The accuracy of the product kernel on the DrugBank dataset ranged from 60% to 100%, with an average of 67.5%. While hardly stunning, these results are remarkably good when considering that we are making predictions using

Table 4 Prediction accuracy on DrugBank dataset, as extrapolated from KEGG dataset based SVM model. The results are divided into five classes, depending on intersection between DrugBank and KEGG: class I contains cases where neither the target nor the drug are in the training set; class II contains cases where the drugs are in the training set, but the targets are absent; class III contains cases where the targets are in the training set, but not the drugs; class IV contains cases where both the drug and the target are in the training set, albeit with different partners; and class V is composed of the 32 interactions common between KEGG and DrugBank

Class	Accuracy	Precision	Sensitivity
I	60.0%	59.3%	71.3%
II	57.1%	41.7%	31.3%
III	68.5%	64.8%	81.4%
IV	76.5%	47.4%	81.8%
V	100.0%	100.0%	100.0%
Total	67.5%	64.5%	77.5%

KEGG without any prior knowledge of DrugBank. This is in fact the method we would use if we were actually performing drug discovery.

Our last application applied the string-graph product kernel to the prediction of Enzyme Commission (EC) numbers, also from the KEGG database. Enzymes are organized according to EC number using a hierarchical classification that assigns unique four-field numbers to different enzymatic activities [71]. The first field of an EC number indicates the general class of catalyzed reaction: 1 denotes oxidoreductases, 2 denotes transferases, 3 denotes hydrolases, 4 denotes lyases, 5 denotes isomerases, and 6 denotes ligases. The second and third fields depend on different criteria related to the chemical features of the substrate and the product of the reaction. The fourth field is substrate and product specific. As an example, the tripeptide aminopeptidases have the number "EC 3.4.11.4". Level 1 "EC 3" enzymes are hydrolases. Level 2 "EC 3.4" enzymes are hydrolases that act on peptide bonds. Level 3 "EC 3.4.11" enzymes are hydrolases that cleave off the amino-terminal residue from a polypeptide, and level 4 "EC 3.4.11.4" enzymes are those that cleave off the amino-terminal end from a tripeptide.

For EC number prediction, datasets were generated by first selecting positive examples from the KEGG database. Positive examples were compiled using all reactions and/or proteins in the KEGG database having a specified EC number. Wild cards were allowed in order to generate datasets at various EC levels. As an example, EC level 2 "1.1.*.*" consisted of 409 reactions and 16225 sequences having an EC number starting with 1.1. Positive example sets having less than 50 elements were not processed. If the positive set included more than 500 examples, then excess examples were removed at random. Next, the datasets were completed by taking equal numbers of negative examples at random. Negative examples were protein sequences and/or reactions not present in the positive class.

Using the reactions and sequences with assigned EC numbers in the KEGG database, a set of 855 772 pairs (out of 3905 reactions and 255 304 enzymes) was compiled. The string-graph product kernel was applied for each EC level using various chemical and protein signature heights. Five-fold cross-validation results are shown for class 1.1.1.1 in Fig. 9. Additional cross-validation results and comparisons with alternative methods [72–74] are summarized in Table 5. Table 5 shows that the product kernel outperforms all other techniques in terms of sensitivity (accuracy on positives). Consequently, the product kernel can be used to accurately process a larger number

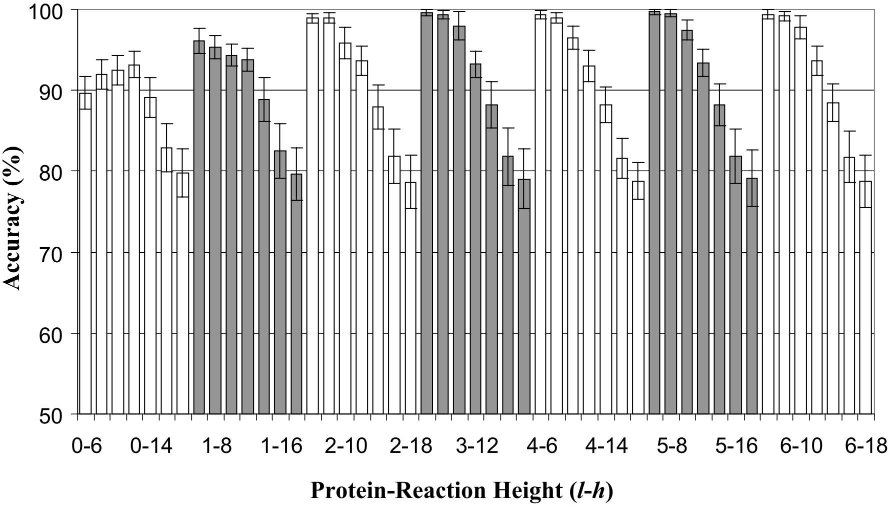


Fig. 9 Five-fold cross-validation accuracies for prediction of the KEGG EC number 1.1.1.1 (alcohol dehydrogenases). The *x*-axis shows different combinations of string-graph product kernel height *l*–*h*

Table 5 Performance of the product kernel when predicting EC numbers. Statistics were computed using five-fold cross-validation and comparisons were made to published results

EC level	Method	Accuracy	Precision	Sensitivity
1	Graph kernels [72]	89.9%	–	40.0%
	Product kernel	88.0%	87.1%	89.6%
2	SVM-Prot [73]	95.2%	97.4%	77.4%
	Product kernel	94.2%	93.6%	93.3%
3	Product kernel	97.9%	97.9%	97.9%
4	Product kernel	99.0%	98.7%	98.7%

of sequences than can be correctly processed with the other techniques listed in Table 5. This observation is important if one is to use any of the techniques listed in Table 5 to complete annotations of newly sequenced genomes.

We also tested the ability of the string-graph product kernel to predict unknown enzyme-metabolite interactions. For this exercise, we compiled a list of enzymes and reactions corresponding to EC numbers accepted in September 2006 by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). This list was removed from the KEGG database and used as a test set to assess the ability of the product kernel to predict when an enzyme will catalyze a reaction when neither the reaction nor the enzyme is characterized by an EC number. A training set from the reduced KEGG database was compiled. Results are given in Table 6 for each removed EC class. These results show that it is possible to predict (with accuracies above 80%) whether or not a given enzyme will catalyze a given reaction, even when the enzyme-reaction pair is not present in the training set.

Table 6 Prediction statistics for reaction-enzyme interactions not classified by an EC number. All reactions and enzymes corresponding to the EC numbers listed in the first column were removed from the KEGG database and stored as test sets. A training set of 3000 examples was constructed from the remaining KEGG database. The training set consisted of 500 examples for each of the six top levels of the EC nomenclature. SVMs were trained with the signature product kernel using height three for reactions, and height ten for proteins

EC class	# Positive pairs	Accuracy	Precision	Sensitivity
1.1.1.290	59	88.7%	82.6%	98.3%
1.13.11.52	13	76.9%	76.9%	76.9%
1.13.11.53	11	86.4%	83.3%	90.9%
1.2.1.71	55	87.5%	82.3%	95.5%
1.2.1.72	46	88.0%	80.8%	100.0%
1.8.4.11	390	79.5%	99.8%	59.1%
2.6.1.81	21	81.0%	72.7%	100.0%
3.1.3.77	160	89.4%	82.5%	100.0%
3.3.2.9	17	84.3%	82.3%	88.2%
3.5.1.96	49	88.6%	81.6%	100.0%
3.5.3.23	51	90.2%	83.7%	100.0%
4.2.1.109	12	87.5%	80.0%	100.0%
Average	74	85.7%	82.4%	92.4%

The traditional method for constructing a metabolic map of a newly sequenced organism is to assign EC numbers to its proteins. Our results indicate that this task may be better performed when both sequence and chemical information are taken into account. Yet many proteins remain un-annotated not only because their sequences have not been mapped to an already classified enzyme, but also because the reactions catalyzed by the proteins have not been characterized in the EC nomenclature. EC number assignment requires published evidence and full characterization of the enzymatic reaction. For this reason, many reactions, although occurring in various pathways, do not have an assigned EC number. However, using the product kernel, we can predict when an enzyme will catalyze a metabolic reaction, even in the absence of any EC nomenclature information.

9 Conclusions

Protein interactions are a primary mechanism in many cellular processes. Thus there are a wide variety of experimental methods designed to gather information related to protein interactions [1]. Ideally, we would use this experimental data in conjunction with structural and dynamic information to make computational predictions of protein-interaction maps, protein–drug interactions, and protein-DNA regulatory interactions. However, structural data is not widely available, and structure computations are difficult and time consuming [14]. On the other hand, sequence information is widely available, and statistical methods are easily computed on large datasets. It therefore makes sense to develop statistical, sequence-based methods for predicting protein–protein interactions [13].

Our method is one such statistical, sequence-based method. Our method generally compares well with competing methods and has the additional advantage of being fairly general. Once the concept of the product kernel has been introduced, we can mix and match kernels to operate on any type of data. We have demonstrated the use of the product kernel for predicting protein–protein interactions, β-strand packing interactions, and protein–drug interactions. While these applications are actually quite different, they are treated nearly identically in the framework of the product kernel method. So far we have used the product kernel framework with amino acid string and chemical graph kernels, but we could just as easily use DNA sequences and structural information to extend our method to other data. We could, for example, use string kernels to predict protein–DNA interactions. Although more ambitious, we could also develop structure-based kernels to make more accurate predictions of protein–protein or protein–drug interactions.

This last example suggests the most obvious avenue for future research using our method, and perhaps a general avenue for improving the predictive ability of statistical methods for protein–protein interaction data. While statistical methods for prediction of protein–protein interactions are fast and fairly accurate, they are, after all, still statistical. Larger datasets can make the predictions more accurate, but eventually, structural or dynamic information must be incorporated. While the ultimate goal of computational prediction of protein–protein interactions may be very accurate methods that use structural and dynamic information from first principles alone, an intermediate goal is the combination of statistical and structural methods. From this point of view, an SVM kernel or product kernel method may be the easiest way to achieve this goal in the near future.

Acknowledgements This work was funded in part by the U.S. Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org). This work was also funded by Sandia Laboratory Directed Research and Development. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part I Experimental techniques and databases. PLoS Comput Biol 3(2):e42
2. Yan Y, Marriott G (2003) Analysis of protein interactions using fluorescence technologies. Curr Opin Chem Biol 7(4):635–640
3. Karlsson R (2004) SPR for molecular interaction analysis: a review of emerging application areas. J Mol Recognit 17(2):151–161
4. Yang Y, Wang H, Erie DA (2003) Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy. Methods 29(1):175–187
5. Baumeister W, Grimm R, Walz J (1999) Electron tomography of molecules and cells. Trends Cell Biol 9(1):81–85
6. Ito T et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98(7):4569–4574
7. Uetz P et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403(6770):623–627
8. Rigaut G et al. (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17(9):1030–1032
9. Eisen MB et al. (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25):14863–14868
10. Jones RB et al. (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439(7073):168–174
11. Ye P et al. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1:2005–0026
12. Smith GP (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 228(4705):1315–1717
13. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part II Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3(3):e43
14. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7(2):188–197
15. Smith GR, Sternberg MJ (2002) Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 12(1):28–35

16. Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 99(8):5896–5901
17. de Rinaldis M et al. (1998) Three-dimensional profiles: a new tool to identify protein surface similarities. J Mol Biol 284(3):1211–1221
18. Sheinerman FB, Al-Lazikani B, Honig B (2003) Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains. J Mol Biol 334(3):823–841
19. Dandekar T et al. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci 23(8):324–328
20. Overbeek R et al. (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96(5):2896–2901
21. Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng 14(8):609–614
22. Pellegrini M et al. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96(7):4285–4288
23. Enright AJ et al. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402(6757):86–90
24. Goh CS et al. (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299(1):283–293
25. Walhout AJ et al. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science 287(5450):116–122
26. Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein–protein interactions. Bioinformatics 21(1):i38–46
27. Martin S, Roe D, Faulon JL (2005) Predicting protein–protein interactions using signature products. Bioinformatics 21(1):218–226
28. Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol 311(3):681–692
29. Deng M et al. (2002) Inferring domain-domain interactions from protein–protein interactions. Genome Res 12(9):1540–1548
30. Jansen R et al. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302(5644):449–453
31. Apweiler R et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29(1):37–40
32. Shawe-Taylor J, Cristianini N (2004) Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge
33. Vapnik V (1998) Statistical Learning Theory. Wiley, New York
34. Shawe-Taylor J, Cristianini N (2000) Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press, Cambridge
35. Smola A, Scholkopf B (1998) A tutorial on support vector regression. NeuroCOLT NC-TR-98–030, Royal Holloway College, University of London, UK
36. Ben-Hur A et al. (2001) Support vector clustering. J Mach Learn Res 2:125–137
37. Ham J et al. (2004) A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the International Conference on Machine Learning (ICML'04). Banff, Canada
38. Weinberger KQ, Saul LK (2006) An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In: Proceedings of the National Conference on Artificial Intelligence (AAAI'06). Boston, MA
39. Bennet K, Campbell C (2000) Support vector machines: hype or hallelujah? SIGKDD Explorations 2(1):1–13
40. Burges C (1998) A tutorial on support vector machines for pattern recogntion. Data Mining Knowledge Discov 2:121–167

41. Xenarios I et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30(1):303–305
42. Alfarano C et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. 33(Database issue), p D418–D424
43. Guldener U et al. (2006) MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. 34(Database issue), p D436–D441
44. Leslie C, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput, pp 564–575
45. Leslie C, Kuang R (2004) Fast string kernels using inexact matching for protein sequences. J Mach Learn Res 5:1435–1455
46. Mahe P et al. (2005) Graph kernels for molecular structure-activity relationship analysis with support vector machines. J Chem Inf Model 45(3):939–951
47. Ralaivola L et al. (2005) Graph kernels for chemical informatics. Neural Netw 18(7):1093–1110
48. Swamidass SJ et al. (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics 21(1):i359–i368
49. Faulon JL, Visco DP Jr, Pophale RS (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci 43(2):707–720
50. Faulon JL, Churchwell CJ, Visco DP Jr (2003) The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. J Chem Inf Comput Sci 43(2):721–734
51. Faulon JL, Collins MJ, Carr RD (2004) The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. J Chem Inf Comput Sci 44(1):427–436
52. Spivak M (1965) Calculus on Manifolds. Perseus Books Publishing
53. Tong AH et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295(5553):321–324
54. Rain JC et al. (2001) The protein–protein interaction map of *Helicobacter pylori*. Nature 409(6817):211–215
55. Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17(4):455–460
56. Brown WM et al. (2006) Prediction of beta-strand packing interactions using the signature product. J Mol Model 12(2):355–361
57. Orengo CA et al. (1999) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins Suppl 3:149–170
58. Przybylski D, Rost B (2002) Alignments grow, secondary structure prediction improves. Proteins 46(1):197–205
59. Hutchinson EG et al. (1998) Determinants of strand register in antiparallel beta-sheets of proteins. Protein Sci 7(10):2287–2300
60. Steward RE, Thornton JM (2002) Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. Proteins 48(1):178–191
61. Zaremba SM, Gregoret LM (1999) Context-dependence of amino acid residue pairing in antiparallel beta-sheets. J Mol Biol 291(1):463–479
62. King RD et al. (1994) On the use of machine learning to identify topological rules in the packing of beta-strands. Protein Eng 7(10):1295–1303
63. Siepen JA, Radford SE, Westhead DR (2003) Beta edge strands in protein structure prediction and aggregation. Protein Sci 12(9):2348–2359
64. Berman HM et al. (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242

65. Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134(2–3):204–218
66. Simossis VA, Heringa J (2004) Integrating protein secondary structure prediction and multiple sequence alignment. Curr Protein Pept Sci 5(3):249–266
67. Faulon J-L, Misra M, Martin S, Sale K (2007) Genome scale enzyme-metabolite and drug-target interaction prediction using the signature molecular descriptor. Bioinformatics
68. Kanehisa M et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34(Database issue):D354–D357
69. Wishart DS et al. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34(Database issue):D668–D672
70. Nagamine N, Sakakibara Y (2007) Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 23(5):2004–2012
71. Webb EC (1992) Enzyme Nomenclature Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Academic Press, San Diego
72. Borgwardt KM et al. (2005) Protein function prediction via graph kernels. Bioinformatics 21(1):i47–i56
73. Cai CZ et al. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31(12):3692–3697
74. Kunik V et al. (2005) Motif extraction and protein classification. Proc IEEE Comput Syst Bioinform Conf, pp 80–85

Adv Biochem Engin/Biotechnol (2008) 110: 247–267
DOI 10.1007/10_2007_089
© Springer-Verlag Berlin Heidelberg
Published online: 18 January 2008

Computational Methods For Predicting Protein–Protein Interactions

Sylvain Pitre[1] · Md Alamgir[2] · James R. Green[3] · Michel Dumontier[1,2] ·
Frank Dehne[1] · Ashkan Golshani[2] (✉)

[1]School of Computer Science, Carleton University, 5304 Herzberg Building,
1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada

[2]Department of Biology and Ottawa Institute of Systems Biology, Carleton University,
209 Nesbitt Building, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada
ashkan_golshani@carleton.ca

[3]Department of Systems and Computer Engineering, Carleton University,
4456 Mackenzie Building, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada

Abstract Protein–protein interactions (PPIs) play a critical role in many cellular functions. A number of experimental techniques have been applied to discover PPIs; however, these techniques are expensive in terms of time, money, and expertise. There are also large discrepancies between the PPI data collected by the same or different techniques in the same organism. We therefore turn to computational techniques for the prediction of PPIs. Computational techniques have been applied to the collection, indexing, validation, analysis, and extrapolation of PPI data. This chapter will focus on computational prediction of PPI, reviewing a number of techniques including PIPE, developed in our own laboratory. For comparison, the conventional large-scale approaches to predict PPIs are also briefly discussed. The chapter concludes with a discussion of the limitations of both experimental and computational methods of determining PPIs.

Keywords Automated tools · Computational techniques · Interactome · PIPE ·
Protein–protein interaction

Abbreviations

AD	Activation domain
BIND	Biomolecular interaction network database
CAPRI	Critical assessment of predicted interactions
DBD	DNA binding domain
DBID	Database of interacting domains
DIP	Database of interacting proteins
InterPreTS	Interaction prediction through tertiary structure
MINT	Molecular interactions database
MIPS	Munich information center for protein sequences
PDB	Protein data bank
PID	Potentially interacting domain pairs
PIPE	Protein–protein interaction prediction engine
PPI	Protein–protein interaction
PRISM	Protein interactions by structural matching
SVM	Support vector machine
TAP	Tandem affinity purification
Y2H	Yeast two-hybrid

1 Introduction

An overwhelming number of biological processes are mediated through the action of proteins. In many cases, these proteins carry out their functions by interacting with each other in either stable or transient protein complexes. The nature and increasing complexity of these interactions is thought to be responsible for the overall biological complexity in higher organisms. Therefore, it is believed that humans, for example, are more sophisticated than the nematode *C. elegans*, not only because we possess marginally greater number of genes, but largely because human proteins form more intricate networks [1, 2]. Recent advances in the field of genomics and proteomics have lead to the discovery and characterization of some of these networks [3, 4]. An organism may have numerous interactomes representing different tissue types, biological states, etc. The complete elucidation of all interaction networks found in an organism will have significant implications for science [5]. For example, the cellular roles and molecular functions for previously ill-characterized proteins may be inferred from the networks of interactions that they participate in. Moreover, the conservation of protein interactomes across organisms will also provide insight into their evolutionary relationships. Practically, knowledge of interaction networks will provide insight into their dependencies and lead to enhanced approaches for drug discovery. For these reasons, the elucidation of protein–protein interactions (PPIs) especially within the context of an interaction network is an important goal in biological research [6, 7].

Until recently, PPIs were determined by carrying out experiments that were specifically designed to identify a small number of specifically targeted interactions. However, the development of novel genomic techniques allows for high-throughput experiments, which can now be carried out to exhaustively probe all possible interactions within an entire genome. *Saccharomyces cerevisiae*, also known as baker's yeast, has emerged as the model organism of choice for functional proteomics due to the elucidation of its genomic sequence in 1996 [8]. Since then, whole PPI maps have been determined using various methods including yeast two-hybrid [9, 10], affinity purification/mass spectrometric identification methods such as TAP-tagging [11, 12], and protein chips [13, 14]. Indirect large-scale approaches such as synthetic lethal analysis [15] and correlated mRNA expression profile [16] have also been used to investigate PPIs.

However, these methods are not without shortcomings. Not only are they labor- and time-intensive, they also have a high cost associated with them. Another important disadvantage is the poor accuracy of the data generated. Significant discrepancies between results of small-scale high-confidence experiments and high-throughput studies have been reported [6, 17]. Interstudy discrepancy is even higher when comparing data generated from different large-scale studies [6, 17]. In addition, the PPI data obtained from biological experiments often include many false positives, which may connect proteins that are not necessarily related. Therefore, it is often necessary to confirm the interactions by other methods. Consequently, there is a growing need for the development of computational tools that are capable of effectively identifying PPIs as well as interpreting and validating the experimentally derived data.

A wide range of computational methods have been developed to build, study, and exploit protein interactomes (reviewed in [6, 17]). First, computational methods have been developed to construct interaction databases within which experimentally determined data is collected and annotated. Automated data mining techniques can then be applied to extract relevant information about potential interactions from the vast amount of PPI information in these databases. As mentioned earlier, a number of experimental techniques have been used to determine large-scale protein interaction maps. Although the significant inconsistencies between interaction maps of the same organism obtained using different techniques can be somewhat justified [6], computational methods have been successfully applied to assess, validate, and carefully scrutinize these experimentally determined protein interactomes. Based on the assumption that physically interacting proteins have a high probability of also being functionally related, a number of computational tools have been developed to exploit protein interaction networks in order to predict functional features of the proteins. Lastly, computational methods can also be used to predict novel PPIs by learning from known interactions [6, 17].

It is the objective of this chapter to provide an overview of these computational methods, with the main focus being on computational tools for the prediction of novel interactions. We also highlight the specific limitations for each of the tools discussed, as well as the systematic shortcomings common to most computational tools. A novel tool recently developed by our bioinformatics group (protein–protein interaction prediction engine, PIPE) is also discussed. For comparison, the advantages and limitations of traditional "wet lab" experimental approaches are also summarized.

Finally, due to space constraints, it is impossible to include all tools relevant to the study of PPIs and the authors apologize in advance to all those researchers whose work has not been cited here.

2
Traditional Methods of Determining PPIs

The yeast two-hybrid (Y2H) method was one of the first methods to be applied to the detection of PPIs. Two protein domains are required in the Y2H assay that have specific functions: (i) a DNA binding domain (DBD) that helps bind to DNA, and (ii) an activation domain (AD) responsible for activating transcription of DNA. Both domains are required for the transcription of a reporter gene [10]. The Y2H assay relies on the fusion of DBD to a protein of interest (X) at its N-terminus and the fusion of AD to another protein of interest (Y) at the C-terminus, which forms DBD-X (bait) and AD-Y (prey). If the bait and prey hybrids interact with each other, the transcription of the reporter gene will be induced and, in this way, the interaction can be detected [18].

Y2H analysis allows the direct recognition of PPI between protein pairs. However, a large number of false positive interactions may arise, while a number of true interactions will be missed (i.e., false negatives). A false positive interaction can occur by activation of RNA polymerase by a bait protein, by the binding of the prey AD-Y protein with upstream activating sequences (UAS), by non-specific binding of bait and prey proteins with some endogenous proteins, or by the binding of "sticky" prey proteins with bait proteins [19]. On the other hand, many true interactions may not be detected using Y2H assay, leading to false negative results. In a Y2H assay, the interacting proteins must be localized to the nucleus; since membrane proteins are typically less likely to be present in the nucleus they are unavailable to activate reporter genes, and hence are excluded. Proteins that require post-translational modifications to carry out functions are also unlikely to behave or interact normally in a Y2H experiment. Furthermore, if the proteins are not in their natural physiological environment, they may not be folded properly to interact [20]. During the last decade, Y2H has been improved by designing new yeast strains containing multiple reporter genes and new expression vectors to facilitate the transformation of yeast cells with hybrid proteins [21].

Tandem affinity purification (TAP) tagging was developed to study PPIs under the native conditions of the cell [22]. Gavin et al. first attempted the TAP-tagging method in a high-throughput manner to analyze the yeast interactome [23]. This method is based on the double tagging of the protein of interest on its chromosomal locus, followed by a two-step purification procedure using Staphylococcus protein A and calmodulin beads separated by a tobacco etch virus (TEV) protease cleavage site. First, a target protein open reading frame (ORF) is fused with the DNA sequences encoding the TAP tag and is expressed in yeast where it can form native complexes with other proteins. The tagged protein along with its associated proteins/complexes is then extracted from the cell lysate. The fused protein and the associated complexes are then purified via a two-step affinity purification procedure. Proteins that remain associated with the target protein can then be analyzed and identified through SDS-PAGE [24] followed by mass spectrometry analysis [22], thereby identifying the PPI partner proteins of the original protein of interest.

An important advantage of TAP-tagging is its ability to identify a wide variety of protein complexes and to test the activity of monomeric or multimeric protein complexes that exist in vivo. Compared to Y2H, TAP-tagging obtains interaction information from a more natural environment since the physiological conditions are more realistic than those created by Y2H, including factors like post-translational modifications and pH requirements. However, the TAP tag may interfere with the formation of some protein complexes (as shown by [23]) by low expression of fusion proteins [25], which can affect the ability of a protein to interact with other proteins or may cause a mutant phenotype [26]. These problems may be minimized by using other complementary techniques that can increase the reproducibility of any large-scale approaches.

The large quantity of experimental PPI data being generated on a continual basis necessitates the construction of computer-readable biological databases in order to organize and effectively disseminate this data. A number of such databases exist (Table 1) and are growing at exponential rates. The biomolecular interaction network database (BIND), for example, is built on an extensible specification system that permits detailed description of the manner in which the PPI data was derived experimentally, often including links directly to the supporting evidence from the literature [27]. The database of interacting proteins (DIP) is another database of experimentally determined protein–protein binary interactions [28]. DIP serves as an access point to a number of other related databases such as LiveDIP, which provides information on the functional aspects of protein complexes as well as links out to other databases such as the database of ligand–receptor partners (DLRP). The general repository of interaction datasets (BioGRID) is a database that contains protein and genetic interactions among proteins from 13 species [29]. Interactions are regularly added through exhaustive curation of the primary literature. Interaction data is extracted from other

Table 1 Databases of experimental protein–protein interactions

Database	URL	Organism	Refs.
BIND	http://bond.unleashedinformatics.com/	Any	[27]
DIP	http://dip.doe-mbi.ucla.edu	Any	[28]
BioGRID	http://www.thebiogrid.org/	Any	[29]
MIPS	http://mips.gsf.de	Yeast	[30]
MINT	http://cbm.bio.uniroma2.it/mint	Any	[31]

databases including BIND and MIPS (Munich information center for protein sequences) [30], as well as directly from large-scale experiments [31]. The molecular interaction database (MINT) is another database of experimentally derived PPI data extracted from the literature, with the added feature of providing the weight of evidence for each interaction [31].

3 Computational Prediction of PPIs

Computational methods provide a complementary approach to detecting PPIs. Indeed, the wide availability of experimental data has spurned the development of numerous computational methods over the past few years. In general, all computational approaches to PPI prediction attempt to leverage knowledge of experimentally determined previously known interactions in order to predict new PPIs. These methods enable one to discover novel putative interactions and often provide information for designing new experiments for specific protein sets.

These approaches can be classified into five general categories: methods based on genomic information, evolutionary relationships, three dimensional protein structure, protein domains, and primary protein structure. Specific approaches that fall within these categories are listed in Table 2 and are discussed below. Figure 1A–E presents the idea behind the five categories of methods.

3.1 Genomic Methods

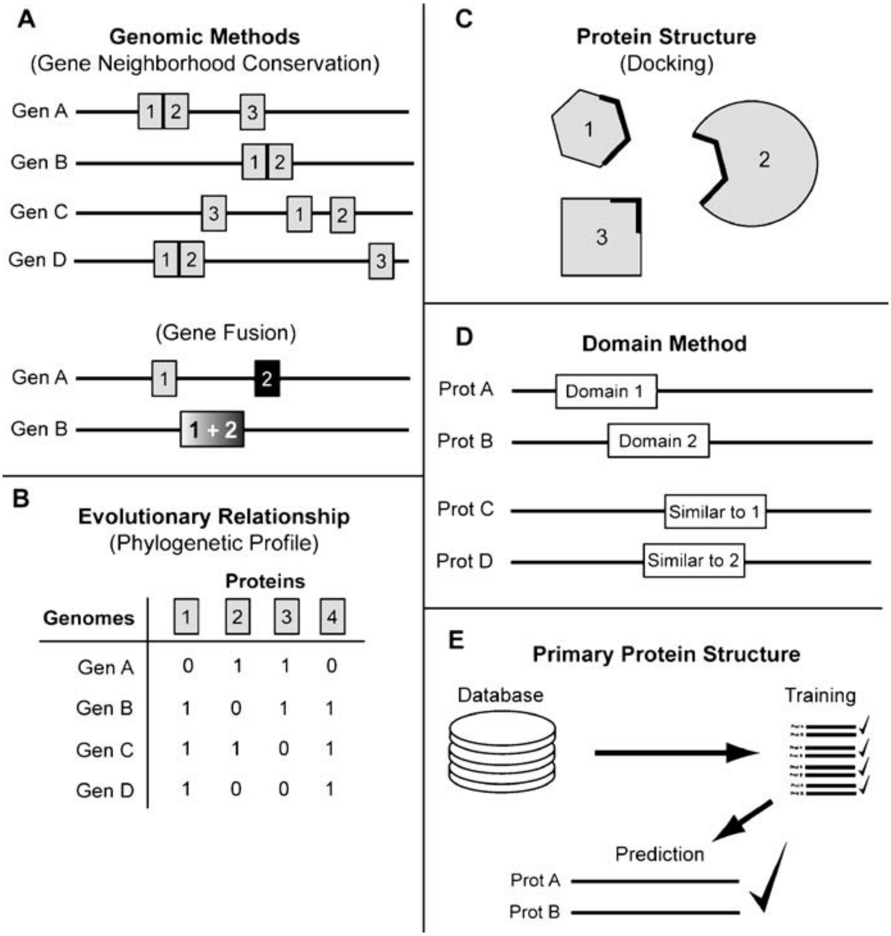
Genomic methods for interaction prediction take advantage of the availability of information obtained by complete genome sequencing. Completely sequenced genomes provide knowledge of which genes are present and how they are organized (gene order). The conservation of gene order across species yields information about the evolution of the genome, and hints at which genes may be functionally correlated. Most computational methods

Table 2 Computational methods for the prediction of protein–protein interactions

Method	Description
Whole genome	Conservation of gene order across genomes [32] Comparison of protein pairs in one genome to its fused single protein product homolog in another genome [33, 34]
Evolutionary relationship	Correlated evolution of functionally related proteins [35] Tree kernel-based computational system to assess similarities between phylogenetic profiles [36, 37]
3D protein structure	Assess fit of two interacting partners on a predetermined complex of known 3D structure; Web-based version InterPreTS [38, 39] Multimeric threading algorithm MULTIPROSPECTOR to recognize partners in protein interactions [40] CAPRI is a community-wide experiment focusing on the performance of protein–protein docking procedures [41] PRISM: protein interactions by structural matching [42]
Domain	Combination of similarity between sequence patches involved in interactions and between domains of interacting partners [43] Maximum likelihood estimation method to determine probability of interactions between evolutionarily conserved protein domains in the Pfam protein domain database [44] Prediction of interaction probability of proteins; ranking system for probability of interactions between multiple protein pairs [45, 46] Database of potentially interacting domain (PID) pairs using a DIP database and InterPro; PID matrix score as a reliability index for accurate analysis of interaction networks [47]
Primary protein structure	Protein interactions mediated through specific short polypeptide sequences [48] Automatic recognition of correlated patterns of sequences and substructure by support vector machine; also uses associated physiochemical parameters [49] Combination of sequence information, experimental data analysis and subsequence paring to generate a "signature product" that is implemented with support vector machine [50] Kernel methods for predicting protein–protein interactions [51] PIPE: protein–protein interaction prediction engine that uses primary protein structure data from MIPS and DIP databases [52]

that use genomic information do not rely solely on the sequence similarity between homologous genes (or their products) [53, 54], but rather assess functional links between pairs or clusters of co-located genes.

Evidence for the evolutionary conservation of gene order can be obtained by systematic comparison of completely sequenced genomes. Dandekhar et al. [32] compared nine bacterial and archaeal genomes and applied



	Proteins			
Genomes	1	2	3	4
Gen A	0	1	1	0
Gen B	1	0	1	1
Gen C	1	1	0	1
Gen D	1	0	0	1

Fig. 1 The five categories of computation PPI methods: **A** Genes of proteins that are close in different genomes are predicted to interact. Proteins 1 and 2 are predicted to interact since the physical locations of their genes are in close proximity to each other in the genomes *A*, *B* and *D*. Two proteins are also predicted to interact if they combine (fuse) to form one protein in another organism. **B** Protein pairs with similar phylogenetic profiles in different genomes are predicted to interact. Proteins 1 and 4 are predicted to interact since they share the same phylogenetic profile. **C** Using the protein structures, docking methods will predict the best compatibility of their interacting regions. Proteins 1 and 2 are predicted to interact since they have the best fit. **D** If two proteins *A* and *B* known to interact share a pair of conserved domains and two other proteins *C* and *D* also share those same conserved domains, *C* and *D* are predicted to interact. **E** Using the primary protein structure and a database containing some other information (such as known interactions), it is possible to train an algorithm to predict protein–protein interactions

a method based on co-localization to determine conserved gene pairs even within relatively low conservation of gene-order. They found that proteins encoded by conserved gene pairs also appeared to interact physically. Physical interactions between encoded proteins have been demonstrated for at least 75% of the conserved gene pairs. A further 20% of the conserved pairs were predicted to encode proteins that interact physically [32]. While promising, the approach fails to identify interactions between products of distantly located genes. Moreover, false predictions are generated because the proximity constraint is not sufficient to determine physical interaction. Finally, this approach may not be applicable to eukaryotes, because the co-regulation of genes is not imposed at the genome structure level [33].

The co-localization of genes encoding interacting or functionally related gene products can be taken a step further. Pairs of interacting or functionally related proteins sometimes have homologs in another genome in which they are fused into a single protein [55]. For example, the Gyr A and Gyr B subunits of *Escherichia coli* DNA gyrase are fused as a single protein in yeast topoisomerase II [33]. Thus, the sequence similarities between Gyr A and Gyr B and different segments of the topoisomerase II might be used to predict that Gyr A and Gyr B may interact in *E. coli* [33]. Marcotte et al. developed a computational method to search for such fusion events within multiple genomes. In their study, they uncovered 45 502 such putative PPIs in yeast. Some proteins that were found to be linked to several other proteins also appeared to interact functionally in pathways. Many of these putative interactions were also confirmed experimentally, as documented in the DIP database.

Similarly, Enright et al. identified 215 genes involved in 64 unique fusion events across *E. coli*, *Haemophilus infuenzae* and *Methanococcus jannaschii* [34]. This gene-fusion analysis approach has since been incorporated into a computational algorithm for the prediction of PPIs and protein function [55].

3.2
Evolutionary Relationship

Evolutionary relationships between two proteins can also be used to infer a physical and functional relationship. The phylogenetic profile of a protein describes the presence of homologs across a series of organisms. Proteins that exhibit similar profiles may be functionally linked. For instance, proteins that make up multimeric structural complexes or that participate in a given biochemical pathway typically exhibit similar phylogenetic profiles. Pellegrini et al. applied phylogenetic profiling to predict the function of previously uncharacterized proteins [35]. The comparison of profiles is further enhanced by including evolutionary information. Vert showed that the accuracy of function prediction using a support vector machine (SVM) is improved with the use of evolutionarily enhanced phylogenetic profiles [36]. A comparative genome phylogenetic analysis approach has also lead to prediction of hundreds of pairs of interactions in *E. coli*, and thousands in yeast [37].

3.3 Protein Structure

As the number of experimentally solved protein structures continues to increase, three-dimensional (3D) structure information has become increasingly applied to the prediction of physical binding [40, 56]. By considering homologous proteins, it has been shown that close homologs (>30% sequence identity) physically interact in the same or similar way [56]. Aloy and Russell describe such a 3D-based method to model putative interactions [56]. The method assesses the fit of two potential interacting partners on a complex of known 3D structure and infers molecular details of how the interaction is likely to occur. In general, it has been shown that residues located at the interface tend to be structurally conserved [38]. Residues that make atomic contacts in a crystallographic complex are analyzed. An interaction is conserved as long as the contacting resides is also conserved. Homologs of both interacting proteins are then examined to see whether these interactions are preserved. All possible pairs between two protein families can then be modeled and the most likely interactions determined. The method also provides a means of assessing the compatibility of a proposed PPI within such a complex, as well as for ranking interacting pairs in studies that involve protein families that show different interaction specificities. The method can be used to model a complex based on the known structure of a similar template complex, and to correctly predict interactions within several systems [56]. Aloy et al. successfully demonstrated how 3D structures can be used to query entire interaction networks so as to validate and infer the molecular details of interactions that have been predicted using other methods. InterPreTS (interaction prediction through tertiary structure) is a web-based version of the above method [39]. Homologs of a test pair of protein sequence are identified from the database of interacting domains (DBID) of known 3D complex structures. The sequences are then scored for how well they preserve sites of contacts at the interaction interface [39]. InterPreTS allows one to visualize the molecular details of any predicted interaction. Combining domain structural similarities and conserved sequence patches among interacting proteins has also led to improved methods for interaction prediction [43].

Lu et al. report a multimeric threading approach to identifying interaction partners and to assign quaternary structures of proteins found in the yeast DIP database [40]. This multimeric threading algorithm, MULTIPROSPECTOR, is able to recognize partners involved in protein interactions and correctly predict a significant number of interacting yeast proteins pairs that have already been identified in the DIP database. The method correctly recognized and assigned 36 of 40 homodimers, 15 of 15 heterodimers, and 65 of 69 monomers that were scanned against a protein library of 2478 structures obtained from the protein data bank (PDB) [57].

The reported prediction accuracy of current methods often varies substantially, and recent efforts have been made to address this issue. CAPRI (critical assessment of predicted interactions) is a community-wide experiment that aims to fairly evaluate the state of the art in protein–protein docking procedures by making predictions on a set of interacting proteins for which the solution has not yet been published [41]. Models are compared to high quality crystallographic interaction data by independent CAPRI assessors. During the course of these experiments, it was found that models exhibiting a high degree of native intermolecular contacts were generally good indicators of true PPIs.

PRISM (protein interactions by structural matching) searches a dataset of protein structures for potential interaction partners by comparing protein structure pairs with a dataset of interfaces [42]. This interface dataset is a structurally and evolutionarily representative subset of biological and crystal interactions present in the PDB. The algorithm calculates the similarity between interfaces by first obtaining structural surface alignments. This measures structural similarity of a target structure to a binding site. If the surfaces of two target proteins contain similar regions to complementary partner chains, it may be inferred that those target proteins interact through similar regions. The PRISM web server allows users to explore protein interfaces as well as predictions of PPIs. One can search a variety of stored interfaces categorized by functional clusters or structural similarity. For example, users can search for proteins involved in cell metabolism, while restricting the results to interfaces of certain sizes. PRISM's interactive visualization tool shows the 3D model along with the desired features. One can also submit protein structures (in PDB format) for interaction prediction. Note that this method is only applicable to proteins with known structure.

3.4 Domain-Based

There are a number of computational techniques that are based solely on the conservation of protein domains. For example, a method developed by Deng et al., employs maximum likelihood estimation to infer interacting domains that are consistent with the observed PPIs [44]. Using evolutionarily conserved domains defined in the Pfam (protein families) protein domain database [58], the probabilities of interactions between every pair of domains are estimated. These inferred domain–domain interactions are subsequently used to predict interactions between proteins. Han et al. provide a similar computational tool that not only predicts the PPIs, but also provides the interaction probability of input proteins and ranks the possibilities of interaction between multiple protein pairs [45, 46].

Another prediction algorithm called PreSPI (prediction system for protein interaction), based on conserved domain–domain interactions, was also described by Han et al. [45]. Here a domain combination-based PPI probabilistic framework is used to interpret PPIs as the result of interactions of multiple domain pairs or of groups. This tool is able to predict the interaction probability of proteins and also provides an interaction possibility ranking method for multiple protein pairs that can be used to determine which protein pairs are most likely to interact with each other in multiple protein pairs. A high sensitivity of 77% and specificity 95% were obtained for the test groups containing common domains when tested using an interacting set of protein pairs found in the yeast DIP database. Correlations were observed between the interacting probability and the accuracy of the prediction, making the output probability a useful indicator of prediction confidence. This method was also somewhat successful when tested on an artificially made random pairing of proteins used as a negative test set of non-interacting protein pairs. This method is particularly advantageous because it also allows for mass prediction of whole protein interactions, which in turn makes it possible to construct entire protein interaction networks.

Finally, Kim et al. developed a database for potentially interacting domain pairs (PID) refined from the DIP database of interacting proteins by making use of InterPro, an integrated database of protein families, domains, and functional sites. A statistical scoring system, "PID matrix score" was developed as a reliability index for accurate functional analysis of interaction networks and a measure of the interaction probability between domains. This method combines various kinds of information such as sequences, interacting regions, and domains of both interacting partners [47]. In order to evaluate the predictive power of the PID matrix, cross-validation was performed with subsets of DIP data (positive datasets) and randomly generated protein pairs from TrEMBL/SwissProt database (negative datasets). The prediction system resulted in approximately 50% sensitivity and more than 98% specificity [47]. The result also showed that mapping of the genome-wide interaction network can be achieved by using the PID matrix.

3.5 Primary Protein Structure

Primary protein structure approaches are predicated on the hypothesis that PPIs may be mediated through a specific number of short polypeptide sequences. These sequences do not span whole domains but are found repeatedly within the proteins of the cell. SVM-based learning methods have shown that the primary sequence of an amino acid chain can effectively identify PPIs [49, 50].

An approach by Spriznak et al. integrates the predictions obtained from different computational approaches together with experimental data, so as to provide functional assignments [48]. It was reported that characteristic pairs of sequence-signatures can be learned from a database of experimentally determined interacting proteins, where one protein contains the first sequence-signature and its interacting partner contains the other sequence-signature. The sequence-signatures that appear together in interacting protein pairs are termed correlated sequence-signatures. This analysis is applied to a database of experimentally identified interacting protein pairs in yeast, from which distinct over-represented sequence-signature pairs were identified. Although not every protein with the one signature is expected to interact with every protein with the other signature, this approach can be used to direct and narrow down experimental interaction screens [48].

Another approach is based on the ability of an SVM learning system to automatically recognize correlated patterns of sequence and substructure in the interacting pairs of proteins found in the DIP database. These patterns typically comprise a small number of functional residues in each protein. This computational tool, developed by Bock and Gough, is based on primary structure information as well as associated physicochemical properties such as charge, hydrophobicity, and surface tension. Reported prediction accuracy was 80%, but the test set size was very small (five previously characterized interactions) [49].

Martin et al. describe an algorithm for PPI prediction [50] that follows the approach of Bock and Gough by combining sequence information and experimental data analysis, while extending the concept of sequence-signatures from Sprinzak et al. by using subsequence pairing. Information from experimental data, sequence analysis, and local descriptions of protein pairs, which are more representative of the actual biology of PPI, are combined to generate a novel and even more general descriptor called a signature product. The signature product is then implemented within a SVM classifier as a kernel function [50]. This method was applied to publicly available yeast datasets among others. The yeast and *H. pylori* datasets used to verify the predictive ability of the method yielded accuracies of 70–80% using tenfold cross-validation. The human and mouse datasets were also used to demonstrate that the method is capable of cross-species prediction. This method is advantageous over that of Bock and Gough because it uses only experimental and sequence information, and does not require physio-chemical information. In addition, this approach, unlike that of Sprinzak et al., does not require prior knowledge of domains.

Ben-Hur and Noble [51] also make use of SVMs to predict PPIs, but introduce a novel pair-wise kernel that measures the similarity between two pairs of proteins. SVMs and kernel methods have the ability to integrate different types of information through the kernel function. Here, kernels make use of a combination of data including protein sequence, homologous interactions, and GO annotations. Ben-Hur and Noble explore a number of different kernel functions using yeast PPI data from the BIND database. At a false positive rate of approximately 1%, the sensitivity was 80%. Future directions may include data incorporation from gene expression studies and transcription factor binding data that have been useful in predicting PPIs.

A recent paper by Shen et al. [59] presents another method based on a SVM with a kernel function using only sequence information to predict PPI in *human*. The authors report an average prediction accuracy of 83.90%.

Finally, a method developed in our own laboratory called PIPE (protein-protein interaction prediction engine) is able to predict with high confidence PPIs for any target pair of yeast proteins given only knowledge of their primary structure data [52]. Like other PPI prediction methods, PIPE relies on previously acquired experimentally derived PPI data and extrapolates this information to predict novel PPIs. This engine compiled the dataset of 15 118 PPI pairs of *S. cerevisiae* from the DIP [28] and MIPS [30] databases. PIPE predicts the probability of interaction between two proteins by measuring how often pairs of subsequences in two query proteins A and B are observed to co-occur in pairs of protein sequences known to interact (see Fig. 2). PIPE showed an overall accuracy of 75%, a success rate that is on par with other commonly used biochemical techniques. PIPE analysis also has other applications in that it can be used to study the internal architecture of yeast protein complexes [52].

To validate the predictive accuracy obtained from PIPE, previously published positive and negative validation datasets were tested. Over a positive database of 100 known protein pairs PIPE displayed a sensitivity of 61% and

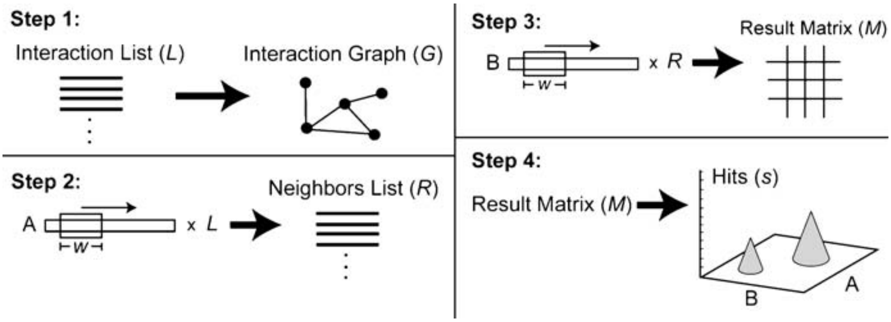


Fig. 2 Design of PIPE algorithm [52]: *Step 1*: The interaction list (dataset of 15 118 known interactions) is used to create an interaction graph *G*. *Step 2*: The first sequence is fragment using a sliding window and used to find all sequences in the database similar to it. For all sequences found, its neighbors in *G* are added to a neighbors list *R*. *Step 3*: The second sequences is also fragmented and is then used to scan the list *R*. For every match a score of 1 is incremented in the result matrix *M*. *Step 4*: Once Step 3 is done we graph the result matrix *M*, which will show visually the peaks representing possible interaction sites

a false negative rate of 39% [52] in predicting yeast PPIs. On the other hand, comparing the data obtained from PIPE with the negative validation dataset helped to verify the false positives rate for PPI. It was found that PIPE falsely detected only 11% non-interacting proteins pairs as interacting pairs. This indicates an 11% false positives rate and 89% specificity rate [52] for the detection of PPI in yeast. Overall, PIPE has the accuracy of 75% [52] and has lower false positive and negative rates than TAP-tagging and Y2H analysis [60].

PIPE also has the ability to identify interacting sites within the sequence of the interacting protein pairs. For example, PIPE also identified previously reported interaction sites between the first 75 amino acid residues of YCR084C and the N terminal region of YBR112C. Figure 3 illustrates that PIPE identified that amino acid region 350–410 of protein YNL243W may interact with the amino acid region 100–250 of protein YBL007C, with a score of 40.

PIPE has been employed to identify and validate a novel PPI between YGL227W and YMR135C. Although yeast gene deletion studies indicated that both YGL227W and YMR135C may be involved in the catabolism of fructose-1,6-bisphosphatase (FBPase) [61], little else is known about them. Following a PIPE prediction that these two proteins may interact, dual TAP-tagging experiments performed in our laboratory identified both of these proteins in co-purification complexes. Moreover, the YGL227W TAP-tagged protein was co-purified with six other proteins in what we termed the vid30 complex. While TAP-tagging does not determine the internal architecture of this complex, PIPE was able to analyze systematically each of the 21 possible PPIs to predict the internal architecture of the vid30 complex. PIPE found that four

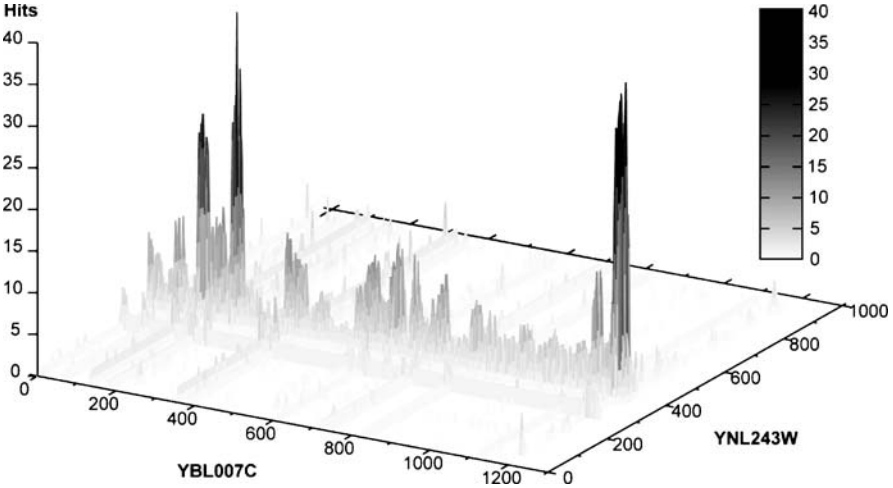


Fig. 3 Possible interaction sites between YNL243W and YBL007C [52]. The highest scoring (*dark*) regions represent the theoretical sites of interaction between the two proteins

proteins formed the core of the complex, whereas three other proteins only interact with YGL227W and YIL017C, but not with each other.

Since the original release [52], we have strived to improve the performance and accuracy of PIPE in order to scan the entire yeast genome. In our most recent work (to be published), we have improved the speed of PIPE over 16 000-fold and increased specificity (∼99.9%) at the expense of a lower sensitivity (∼15%). These improvements, together with the use of a high performance cluster computer, allowed us to do an all-to-all examination of the entire yeast genome (6304 proteins, 19 867 056 possible pairs) in order to detect novel PPIs. Our improved method detected a total of 29 589 interactions, of which 14 438 have not been previously reported in any large-scale database.

4 Validation of Experimentally Determined Interactomes

Reports show that the intersections between various interaction maps obtained using different methods are very small. A comparison study carried out by Aloy and Russell in 2002, showed a low level of overlap among two-hybrid, affinity purification, mass spectrometry, and bioinformatics methods [6, 17]. One such measure for the validation of computational methods is the "interaction generality" measure (IG1) [62]. IG1 is the number of proteins involved in a given interaction or the number of proteins that directly interact with the target protein pair. This measure is based on the assumption that interactions observed in a complicated interaction network are likely to be true positives, while interacting proteins that appear to have many other interacting partners that have no further interactions are likely to be false positives. Interactions with low generalities were more likely to be reproducible in other independent assays and these protein pairs are likely to be co-expressed and are therefore physically related. In [62], Saito et al. were able to refine the existing networks as determined by Uetz et al. [9] and Ito et al. [10]. The authors also developed a new "interaction generality" measure (IG2) that considered the topological properties of the protein interaction network beyond the target pair of proteins. IG2 was found to assess the reliability of putative PPIs with higher accuracy [62].

Another measure used to determine the reliability of an interaction between two proteins is the correlation of their mRNA expression levels. This is then used to determine an expression profile reliability index (EPR), which monitors the fraction of interacting proteins [63]. A paralogous verification method (PVM) was also developed in which paralogous interacting proteins are searched in the DIP database and counted. The reliability of their interaction is then determined on the basis of this count [63].

5 Strengths, Weaknesses, and Challenges of Computational PPI Predictions

Researchers have embraced the use of computational methods in the elucidation of PPIs. Computational PPI prediction methods are an invaluable source of information that complement labor-intensive experimental approaches such as Y2H and TAP-tagging. However, the high-throughput nature of bioinformatics tools should require that computational predictions be deemed reliable only after proper scrutiny. Appropriate measures to evaluate the significance of the interactions should be developed to minimize the number of results that give false positives and negatives. While it is often difficult to differentiate between novel interactions and false positives, additional contextual clues including function, expression, and localization should be brought into consideration. As computational methods are based directly or indirectly on experimentally obtained data, the inaccuracies in the original data will likely be propagated into the predictions.

Several other factors contribute to the challenges that face computational PPI predictions. False positives are prevalent in most computational methods, but we can easily find an explanation. The model organism used for testing in many methods, yeast, contains roughly 6300 proteins [64], which yields approximately ∼19 million possible pairs. Even with a false positive rate as low as 1%, we would anticipate 190 000 falsely predicted interactions. It has been estimated that, in actuality, there are anywhere between 10 000 and 30 000 interactions in yeast [64–70]. Recent large-scale studies contain datasets of a size closer to the bottom end of that range (7123 in Krogan et al. [71]). We can therefore see that the positive interactions are vastly outnumbered by the number of negative interactions. Even if we assume there are 30 000 possible interactions there is still more than a 600:1 ratio of negative to positive interactions (∼0.158%). Therefore it is extremely difficult to recognize the true positive predictions among the overwhelming background of false positive predictions.

The lack of reliable a gold standard makes the assessment of prediction accuracy by the various tools somewhat arbitrary. The establishment of a gold standard is essential to measure progress in the field and will also serve as training material for the next generation of prediction methodologies. Strong gold standard datasets need to be constructed from multiple lines of evidence, including structure where possible, and made freely available.

Recent developments in computational interaction prediction have opened the door to predicting entire interactomes for a variety of organisms. For the most sophisticated approaches, this objective is very computationally expensive and time-consuming. However, algorithmic optimizations and continued improvements in hardware performance will help overcome these challenges.

6 Future Work

It is expected that the number of computational tools for predicting novel PPIs will continue to grow for at least another decade. The increasing prediction accuracy of such tools makes them even more useful for the validation and analysis of diverse interactomes. The growing availability of high quality system biology data may provide the basis for even higher prediction accuracy for such methods. For example, regardless of the hypotheses from which computational tools are originated, the increasing availability of 3D structures of proteins and protein complexes should provide a highly improved starting dataset, which in turn can increase the accuracy of future tools to predict novel PPIs.

One possible direction for development of future tools is to include multiple categories of characteristics/approaches to predict an interaction. In fact, some recently published tools make use of a combination of characteristics to make their predictions [72, 73]. Other investigations may focus on the elimination of false positives associated with computational tools. The presence of false positives in almost all computational methods has provided a challenge for computational biologists. This might be overcome by using vigorous filters that may consider other information about the target interaction. Evidence for the development of such tools can already be seen in the literature, where for example GO ontology has been used as a filter [74].

7 Conclusions

In spite of the number of challenges that are faced in the use of computational methods, one can only expect that they will have even wider applications in the genome-wide analysis of interactomes. The most obvious result of this will be the enlargement of protein databases. It is also expected that the efficiency of these methods will improve. At present, there is an emergence of a more integrated strategy in which genomic, proteomic, and other forms of data are incorporated into the process of generating protein interaction maps. It appears that these strategies will also be able to take other cellular processes such as post-translational protein modification and protein degradation into consideration.

It is impossible to deny the invaluable insight into the organization of living organisms that has been provided by even the simplest of protein interaction models. As these models become more sophisticated, computational methods will become of even more importance.

References

1. Alm E, Arkin AP (2003) Curr Opin Struct Biol 13:193
2. Claverie JM (2001) Science 291:1255
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, 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 JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M et al. (2001) Nature 409:860
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, 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 AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C et al. (2001) Science 291:1304
5. Price ND, Papin JA, Schilling CH, Palsson BO (2003) Trends Biotechnol 21:162
6. Franzot G, Carugo O (2003) J Struct Funct Genomics 4:245
7. Salwinski L, Eisenberg D (2003) Curr Opin Struct Biol 13:377
8. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Science 274:546
9. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) Nature 403:623
10. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) Proc Natl Acad Sci USA 98:4569
11. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD,

Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Nature 415:180

12. Mann M, Pandey A (2001) Trends Biochem Sci 26:54
13. 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 RA, Gerstein M, Snyder M (2001) Science 293:2101
14. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, Quondam M, Zucconi A, Hogue CW, Fields S, Boone C, Cesareni G (2002) Science 295:321
15. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Science 294:2364
16. Ge H, Liu Z, Church GM, Vidal M (2001) Nat Genet 29:482
17. Aloy P, Russell RB (2002) Trends Biochem Sci 27:633
18. Fields S, Song O (1989) Nature 340:245
19. Stephens DJ, Banting G (2000) Traffic 1:763
20. Semple JI, Sanderson CM, Campbell RD (2002) Brief Funct Genomic Proteomic 1:40
21. James P, Halladay J, Craig EA (1996) Genetics 144:1425
22. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) Nat Biotechnol 17:1030
23. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, 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 MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Nature 415:141
24. Rohila JS, Chen M, Cerny R, Fromm ME (2004) Plant J 38:172
25. Rubio V, Shen Y, Saijo Y, Liu Y, Gusmaroli G, Dinesh-Kumar SP, Deng XW (2005) Plant J 41:767
26. Werler PJ, Hartsuiker E, Carr AM (2003) Gene 304:133
27. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) Nucleic Acids Res 29:242
28. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) Nucleic Acids Res 30:303
29. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) Nucleic Acids Res 34:D535
30. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B (2002) Nucleic Acids Res 30:31
31. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) Nucleic Acids Res 35:D572
32. Dandekar T, Snel B, Huynen M, Bork P (1998) Trends Biochem Sci 23:324
33. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Science 285:751
34. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Nature 402:86
35. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Proc Natl Acad Sci USA 96:4285
36. Vert JP (2002) Bioinformatics 18(1):S276
37. Pazos F, Valencia A (2001) Protein Eng 14:609
38. Ma B, Elkayam T, Wolfson H, Nussinov R (2003) Proc Natl Acad Sci USA 100:5772
39. Aloy P, Russell RB (2003) Bioinformatics 19:161

40. Lu L, Lu H, Skolnick J (2002) Proteins 49:350
41. Wodak SJ, Mendez R (2004) Curr Opin Struct Biol 14:242
42. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A (2005) Nucleic Acids Res 33:W331
43. Espadaler J, Romero-Isart O, Jackson RM, Oliva B (2005) Bioinformatics 21:3360
44. Deng M, Mehta S, Sun F, Chen T (2002) Genome Res 12:1540
45. Han DS, Kim HS, Jang WH, Lee SD, Suh JK (2004) Nucleic Acids Res 32:6312
46. Han DS, Kim HS, Jang WH, Lee SD, Suh JK (2004) Genome Inform 15:171
47. Kim WK, Park J, Suh JK (2002) Genome Inform 13:42
48. Sprinzak E, Margalit H (2001) J Mol Biol 311:681
49. Bock JR, Gough DA (2001) Bioinformatics 17:455
50. Martin S, Roe D, Faulon JL (2005) Bioinformatics 21:218
51. Ben-Hur A, Noble WS (2005) Bioinformatics 21(1):i38
52. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Golshani A (2006) BMC Bioinformatics 7:365
53. Marcotte EM (2000) Curr Opin Struct Biol 10:359
54. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Proc Natl Acad Sci USA 95:14863
55. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) Nature 402:83
56. Aloy P, Russell RB (2002) Proc Natl Acad Sci USA 99:5896
57. Berman H, Henrick K, Nakamura H, Markley JL (2007) Nucleic Acids Res 35:301–303
58. Sonnhammer EL, Eddy SR, Durbin R (1997) Proteins 28:405
59. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H (2007) Proc Natl Acad Sci USA 104:4337
60. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Trends Genet 18:529
61. Regelmann J, Schule T, Josupeit FS, Horak J, Rose M, Entian KD, Thumm M, Wolf DH (2003) Mol Biol Cell 14:1652
62. Saito R, Suzuki H, Hayashizaki Y (2002) Nucleic Acids Res 30:1163
63. Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Mol Cell Proteomics 1:349
64. Grigoriev A (2003) Nucleic Acids Res 31:4157
65. Bader GD, Hogue CW (2002) Nat Biotechnol 20:991
66. Legrain P, Wojcik J, Gauthier JM (2001) Trends Genet 17:346
67. Tucker CL, Gera JF, Uetz P (2001) Trends Cell Biol 11:102
68. Sprinzak E, Sattath S, Margalit H (2003) J Mol Biol 327:919
69. Walhout AJ, Boulton SJ, Vidal M (2000) Yeast 17:88
70. Hazbun TR, Fields S (2001) Proc Natl Acad Sci USA 98:4277
71. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Nature 440:637
72. Wang H, Segal E, Ben-Hur A, Li Q, Vidal M, Koller D (2007) Genome Biol 8:R192
73. van Berlo RJP, Wessels LFA, de Ridder D, Reinders MJT (2007) J Bioinform Comput Biol 5:839
74. Mahdavi MA, Lin YH (2007) BMC Bioinformatics 8:262

Subject Index