Mass Spectrometry in Chemical Biology

Evolving Applications

Chemical Biology

Editor-in-chief:
Tom Brown, *University of Oxford, UK*

Series editors:
Kira J. Weissman, *Lorraine University, France*
Sabine Flitsch, *University of Manchester, UK*
Nick J. Westwood, *University of St Andrews, UK*

Titles in the series:
1: High Throughput Screening Methods: Evolution and Refinement
2: Chemical Biology of Glycoproteins
3: Computational Tools for Chemical Biology
4: Mass Spectrometry in Chemical Biology: Evolving Applications

How to obtain future titles on publication:
A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

For further information please contact:
Book Sales Department, Royal Society of Chemistry, Thomas Graham House,
Science Park, Milton Road, Cambridge, CB4 0WF, UK
Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247
Email: booksales@rsc.org
Visit our website at www.rsc.org/books

Mass Spectrometry in Chemical Biology

Evolving Applications

Edited by

Norberto Peporine Lopes
University of Sao Paulo, Brazil
Email: npelopes@fcfrp.usp.br

and

Ricardo Roberto da Silva
University of California, San Diego, USA
Email: ridasilva@ucsd.edu

ROYAL SOCIETY
OF CHEMISTRY

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2013

Chemical Biology No. 4

Print ISBN: 978-1-78262-527-8
PDF ISBN: 978-1-78801-039-9
EPUB ISBN: 978-1-78801-346-8
ISSN: 2055-1975

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

Whilst this material has been produced with all due care, The Royal Society of Chemistry cannot be held responsible or liable for its accuracy and completeness, nor for any consequences arising from any errors or the use of the information contained in this publication. The publication of advertisements does not constitute any endorsement by The Royal Society of Chemistry or Authors of any products advertised. The views and opinions advanced by contributors do not necessarily reflect those of The Royal Society of Chemistry which shall not be liable for any resulting loss or damage arising as a result of reliance upon this material.

The Royal Society of Chemistry is a charity, registered in England and Wales, Number 207890, and a company incorporated in England by Royal Charter (Registered No. RC000524), registered office: Burlington House, Piccadilly, London W1J 0BA, UK, Telephone: +44 (0) 207 4378 6556.

For further information see our web site at www.rsc.org

Printed in the United Kingdom by CPI Group (UK) Ltd, Croydon, CR0 4YY, UK

Foreword

Mass spectrometry (MS) is one of the core analytical techniques for the identification and confirmation of molecules. The origins of the technique can be traced back to the pioneering work of Nobel prize-winners J. J. Thomson and F. W. Aston at the University of Cambridge. Since then MS has undergone over one hundred years of continuous development and has won three Nobel prizes for key developments along the way. The analysis of labile, thermally unstable analytes (for example peptides, proteins, saccharides and complex natural products) has always been the driving force for the development of new MS approaches and methods. Moreover, the ability to detect and quantify metabolites from natural or physiological sources, or even *in situ* in plant material or animal tissue, has opened up the whole world of synthetic biology to MS analysis.

My own first experiences of MS were at Warwick University where, for my final year research project, I studied the analysis and sequencing of simple peptides by fast-atom bombardment (FAB)-MS. This technique was severely limited by molecular weight (less than 1500 Da) due to the competition between energy dissipation along the analyte molecule *vs.* bond fission. Additionally, not all peptides wanted to ‘play ball’ and just simply refused to fragment. When it worked, this methodology was, of course, a lot faster and more sensitive than non-MS approaches, but it was very unreliable and as a result didn't exactly ‘set the world on fire’! After completing my degree, I returned to my home town of Cambridge and spent a couple of years working at the Mass Spectrometry service in the Department of Chemistry at Cambridge University, spending most of my time analysing heavy metal catalysts for groups of the ilk of Professor The Lord Lewis. Most of the analyses were performed by FAB-MS along with the allied technique of liquid secondary ion mass spectrometry (LSIMS). At the same time, I was often asked to analyse some of the more intractable samples from the natural product groups of

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-FP005

James Staunton, Dudley Williams and Ian Fleming. This was very much my formative period and during this time I learned a considerable amount about both the theory and applications of MS, and the seeds of my future career were very much being sown. The biggest lesson I learned was that it was all too easy to give up on an analysis and blame the sample for poor quality spectra. However, I was a fast learner and also a bit of a perfectionist (comes with the job I think), and I soon learned that sample preparation was often as much to blame. This is an important lesson, especially when you are working for such distinguished chemists as Lewis, Staunton, Williams and Fleming! In particular, with FAB-MS, the quality of the resulting spectra was often very much due to the formation of the matrix/analyte drop on the probe tip. Often the drop would ‘skin over’ and no spectra could be obtained. Sometimes you would have to repeat the analysis five or six times to get a spectrum. We never really understood the skin formation process, but it was most probably down to the way the matrix dried, sometimes forming a crust that was impenetrable to the xenon beam. On a good day, I would go home happy if I had obtained more than ten good quality spectra.

After this period at Cambridge I returned to Warwick University to work with Peter Derrick towards my MSc by research. He had just obtained one of the first commercial matrix-assisted laser desorption/ionisation (MALDI) instruments using a short time-of-flight analyser. My project was to test the utility of this instrument and technique for the analysis of saccharides. Again, this project taught me about the importance of sample preparation. With MALDI you have the complexity of matrix choice, matrix/analyte ratios and spot preparation. In this time period (mid 1990s) little was really known about how MALDI worked and sample preparation was like a ‘black art’. To the uninitiated, MALDI-MS can still seem like magic, but the extent of knowledge is now considerable, and there are matrices and successful sample preparation methodologies for almost all sample types. However, a universal matrix and sample preparation methodology remains elusive.

After completing my MSc, I stayed on for a short period at Warwick and helped with the establishment of the high-field Fourier-transform ion cyclotron resonance (FT-ICR) facility. This was a real eye opener for me and was a major step-change for the analysis of natural systems by MS in the UK. This instrument used electrospray ionisation (ESI) and, coupled with the high-resolution abilities of the instrument, intact biological compounds could be analysed rapidly with relative ease. FT-ICR-MS also brought an unsurpassed ability to perform controlled tandem and sequential MS as well as gas-phase chemistry experiments. After six months, I returned to the University of Cambridge to study for a BBSRC funded PhD with James Staunton on natural product MS using their newly acquired 4.7 Tesla FT-ICR-MS instrument. Although the ESI source on this instrument was fairly crude by today’s standards, it allowed for a whole realm of analyses not previously available with older ‘soft’ ionisation techniques such as FAB, LSIMS or plasma desorption.

My PhD project was to try to develop a routine methodology for the analysis of target natural products being developed as potential new antibiotics generated through a combination of genetic engineering of the bacteria, which produced the synthase enzymes, and biochemical techniques to manipulate the tailoring of enzymes at the protein level, which were in turn responsible for oxidation, reduction and glycosylation of the final (non-)natural products. This work produced a lot of very similar structural variants and co-metabolites and the group needed a fast, sensitive and reliable method of confirming structure to then feed back to the biochemists that their chemistry was (or was not) working. This required me to develop a high-resolution sequential MS protocol and a parallel understanding of molecular fragmentation. To cut a long story short, with my methodology, I was not only able to confirm that loading modules of the enzyme were functioning correctly, but also that the various tailoring enzymes were producing the correct final molecules from a five-minute experiment using accurate-mass, high-resolution ESI–FT-ICR-MS/MS.

After completion of my PhD, I stayed at Cambridge for a four-year post-doc, working with James Staunton and Steven Ley. During this time, I worked on increasing the understanding of natural product fragmentation in ESI-MS/MS and the exploitation of this increased understanding of structural elucidation. One of the more bizarre memories I have is of Dr Lopes (now Professor) climbing on a stepladder with a heat gun to heat the glass inlet. In the inlet was solid D_4-methanol at vacuum. We needed to generate a low vapour pressure of gaseous D_4-methanol to bleed into the FT-ICR cell to try to prove an unusual gas-phase substitution reaction. Needless to say, the experiment worked and it led to a paper in *Chemical Communications*. It was an unusual experiment, and a great lasting memory of a very productive and fun four years of academic research.

In 2003, I moved to my current position as manager of the School of Chemistry MS facility at the University of Bristol. During my time here, the facility has undergone substantial investment resulting in expansion from two aging (and failing) instruments to more than ten, including state-of-the-art Orbitrap, nanospray IMS-MS/MS and high-resolution MALDI-TOF-MS/MS instruments. The facility is used to carry out cutting edge MS-based research in natural products, proteomics, metabolomics (both biological and environmental) and petroleomics, as well as studies of protein structure and function. The facility also supports cutting edge research across a whole range of chemistries, including studies of synthetic life, superconductors, clean fuels, environmentally friendly catalysts, self-assembly nanomaterials and metallopolymers. The facility also hosts the departmental MS service, which primarily supports the synthetic and materials chemists in the School. The resulting samples range from simple low-mass organic compounds to metal catalysts, synthetic polymers, non-natural peptides and dendrimeric cage systems designed for drug transport. This diverse range of analytes brings with it its own challenges for the facility, not least of which is sample preparation. Somehow, I also find time to conduct my own research into natural product ionisation and fragmentation, and increasing our understanding of gas-phase fragmentation. It is my long-term hope that such knowledge can be applied to ever increasingly complex samples such as blood plasma and crude oil. At the same time, I have also developed a totally new methodology for the MALDI analysis of small molecules using colloidal graphite as the matrix.

This book brings together an international array of renowned scientists active in a diverse range of MS research areas from multi-omics (proteomics, metabolomics, lipidomics) to enzyme action and natural product discovery. As a result, this book should serve as a comprehensive review of current MS applications in analytical synthetic biology including enzymology, biomedical research, natural products and drug discovery.

Paul J. Gates
University of Bristol, UK

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-FP005

Contents

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-FP009

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-FP009

Acknowledgments

The editors are grateful to University of São Paulo, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for financial support. R. R. d. S. was supported by the Fellowship (FAPESP-2014/01884-2).

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

CHAPTER 1

Introduction

V. ZAGORIY

MetaSysX GmbH, Am Muehlenberg 11, 14476 Potsdam, Germany
*E-mail: zagoriy@metasysx.eu

Chemical biology is a young discipline without a rigid definition. *Nature Chemical Biology*, a leading journal in the field, defines 'chemical biology' as the application of chemical methods to solve biological problems. Long before the term 'chemical biology' was coined, many fascinating discoveries in biochemistry had been made by applying chemical methods to biological phenomena. Biochemistry as a science can be traced back to the synthesis of urea by Wöhler in 1828 who, for the first time, was able to prepare a compound originating from living organisms through chemical synthesis.

Modern biochemistry relies heavily on organic mass spectrometry (MS), a method that originates from physics and that further evolved in the hands of chemists. Neither field describes the whole history of chemical methods in biology that finally led to the establishment of MS as a routine biochemical technique. The description of biological breakthroughs that were achieved with MS cannot be given in a single book, even less so a single chapter. In this introduction we present some selected discoveries that, by demanding exact molecular mass measurements, paved the way for biological MS, starting with the history of the discovery of lipophilic vitamins D and E at the beginning of the 20th century.

In 1925 it was known that two possibilities existed for the prevention of rachitis. One was the administration of cod liver oil and the other was irradiation of the skin with ultraviolet light. At that point it was considered that

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

only two distinct antirachitic factors existed, until it unexpectedly turned out that UV-irradiation of food was also sufficient to prevent rachitis. It was concluded that a certain chemical factor exists, a pro-vitamin, which upon UV-irradiation is converted to an active antirachitic compound. In initial attempts to isolate this vitamin it was shown that the fraction that contained this pro-vitamin mainly contained cholesterol, suggesting that the pro-vitamin might also be a steroid. The biggest expert in the field of steroid chemistry at the time was Adolf Windaus from the University of Göttingen, who was invited to join the research in order to isolate and identify the antirachitic vitamin. Initially, the direct isolation of the vitamin from natural sources did not prove to be successful, so Windaus turned to an empirical approach, testing the antirachitic effect of known steroids after UV-irradiation. UV-treated cholesterol did not have any antirachitic effects (thus the expected vitamin D1 was not discovered), but ergosterol and 7-dehydrocholesterol produced antirachitic compounds after UV-treatment, which were named vitamins D2 and D3 respectively.[1] Final proof of the vitamin D identity was obtained when Hans Brockmann, a student of Windaus, managed to isolate a natural antirachitic compound from tuna liver oil using liquid–liquid partitioning and column chromatography. Isolation was guided by an activity assay using rachitic rats and the isolated active compound was shown to be identical to the product of UV-treatment of 7-dehydrocholesterol.[2]

The history of the identification of vitamin E is closely related to vitamin D research. The existence of vitamin E was shown in an experiment by the laboratory of Herbert Evans at UC-Berkeley in 1922. They observed that rats that were fed a diet of purified protein, fat, carbohydrates, and vitamins A, B and C, which were already known at the time, were sterile,[3] unlike animals given normal complete food. Isolation of the unknown factor necessary for reproduction was attempted by a number of research groups, mainly by means of liquid separation and adsorption chromatography. This approach did not prove to be successful. In 1932 Herbert Evans and two of his group members, Oliver and Gladys Emerson, travelled to Göttingen for a research stay at the laboratory of Alfred Windaus.[4] After this visit, Evans and the Emerson couple made a new attempt at isolating vitamin E using an approach used by Adolf Butenandt, a student of Windaus, in many of his works on the isolation of natural compounds. Instead of isolating the intact vitamin E, they tried to crystalize it out of an enriched extract after chemical derivatization. Wheat germ was used as a starting material and an extract of non-saponified lipids turned out to be particularly active in supporting the reproductive function of rats. Evans and his colleagues did manage to crystalize the active compound out of this extract after treating it with cyanic acid,[5] suggesting the presence of hydroxyl groups in vitamin E, which form a crystallizable allophonate. Final structure elucidation was performed by another Windaus student, Erhard Fernholz, who in 1938 showed that purified vitamin E under pyrolytic conditions decomposes into durohydroquinone and a hydrocarbon residue $C_{19}H_{38}$ (Figure 1.1). Under oxidative conditions, vitamin E produced a five-membered lactone bearing a hydrocarbon residue. Based on these findings Fernholz correctly identified the structure of vitamin E.[6]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

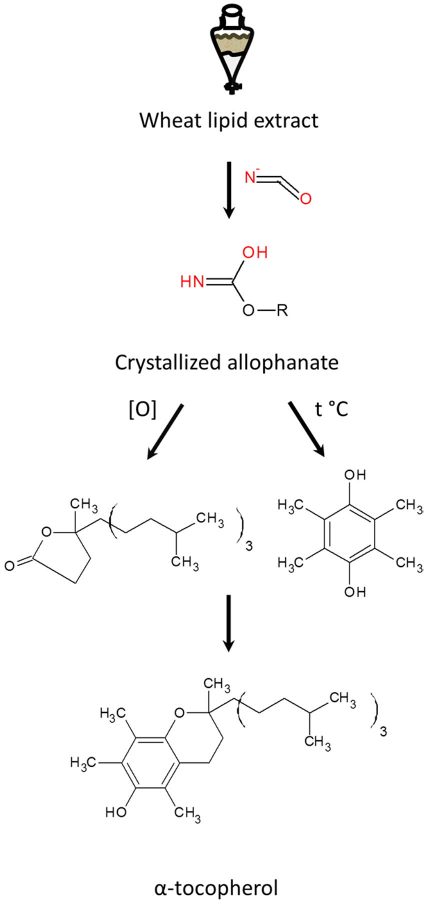

Figure 1.1 Isolation and identification of α-tocopherol.

In these examples of vitamin research, low-efficiency isolation separation methods such as liquid–liquid partitioning or adsorption chromatography were used. These methods were appropriate for isolation of vitamins that were relatively abundant in the starting material, but as biological research started to deal with minute amounts of target compounds a need for new separation methods that could deal with much smaller amounts became clear.

At the same time as the group of Herbert Evans was developing an approach to crystallizing vitamin E, Archer Martin, a young PhD student from Cambridge, attempted isolation of non-modified vitamin E using counter-current liquid–liquid extraction. However, the group of Evans was the first to publish a report on its isolation and the results of Martin never appeared on paper.[7] However, he built an extremely sophisticated counter-flow machine for his project, which gained him popularity as a solvent–solvent partitioning expert.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

As such, in 1938 he was approached by Richard Synge who was then working on the separation of amino acids. Among other methods, Synge tried the separation of amino acids using adsorption chromatography, which implements partitioning between a liquid mobile phase and a solid sorbent. However, the interaction of all amino acids with sorbents available at that time turned out to be very similar and did not permit separation of an amino acid mixture. Therefore, Synge sought the help of Archer Martin in trying to achieve amino acid separation using counter-current extraction. He had already shown that different acetylated amino acids had different partition coefficients between chloroform and water, and was looking for a way to use this fact for separation purposes. Despite certain success, such an approach turned out to be extremely cumbersome and technically demanding. A breakthrough was achieved when Martin decided to immobilize one liquid phase while moving the other. This was done by soaking silica gel with water and packing it into a column. Chloroform running through this column served as a mobile liquid phase. Partitioning of the analyte between the column and the mobile phase in this apparatus turned out to be very similar to the partitioning of analyte between two liquid phases.[8] This method, which was later named partition chromatography, significantly improved the separation of acetylated amino acids, but it was still not good enough due to analyte adsorption on the silica, which interfered with the chromatography. Martin and Synge next tried using paper as a carrier for the immobilized water. Instead of packing paper soaked with water into a column they put a drop of the mixture to be separated onto the corner of a paper sheet and then "developed" it by putting one edge of the paper into a beaker with mobile phase. When the mobile phase reached the opposite edge, driven by capillary forces, the paper was taken out, dried and inserted with a neighboring edge into a different mobile phase, thus permitting "two-dimensional" separation. This method turned out to be so applicable to freeing amino acids that Synge used it to determine the sequence of gramicidin S. His research group first determined that this biologically active compound consists of valine, ornithine, leucine, phenylalanine and proline in equimolar amounts.[9] The molecular mass suggested that gramicidin S is a decapeptide and its partial hydrolysis did yield a number of di- and tripeptides, which Synge identified using 2D partition chromatography on paper comparing the retention of hydrolysis products with synthetic dipeptides.[10] Matching the sequence of dipeptides in the hydrolysate indicated a repeating pentapeptide sequence Val–Orn–Leu–Phe–Pro. The total sequence was suggested as being cyclic (Val–Orn–Leu–Phe–Pro)$_2$ with the first valine linked to the last proline.[11] This was the first-ever identified peptide sequence! Partition chromatography on paper was further used by Fred Sanger in his work on the sequence identification of insulin and generally became an extensively used analytical method. In the form of high-throughput thin layer chromatography, it is still used by biochemists today. In 1949 Archer Martin started working on the separation of fatty acids and came to the conclusion that a polar stationary phase did not permit a good enough separation with any of the available solvents. In order to improve the situation he tried to switch the chromatographic phases. Treatment of silicon dioxide with highly

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

hydrophobic dichlorodimethylsilane covalently modified the former, changing it to an immobilized hydrophobic phase. Using water solutions of different alcohols as a mobile polar phase, Martin was able to achieve separation of long chain fatty acids.[12] This separation method, later termed reversed-phase liquid chromatography (LC), is one of the most popular separation methods in bioanalytical chemistry today, and Martin and Synge received the 1952 Chemistry Nobel Prize for their work.

Reversed-phase chromatography proved to be an extremely powerful method for the separation of even small amounts of target substances from complex mixtures. One elegant example of the use of reversed-phase chromatography was the identification of the sex-attracting pheromone of the silkworm *Bombyx mori*, the first ever identified semiochemical. Male silkworm moths sense the presence of unfertilized female moths over very long distances. Up to the middle of 20th century this effect was explained as some kind of electromagnetic phenomena, ranging from infrared to X-ray radiation. Adolf Butenandt, at that time already a Nobel Prize winner for his work on sexual steroid hormones, made a suggestion that sexual attraction in silkworms is mediated by a chemical. Male silkworm moths when kept alone may not move for hours, but if a female moth is brought into the vicinity they start to move in a certain zigzag pattern. The same behavior is induced if female scent glands or their extracts are used instead. The group of Butenandt used extracts of scent glands obtained after the dissection of 500 kg silkworm female moths[13] as a starting material for the isolation of the active compound. Every round of isolation was controlled using an activity assay based on the behavior of the male. After preliminary separation using standard liquid–liquid partitioning techniques the activity was isolated using multiple rounds of reversed-phase partition chromatography. Butenandt further used chemical degradation to show that, structurally, the pheromone is a long-chain unsaturated alcohol (Figure 1.2), which he named “bombykol”, thus identifying the first known pheromone.

Liquid partitioning column chromatography proved to be an excellent tool for separation purposes because, among other reasons, it operates under mild conditions and sample decomposition was rather insignificant. However, the chromatographic properties of early liquid–liquid partitioning columns were far from ideal. In 1951 it was once again Archer Martin who suggested that, instead of using liquid as a mobile phase, a compressed gas could be used instead. If the chromatographic column is heated so that the analytes can be vaporized without breaking down, partitioning between a gas and a liquid occurs. This permits chromatographic separation while gas is moving along the column filled with a carrier with bound stationary liquid phase.[14] Due to the lower viscosity of the gas compared to the liquid, much longer columns may be used, significantly increasing the separation efficacy. This new method was given the name gas chromatography. Detection of separated compounds could be performed in many ways. For example Martin himself used titration to detect the separated fatty acids. Soon afterwards, combinations of gas chromatography with MS appeared.

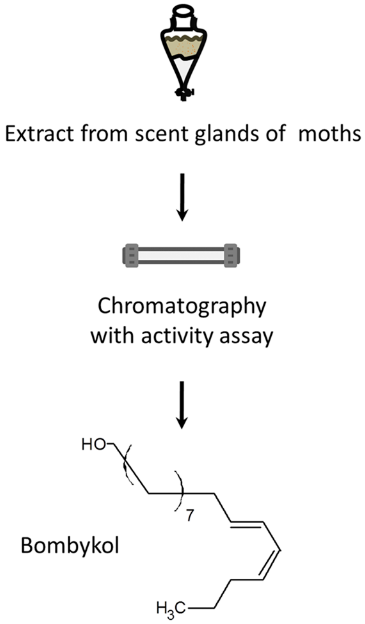

Figure 1.2 Isolation and identification of bombykol.

Measurement of molecular masses using deflection of a charged moving particle in a magnetic field was first performed by Joseph Thompson in 1910. This type of MS was primarily a tool for physicists, for example in works on the study of isotopes. In 1936 Arthur Dempster, at the University of Chicago, introduced electron bombardment as a method for producing positively charged ions. Combining his ionization source with an improved magnetic mass separator he constructed an instrument that became a prototype for commercially available magnetic sector MSs,[15] quickly finding its way into organic chemistry. For over a decade it was almost exclusively used for the analysis of volatile hydrocarbons. Its major drawback for bioorganic chemistry was multiple fragmentation of the parent ion in the electron bombardment ionization source, so that the molecular structure often had to be "reconstructed" from analysis of fragments. For analysis one would need a pure substance, which was possible in the case of synthetic organic chemistry, but was rarely the case with biological samples. MS had to be preceded by sample separation.

Attempts to hyphenate chromatography with a magnetic sector MS were made in the early 1950's by inserting a small fraction (usually less than 1%) of the gas flow from a gas chromatograph (GC) column into the electron impact ionizer of the MS. However, the scanning speed of the magnetic MS was not sufficient to produce spectra of suitable resolution "on-line" and cumbersome techniques were needed in order to perform the mass measurement. For instance, the gas eluting from the column had to be collected in a cooled glass tube and the condensate injected into a MS. One of the first, and simultaneously brightest, examples of gas chromatography in subsequent, but not directly coupled, MS analysis is the discovery of prostaglandins.

At the beginning of the 1930's, the Swedish scientist Ulf von Euler made the observation that an extract from human seminal fluid and from a sheep vesicular gland (which is one of the male genital organs) induces contraction of smooth muscle. This effect plays an important role in the propagation of semen though the female genital tract. Von Euler suggested the existence of a bioactive factor responsible for this effect, which he termed prostaglandin. Its isolation from a sheep prostate was carried out by Sune Bergström and his colleagues at the Karolinska Institute, with classic activity-driven separation using contraction of a gut strip, which contained a smooth muscle layer, as a measure of biological activity. After numerous rounds of counter-current and partition chromatography they crystallized the active substance. In contrast to the habit of most bioorganic chemists of the time, the mass of the crystallized compound was determined using an electron impact (EI) magnetic sector MS yielding the chemical formula $C_{20}H_{34}O_5$.[16] Structure determination was performed by subjecting the purified prostaglandin to ozonolysis and separation of the reaction mixture using GC with fraction collection. Collected fractions were further analyzed using a magnetic MS in order to establish the putative oxidation products with retrospective comparison of the column retention to available synthetic standards. Such an approach was revolutionary in the field of natural product chemistry at that time. The structure of the isolated prostaglandin E, which turned out to be the first-identified representative of a big family of bioactive compounds, was "assembled"[17] based on structures of the identified oxidation products (Figure 1.3).

Establishment of an on-line GC-MS instrument became possible only after a principal scheme of a MS thatcould perform measurements in a broad mass range simultaneously was proposed in 1946 by W. E. Stephens from the University of Pennsylvania. According to his suggestion it was possible to use space dispersion of accelerated charged particles according to their mass-to-charge ratios. After acceleration with a pulsed electric field, the time needed to travel over to the detector was longer for heavier particles of the same charge as compared to lighter ones. Once this travel time was measured the mass-to-charge ratio could be deduced from the length of the travel path and the voltage of the acceleration pulse. In 1948 A. Cameron and D. Eggers from Clinton Engineering Works presented the first working prototype of a MS that allowed discrimination between atoms of a heavy metal bearing different charges. In 1953 Stephens and Wolff presented a working time-of-flight instrument that allowed mass measurement of different hydrocarbons, though with a very low resolution[18] because in the electron impact ion source the accelerating pulse would hit a cloud of ions with different initial velocities and spatial position. Isomass ions acquired different velocities and arrived at the detector at different times. In 1955 Wiley and McLaren produced an improved scheme of the electron impact ionization source that would account for the initial spatial and velocity distribution of ions,

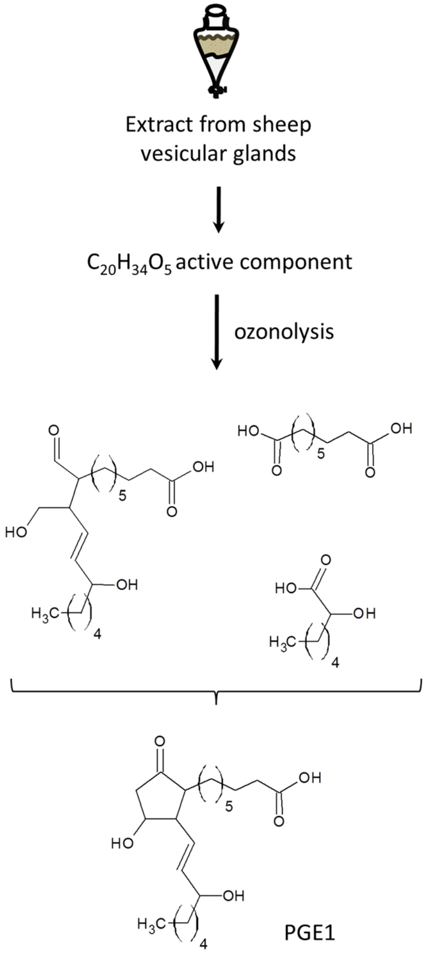

Figure 1.3 Isolation and identification of prostaglandin PGE1.

increasing the resolution of a time-of-flight MS over one hundred-fold.[19] The experimental instrument was tested using xenon isotopes. Wiley wanted to apply his new machine to mixtures of organic compounds and invited two young scientists to his laboratory, McLafferty and Gohlke from Dow Chemicals, who had vast experience in the separation of organic compounds, with their GC, in order to try to connect the two machines by direct infusion of the GC column eluate into the electron impact ionization source of his time-of-flight MS. The acquisition time of one mass spectrum with mass range up to 400 Da was around 20 µs for the instrument they used,[20] allowing sustained mass measurement of the chromatographic eluent. The combination worked, producing MS spectra of separated methanol, acetone, toluene and benzene as the compounds were eluting from the GC column, with the quality comparable to the spectra of pure compounds acquired on a magnetic sector

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

instrument;[21] this gave birth to GC-MS. Later on GC-MS evolved into a leading method for high-throughput profiling of small molecules and, in particular, metabolites. An early example was metabolic profiling of urine, which covered steroids and organic acids.[22] Current developments in metabolic profiling with GC-MS allow simultaneous measurement and identification of hundreds of metabolites from various sources,[23] permitting sophisticated mathematical correlation of the changes in the level of metabolite with amounts of mRNA and proteins. Such a global approach to metabolism is now referred to as metabolomics, and is at the forefront of small molecule biochemistry.

Back in the early 50's, in parallel with the development of analytical methods for small organic molecules, deciphering of metabolic pathways and the discovery of novel biologically active compounds, molecular biology was created by a seminal paper of Watson and Crick in 1953 on the double-helical structure of DNA. Further developments of this new field of life science shifted the interest of the bioorganic community to genes and their products—proteins. At that time, methods of protein purification such as ultracentrifugation and gel electrophoresis were established, liquid-phase synthesis permitted preparation of pure oligopeptides and chemical sequencing of proteins was possible using Edman's degradation. Chemical sequencing, however, was complicated and time-consuming, and MS was a potential alternative. The first biological peptide whose structure was established with the use of MS was fortuitine, a small acylated peptide from the microorganism *Mycobacterium fortuitum*. This nonapeptide carries two methylated leucines, an N-terminal acyl group and a C-terminal methyl group. On the one hand these modifications made the oligopeptide very volatile and on the other the N-terminal acyl group allowed the identification of the end of the chain at which the EI fragmentation took place, since the free amino acids can be formed only if they cleave off at the C-end of the peptide chain, while the amino acids cleaving off at the N-terminus will additionally carry an acyl residue. The sequence was established by following the decreasing masses of the fragments originating from the C-terminus in a magnetic sector MS with high mass accuracy. Special care had to be taken with two possible fragmentations at –CHR–CO– or at –CO–CN– around the amide bond.[24] The observation that the terminal hydrophobic modifications increased the peptide volatility resulted in oligopeptide chemical derivatization with hydrophobic functions prior to introduction into the EI source. Automated analysis of the obtained MS spectra, and subsequent oligopeptide sequence reconstitution using computational algorithms, was first introduced in 1966 by the group of Klaus Biemann at MIT. For oligopeptides of up to five amino acids they devised all their possible sequences and then simulated all possible fragments arising from the breaking of the amide bonds. Matching the simulation with the real spectrum they were able to pick a correct sequence from the list of predicted sequences.[25] However, in general, the electron impact ionization could not be applied to large biomolecules, such as proteins, due to limitations in their volatility, significantly restricting the use

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

of MS for peptide analysis. In 1968 Malcolm Dole from Northwestern University showed that if the solution of intact macromolecules was nebulized in such a way that the formed droplets bore a surface charge, as the solvent evaporated from the droplets the charge repulsion overcame the surface tension which caused the droplets to disintegrate until the surface charge was transferred onto a single macromolecule;[26] importantly, the macromolecules did not fragment. Dole himself oversaw a number of technical details that prevented him from implementing this idea, which was only performed in the mid-1980's by the group of John Fenn at Yale in the form of ESI. Using ESI coupled to a quadrupole mass analyzer, Fenn and his group obtained the mass spectra of non-derivatized gramicidin S as a double-protonated ion.[27] Not only did ESI turn out to be an extremely successful MS interface for protein analysis, it very soon found its way as a hyphen between LC and MS in the form of LC-ESI-MS.[28]

When ESI was introduced, LC columns had developed from the partitioning of chromatography columns of Martin and Synge, with at most a few hundred theoretical plates, to high-performance LC (HPLC) columns with thousands of theoretical plates. This became possible, on the one hand, due to the introduction of chromatographic pumps able to sustain high eluent pressure, and on the other hand due to the introduction in 1968 of silicon-coated glass microbeads as normal stationary phase carriers by Csaba Horvath from Yale University.[29] The small particle size compared to the resins conventionally used significantly improved the separation performance. The silicon coating, representing a thin layer compared to the microbeads diameter, was sufficient to retain the polar separation phase but prevented adsorption of the analyte on the carrier, thus improving the chromatographic resolution further. Shortly after the publication of Horvath and colleagues, HPLC became commercially available from multiple manufacturers and in the late 1980's this separation technique was extremely well established. ESI was ideally suited for introducing the solvent that eluted from the HPLC column into a MS and today HPLC-MS is successfully used for the separation, detection and quantification of virtually all classes of bioorganic molecules. It became an essential bioanalytical method, used in many thrilling discoveries that shaped chemical biology. One such story is the unravelling of developmental regulation with small molecules in the roundworm *Caenorhabditis elegans*.

This organism was introduced by Sidney Brenner in 1968 as a model for his Nobel prize-winning work on neural development. The sexual life cycle of *C. elegans* starts with an egg laid by an adult hermaphrodite. Under favorable conditions, such as the presence of food and moderate population density, the eggs hatch into larvae (so-called L1 larvae), which develop into adult hermaphrodites through a number of other larval stages, L2 to L4. If these later larval stages encounter overcrowding, which means exhaustion of the food source, they die. However if L1 larvae encounter such unfavorable conditions, they undergo a molt into a dauer larva ("dauer" meaning enduring in

German, the language in which this observation was first published). Dauer larvae, or just dauer, are less metabolically active than normal larvae and can survive for many months without food. Additionally, dauers accumulate a number of protective compounds, which makes them remarkably resistant to harsh environmental conditions. After the overcrowding is surpassed or the food source reappears, dauers molt into L4 larvae and resume reproductive development. It was long supposed that small molecules played a crucial role in the regulation of the *C. elegans* life cycle, but only recently has the elegant mechanism that controls reproductive *vs.* dauer developmental switching been elucidated.

In 1982 Golden and Riddle showed that polar extracts from a *C. elegans* culture can induce dauer formation, and suggested the existence of a regulatory pheromone. Its isolation was performed by a group from South Korea using a three hundred liter culture of *C. elegans* as starting material, in which most of the worms did eventually arrest their development as dauer larvae due to overcrowding. The culture medium tested positively for the presence of the active compound, which could be extracted into ethyl acetate. Activity-guided three-round HPLC separation of the organic extract using different columns permitted isolation of a fraction that contained a single compound (Figure 1.4a) according to the MS analysis, and which displayed the dauer-inducing activity in a so-called daumone assay. In this assay *C. elegans* eggs developed through four larval stages into an adult hermaphrodite in the presence of bacteria on the agar surface in a Petri dish. Daumone activity induced the formation of dauers despite the presence of food. Tandem MS fragmentation of the plausible daumone parent ion showed a sugar fragment and heptanoic acid fragment. Further NMR analysis showed that the sugar was ascarylose connected to the ω-1 carbon of the heptanoic acid, identifying the compound to be (6R)-(3,5-dihydroxy-6-methyltetrahydropyran-2-yloxy) heptanoic acid (**1**). The compound got the name ascaroside because of its sugar moiety, and its dauer-inducing activity was confirmed through total synthesis with a positive result in the daumone assay.[30] However, the amount of the synthetic daumone needed to induce dauer formation was much bigger than the actual amount of the natural ascaroside daumone measured in the culture medium, which suggested the presence of other compounds necessary for daumone activity.

In order to identify further dauer inducing ascarosides the group of Frank Schroeder from Cornell University used synthetic ascarosides in order to establish their MS/MS fragmentation pattern. All of the reference compounds produced an ascarylose-derived $C_3H_5O_2$ fragment in negative ionization mode. Using this information the authors performed an LC-MS/MS based screen of the dauer culture medium in which they identified all precursors of the $C_3H_5O_2$ fragment. The structures of the identified precursors were either confirmed with synthetic standards or by NMR.[31] Newly identified ascarosides could be subdivided into ω-1 hydroxylated ascarosides (**2**), their 2,3-enoyl derivatives (**3**) and their 3-hydroxyderivatives (**4**), with each

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

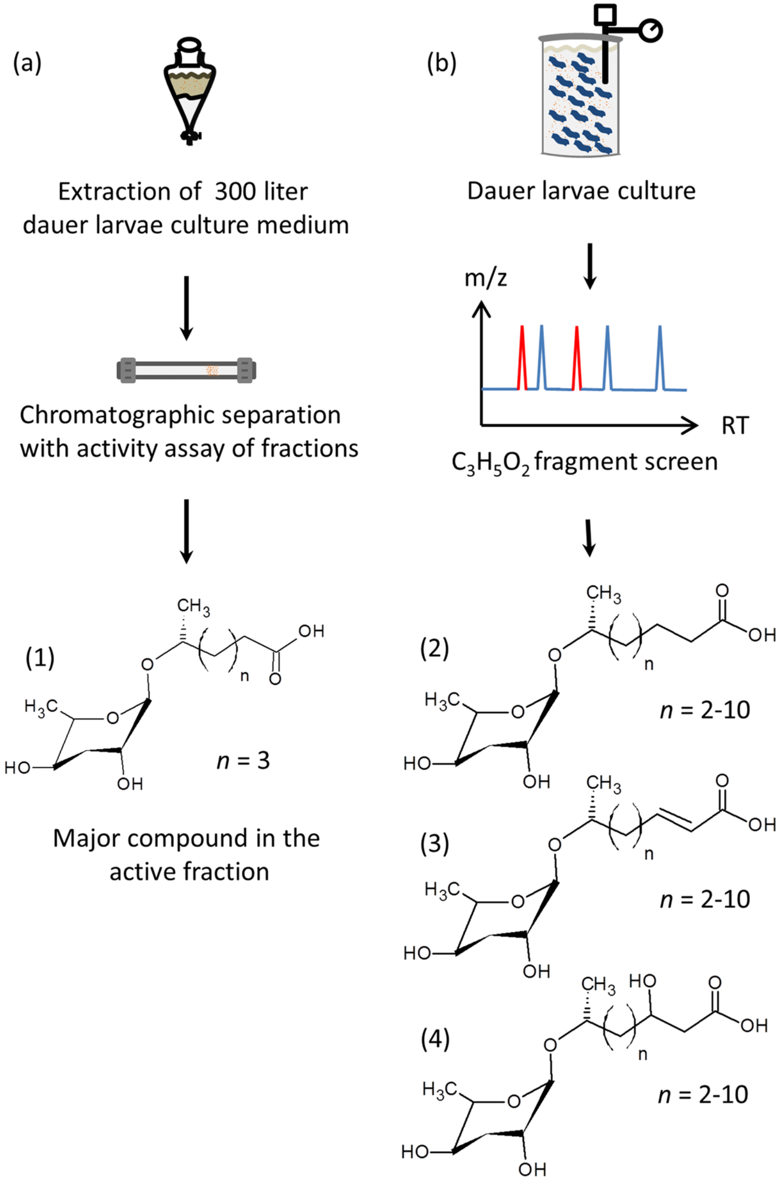

Figure 1.4 Isolation and identification of the first-discovered dauer-inducing pheromone (a) and discovery of further ascarosides with dauer-inducing activity (b).

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

class containing homologues of different side chain length (Figure 1.4b). In particular, representatives of **3**, with eight carbons in the side chain, were orders of magnitude more potent as well as more abundant than the original ascaroside **1**.[32] The blend of ascarosides acting as a dauer-inducing pheromone was named daumone.

Sophisticated reverse genetics experiments have shown that daumone binds to the specific G-protein coupled receptor at the cilia of *C. elegans* chemosensory neurons, leading to inhibition of TGF-β expression through a cascade composed of guanylyl cyclase and heat shock factor 1. It was further shown that TGF-β produced in the chemosensory neurons affects development through binding to an appropriate receptor on the somatic cells. Intracellular TGF-β signaling converges on a nuclear hormone receptor DAF-12, promoting reproductive development. However, DAF-12 deficient animals are unable to form dauers, even under unfavorable conditions, suggesting that DAF-12 itself promotes the expression of dauer genes and TGF-β regulates the production of a certain DAF-12 inhibitor. A member of the P450 oxidase family was identified as a potential enzyme involved in the production of such a factor. The nature of this factor was suggested as being a sterol, because the presence of cholesterol in the culture medium is an essential requirement for reproductive development of these roundworms. *C. elegans* are not able to synthesize cholesterol themselves and in its absence they form dauers. The arrest of reproductive development can be overcome by adding lathosterol to the medium,[33] suggesting that a lathosterol derivative may be involved in DAF-12 regulation. This is plausible since ligands of known homologs of DAF-12, such as mammalian liver X receptor or retinoid receptor, are lipids. In an attempt to identify the DAF-12 ligand a group from Texas first conducted a reporter-based assay with the known ligands of nuclear hormone receptors using a DAF-12/GAL-4 promoter. They showed that 3-keto lithocholic acid could induce the reporter expression and suggested that the real DAF-12 ligand should contain a 3-keto group as well as a terminal carboxygroup. A mixture of lathosterone, which already contains a 3-keto group, with microsomes containing the P450 oxidase, supported the reproductive development of corresponding oxidase mutants. Comparison of LC-MS chromatograms of lathosterone mixed with loaded microsomes or with empty microsomes showed the presence of an additional chromatographic peak if the P450 oxidase was present in the mixture (Figure 1.5). Mass differences between the original lathosterone and the new compound indicated the presence of a carboxyl group, probably at the end carbon of the side chain, analogous to 3-keto lithocholic acid.[34] This new DAF-12 ligand was named Δ7-dafachronic acid and its *in vivo* presence in *C. elegans* was further confirmed using NMR.

Thus, TGF-β promotes the reproductive development of *C. elegans* through induction of dafachronic acid synthesis genes. After synthesis from cholesterol, dafachronic acid binds to the DAF-12 nuclear receptor, preventing its translocation to the nucleus. Under unfavorable conditions daumone accumulation leads to inhibition of TGF-β production in the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00001

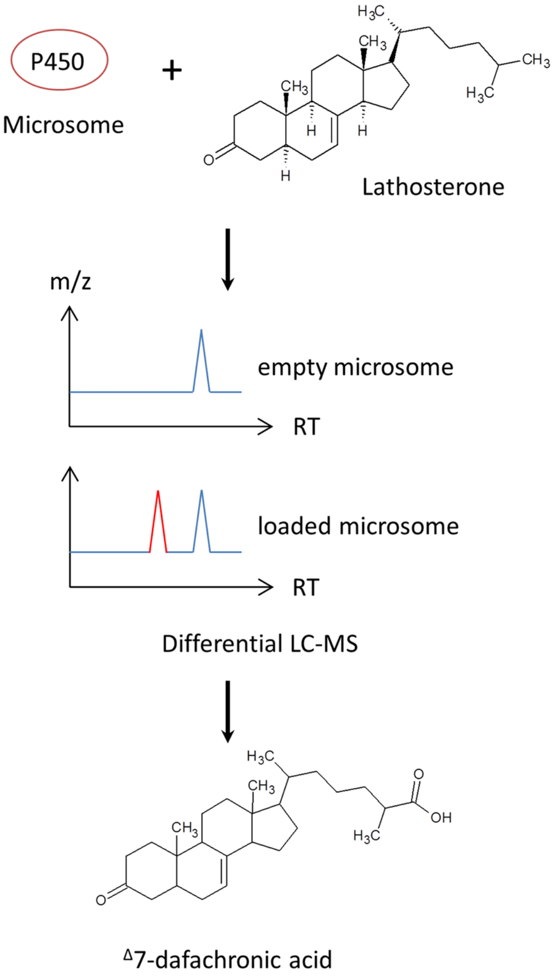

Figure 1.5 Identification of lathosterone derivative Δ7-dafachronic acid.

chemosensory neurons, which results in decreased dafachronic acid production in somatic cells, nuclear translocation of DAF-12 and expression of dauer-inducing genes and the arrest of the reproductive life cycle. Such a complicated regulatory mechanism ensures that the worms only continue reproductive development when there is a sufficient food source to ensure survival of the next generation. This elegant mechanism, which would not have been discovered without the use of MS, brilliantly demonstrates its use in chemical biology. Novel developments in the field of MS and related techniques will be discussed in further chapters of this book and they will hopefully contribute to many more intriguing discoveries, and to the further development of chemical biology.

References

1. A. Butenandt, *Angew. Chem.*, 1960, **18**, 643.
2. H. Brockmann, *Hoppe-Seyler's Z. Physiol. Chem.*, 1936, **241**, 104.
3. H. M. Evans and K. S. Bishop, *Science*, 1922, **56**, 650.
4. H. M. Evans, in *Vitamins and Hormones*, ed. R. S. Harris and I. G. Wool, Academic Press, New York and London, 1962, vol. 20, p. 379.
5. H. M. Evans, O. M. Emerson and G. A. Emerson, *J. Biol. Chem.*, 1936, **113**, 319.
6. E. Fernholz, *J. Am. Chem. Soc.*, 1938, **60**, 700.
7. A. J. P. Martin, *Nobel Lecture*, 1952.
8. A. J. P. Martin and R. L. M. Synge, *Biochem. J.*, 1941, **35**, 1358.
9. A. H. Gordon, A. J. P. Martin and R. L. M. Synge, *Biochem. J.*, 1942, **37**, 86.
10. R. L. M. Synge, *Biochem. J.*, 1948, **42**, 99.
11. R. Consden, A. H. Gordon and A. J. P. Martin, *Biochem. J.*, 1947, **41**, 596.
12. G. A. Howard and A. J. P. Martin, *Biochem. J.*, 1949, **46**, 532.
13. A. Butenandt, R. Beckmann and E. Hecker, *Hoppe-Seyler's Z. Physiol. Chem.*, 1961, **324**, 71.
14. A. T. James and A. J. P. Martin, *Biochem. J.*, 1951, **50**, 679.
15. I. Jarchum, *Nat. Methods*, 2015, **12**, S5.
16. S. Bergström and J. Sjöval, *Acta Chem. Scand.*, 1960, **14**, 1701.
17. S. Bergström, R. Ryhage, B. Samuelsson and J. Sjövall, *J. Biol. Chem.*, 1963, **238**, 3555.
18. N. Mirsaleh-Kohan, W. D. Robertson and R. N. Compton, *Mass Spectrom. Rev.*, 2008, **27**, 237.
19. W. C. Wiley and I. H. McLaren, *Rev. Sci. Instrum.*, 1956, **26**, 1150.
20. W. C. Wiley, *Science*, 1956, **26**, 817.
21. R. S. Gohlke and F. W. McLafferty, *J. Am. Soc. Mass Spectrom.*, 1993, **4**, 367.
22. E. C. Horning and M. G. Horning, *Clin. Chem.*, 1971, **17**, 802.
23. O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey and L. Willmitzer, *Nat. Biotechnol.*, 2000, **18**, 1157.
24. K. Biemann, in *Fortschritte der Chemie organischer Naturstoffe*, ed. L. Zechmeister, Springer Verlag, Wien New York, 1966, vol. 24, p. 2.
25. K. Biemann, C. Cone, B. R. Webster and G. P. Arsenault, *J. Am. Chem. Soc.*, 1966, **88**, 5598.
26. M. Dole, L. L. Mack, R. L. Hines, R. C. Mobley and L. D. Ferguson, *J. Chem. Phys.*, 1968, **49**, 2240.
27. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse, *Science*, 1989, **246**, 64.
28. T. Faust, *Nat. Methods*, 2015, **12**, S10.
29. C. G. Horvath, B. A. Preiss and S. R. Lipsky, *Anal. Chem.*, 1968, **39**, 1422.
30. P. Y. Jeong, M. Jung, Y. H. Yim, H. Kim, M. Park, E. Hong, W. Lee, Y. H. Kim, K. Kim and Y. K. Paik, *Nature*, 2005, **443**, 541.
31. S. H. von Reuss, N. Bose, J. Srinivasan, J. J. Yim, J. C. Judkins, P. W. Sternberg and F. C. Schroeder, *J. Am. Chem. Soc.*, 2012, **143**, 1817.

32. R. A. Butcher, M. Fujita, F. C. Schroeder and J. Clardy, *Nat. Chem. Biol.*, 2007, **3**, 420.
33. V. Matyash, E. V. Entchev, F. Mende, M. Wilsch-Bräuninger, C. Thiele, A. W. Schmidt, H. J. Knölker, S. Ward and T. Kurzchalia, *PLoS Biol.*, 2004, **2**, e280.
34. D. L. Motola, C. L. Cummins, V. Rottiers, K. K. Sharma, T. Li, Y. Li, K. Suino-Powell, H. E. Xu, R. J. Auchus, A. Antebi and D. J. Mangelsdorf, *Cell*, 2007, **124**, 1209.

CHAPTER 2

Introduction to Mass Spectrometry Instrumentation and Methods Used in Chemical Biology

P. HERRERO[a], A. DELPINO-RIUS[a], M. R. RAS-MALLORQUÍ[a], L. AROLA[b] AND N. CANELA*[a]

[a]Universitat Rovira i Virgili, Centre for Omic Sciences (COS), Group of Research on Omics Methodologies (GROM), Av. Universitat 1, Reus, 43204, Spain; [b]Centre Tecnològic de Nutrició i Salut (CTNS), Av. Universitat 1, Reus, 43204, Spain
*E-mail: nuria.canela@urv.cat

2.1 Introduction to Mass Spectrometry (MS)

MS is a powerful analytical technique that is used to determine the molecular mass of compounds by their conversion to gas-phase ions that are then characterized by their mass to charge ratios (m/z) and abundances. MS provides information to quantify known molecules, identify unknown compounds and elucidate the structure and chemical properties of different molecules, including proteins, peptides, nucleotides, carbohydrates, lipids, and many other biologically relevant molecules.

The history of MS began with the discovery of the electron in 1897 by J. J. Thomson from his studies on the conduction of electricity by gases.

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

Later, in 1907, he constructed the first MS that could determine the mass of particles by measuring how far they were deflected by a magnetic field.[1] These initial instruments were improved in the 20th century, amassing numerous advances that were awarded Nobel Prizes. Furthermore, highlighted as essential hits for their application in biomolecular studies, the development of soft ionization methods, such as electrospray ionization (ESI) by Yamashita and Fenn in 1984 to ionize biomolecules,[2,3] matrix-assisted laser desorption/ionization (MALDI) to ionize intact proteins by Tanaka in 1987,[4] and the Orbitrap MS by Makarov in 1999,[5] led to many types of laboratory equipment with performances that had never been observed before.

This chapter offers an overview of current MS instrumentation focusing on the sample introduction methods, ionization sources and mass analysers used in the different MS-based applications for biomolecular analysis.

2.1.1 The MS

An MS instrument accomplishes three essential functions: it generates multiple gaseous ions from the sample, separates the ions according to their specific mass to charge ratios (*m*/*z*) and finally records their relative abundances. These tasks are basically associated with the following three major components of the instrument: the ionization source, the mass analyser and the detector. Moreover, a sample introduction system to the ion source and software to control the instrument and process all of the acquired data are needed. It is important to note that the analyser, detector and some ionization sources are operated under high vacuum conditions (10^{-5} to 10^{-8} torr) because the gas-phase ions are highly reactive. Figure 2.1 shows a simplified representation of an MS.

The introduction of the sample into the MS can be performed with direct insertion probes; however, a prior separation step has become routine analytical practice for complex samples. Coupling MSs with liquid chromatography (LC), gas chromatography (GC) or other separation techniques tailored

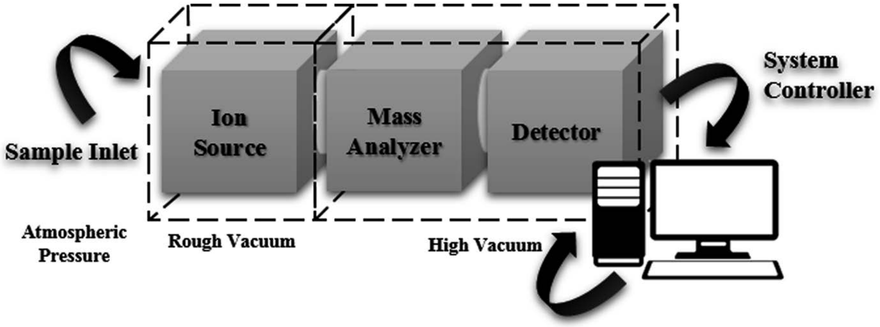

Figure 2.1 Building blocks of an MS.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

to different ionization sources has contributed to the development of new methodologies for biomolecular analysis.

Currently, a wide variety of ionization techniques are available, including ESI and MALDI, which have become exceptional tools in the biological sciences for the analysis of highly polar and large biomolecules, such as peptides, nucleic acids and other organic compounds. MS analysis can be performed using several types of instruments, hybrid MSs that combine more than one mass analyser being the most widely used. Finally, the ions emerging from the analyser are detected electronically, and the resulting information is stored and analysed in a computer, displaying the results on a chart or mass spectrum, which plots the ion abundance against the m/z.

In the next sections, the fundamentals, capabilities and different components of an MS that can be applied to biomolecular analyses are described.

2.2 Sample Introduction and Separation Methods for Coupling to MS

Different methods are available for introducing the sample into an MS and among them direct infusion is widely used due to its short time consumption and high throughput capabilities. However, there are some limitations for complex sample analysis, such as signal suppression or enhancement caused by the matrix components and the inability to resolve isobaric species.

Combining MS with a previous separation method, such as chromatography, either with an on-line or off-line configuration, can solve these problems. Furthermore, on-line separation techniques offer the advantage of increasing the high throughput of the analytical methodology.[6] The following subsections are focused on the most used separation techniques, including LC, GC, supercritical fluid chromatography, and electric-field driven separations, such as capillary electrophoresis (CE).[7]

2.2.1 LC

In LC, the mobile phase is a liquid, and the stationary phase typically consists of small porous particles with a large surface area packed into a column coated with different bonded phases that interact chemically and/or physically with the sample molecules. Currently, high-performance LC (HPLC) is the main LC separation technique used in combination with MS where the operational pressures are up to 400 bar.

Ultra-performance LC, first introduced in 2004 by Waters Corporation, and ultra-high-performance LC are both an evolution of HPLC, which allow working at higher pressures (up to 1000 bar), permitting the use of sub-2 μm particles, along with an increased separation efficiency at higher flow rates, enabling an equal or even better chromatographic resolution with shorter run times.[8,9]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

The necessity of high sensitivity and chromatographic resolution to determine a wide range of compounds at very different concentrations in complex biological samples has revealed capillary and nano-LC coupled with MS to be the best-suited solution to reduce matrix effects. Nano-LC is the maximum miniaturization of an LC system and is widely used to analyse complex digested samples in proteomic applications. Due to the smaller column diameter (75 µm), this technique leads to very small chromatographic dilution, increasing the on-column concentration of eluted compounds and, subsequently, the sensitivity of the whole LC-MS analysis.[10–13]

In general, there are some considerations when LC is coupled with MS. The column eluent needs to be compatible with the ionization sources typically used in LC-MS, such as ESI or atmospheric pressure chemical ionization (APCI). The column diameter and mobile phase flow rate must be reduced enough to allow for the desolvation of mobile phase droplets containing the molecules that will be ionized. In some cases, the flow to the MS can be reduced by splitting it to waste or another detector. Another important issue is the substitution of non-volatile additives in the mobile phases, and replacing LC ion suppression agents such as TFA, which is a popular ion-pairing LC agent but bad for MS. The most common additives used in LC-MS are formic acid, acetic acid and ammonium formate or acetate.

Figure 2.2a shows a schematic diagram of an HPLC instrument that typically includes solvent reservoirs, a degasser, pumps, a mobile phase mixer, a thermostatized sample injector and a column compartment. Additionally, the configuration can include alternative detectors between the column oven and MS, such as absorbance, light scattering or fluorescence detectors, since they are non-destructive.

There are different types of separation methods available for LC depending on the nature of the molecules that will be determined. Thus, normal-phase LC (NPLC), which uses a highly polar stationary phase (*e.g.*, silica gel) and a

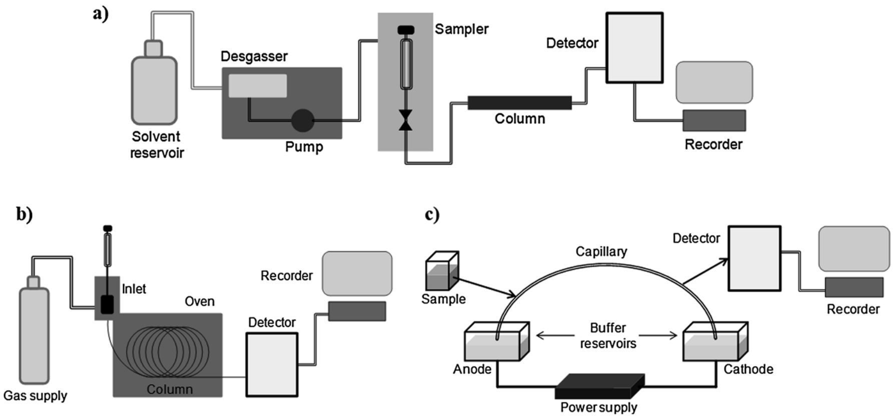

Figure 2.2 Separation system schematic diagrams of HPLC (a), GC (b) and CE (c).

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

non-polar (*e.g.*, hexane) mobile phase, is mainly applied for organic soluble molecules. Despite the utility of NPLC-based analysis, it remains a somewhat underused technique coupled with MS. In fact, the use of non-polar organic solvents may not provide optimal ESI ionization conditions due to its inability to supply ions for conjugation to the anlayte. Therefore some researchers have coupled NPLC separation with APCI, which permits the use of non-polar mobile phases, such as hexane or heptane, to enhance the ionization efficiency of non-polar compounds.[6,14]

The most popular LC mode is reversed-phase LC (RPLC), which uses columns packed with an inert non-polar material (*e.g.*, octadecyl carbon chain (C18), C8, cyano or phenyl), resulting in a hydrophobic stationary phase. The mobile phase is a polar mixture of water and water-miscible organic solvents (*e.g.*, methanol or acetonitrile). This chromatography has become the preferred separation technique to be coupled with MS because of its high mobile phase compatibility with ESI ionization, allowing the separation and determination of a wide range of analytes of varying polarities.

Hydrophilic interaction LC (HILIC) is useful for separating polar compounds. The stationary phases are polar sorbents, such as silica, amino, amide or cyano, but the mobile phase composition is similar to that used in RPLC. The separation mechanism in HILIC is influenced by the partitioning of the analyte between the layer of the aqueous mobile phase modifier onto the adsorbent surface and the organic modifier (frequently acetonitrile) of the mobile phase.[15] HILIC is suitable for polar compounds that are not retained in RPLC or are not soluble in the solvents required for NPLC. However, the required buffer concentration reduces the ionization efficiency of MS.

Ion exchange chromatography is based on interactions between charges. In cation exchange chromatography, positively charged molecules are attracted to the negatively charged stationary phase. Conversely, in anion exchange chromatography, negatively charged molecules are attracted to the positively charged stationary phase. Then, by increasing the ionic strength of the mobile phase using a salt concentration gradient, the molecules elute from the column in order, from the weakest to the strongest ionic interaction.[16]

2.2.2 GC

GC is broadly used to increase the separation of volatile compounds depending on their volatility and polarity. Additionally, non-volatile molecules can be determined using a previous derivatization reaction. In GC, the mobile phase is an inert gas while the stationary phase is a layer of liquid or polymer on an inert solid support. Helium is the most commonly used gas, but nitrogen and hydrogen are also employed. Some stationary phases used in GC are diphenyl, cyanopropylphenyl, dimethyl polysiloxane and polyethyleneglycol, which present different polarity and selectivity characteristics.[17]

GC columns can be either packed or capillary. Packed GC columns are stainless steel tubes filled with the stationary phase with inner diameters of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

up to 5 mm and lengths from 1 to 5 m, and have higher sample capacity compared to capillary columns; however, they are only used in a few applications due their decreased separation efficiency. Capillary or open tubular columns consist of several metres (up to 60m) of a fused silica tube with internal diameters of only 0.18–0.53 mm with inner walls coated with a thin 0.1–5 µm film thickness. Three types of capillary columns can be distinguished: the support coated open tubular column with an inert wall that is coated with a solid support where the stationary phase is adsorbed, the porous layer open tubular (PLOT) column with an inert wall that is coated with a solid stationary phase based on an adsorbent or porous polymer, and the wall-coated open tubular (WCOT) column with an inert wall that is coated with a liquid stationary phase. While PLOT columns are suitable for a few specific applications, such as the analysis of gases and low molecular weight compounds, WCOT columns are the most-used capillary columns, with a wide range of applications in biomolecular analysis.

Moreover, the GC column dimensions affect the sample analysis performance, and narrower columns are the most appropriate for coupling with MS since the lower carrier gas flow needed is more compatible with the MS vacuum conditions. Additionally, microbore columns used with the fast GC technique provide shorter retention times, and taller and narrower chromatographic peaks, which increase the sensitivity, although a fast MS scanning speed (5 scans s^{-1}) is needed to accurately define the extremely narrow peaks obtained.

The GC instrument components are the injection device, the inlet system, the column oven and the detector. The injection device can be an autosampler, and several formats are available depending on the type of sample to be analyzed: (1) an automatic liquid sampler for biofluids or other liquid extracts; (2) a head-space autosampler for the volatile fraction of a liquid or solid sample; (3) a solid phase microextraction autosampler, which extracts the compounds on a fibre that are thermally desorbed into the GC inlet; and (4) a thermal desorption system for gaseous samples.

Figure 2.2b shows a GC system. The inlet is the component that provides the introduction of the sample into a continuous flow of carrier gas. Common inlet types are split/splitless when the sample is introduced into a heated inert liner through a septum. The sample is vaporized, and the carrier gas leads the sample in its entirety (splitless mode) or only a portion of the sample (split mode) into the column. The programmable temperature vaporizing inlet is used to introduce large sample volumes, up to 250 µl, into a capillary column. Then, the oven programme is applied, increasing the inlet and column temperature, allowing the sample to gradually evaporate, preventing the compound from thermally degrading above the boiling point.

The combination of GC with MS interfaced with electronic impact ionization has become a powerful technique, with high sensitivity and identification capability levels that are applicable to the analysis of small biomolecules such as aminoacids, carbohydrates and organic acids, among others.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

2.2.3 Capillary Electrophoresis (CE)

CE is a variant of classical electrophoresis in which analyte migration occurs in a capillary tube. In CE, the analytes migrate through electrolyte solutions under the influence of an electric field, and they can be separated according to their ionic mobility. Additionally, analytes may be concentrated or 'focused' by gradients in conductivity and pH. Usually, CE refers to capillary zone electrophoresis in which the sample is applied to a narrow zone (band) surrounded by an electrolyte or buffer.[18]

CE provides higher resolution separations than LC and GC since electroosmotic flow does not have a laminar flow motion, which strongly reduces peak bordering. However, CE has limited throughput and sensitivity compared to the aforementioned separation techniques due to the small injection volume (a few nL) in the system.[19]

Figure 2.2c shows a schematic diagram of a CE instrument that basically comprises a high-voltage power supply and two buffer containers in which the electrodes are immersed and connected by a capillary. The sample is introduced into the capillary by changing one of the buffer containers using the sample container and applying a positive or negative pressure depending on the system. Over the last decade, several imaginative methods for increasing the amount of analyte introduced into the capillary have been proposed to increase the sensitivity; however, reproducibility has been one of the major drawbacks of this technique.[20]

In the coupling of CE with MS detection, the capillary outlet is introduced into a modified ESI ion source that acts as one of the anodes using sheath flow assisted ionization. However, most of the separation media employed in CE are not fully compatible with ESI interfaces due to the use of non-volatile additives.

CE-MS has been applied in metabolomics for charged molecules such as amino acids and nucleotides,[21] or drug screening in urine.[22]

2.3 Ionization Methods

The function of the ion source is to convert the atoms or molecules from the sample to gaseous-phase ions and introduce them into the vacuum region of the MS. Ionization methods can be classified according to the type and origin of the obtained ions,[23] and their choice depends on the nature of the sample, the physicochemical properties of the analytes of interest and the type of information required.

If biomolecules are volatile and thermostable, the preferred techniques are those that can obtain the molecular ion directly from the gaseous samples. However, when biomolecules are prone to degradation from heating, desorption assisted ionization techniques are necessary to obtain molecular ions from liquid and solid samples. The most commonly used ionization sources in biological research include electron ionization (EI) and chemical ionization (CI) for gas phase samples and ESI and MALDI for condensed phase samples. Currently, there are many methodologies that are variants or combinations of these techniques, which will be detailed and described later with particular attention to ambient ionization techniques. In the following subsections, the ionization theory, instrumentation, advantages, drawbacks and applicability are detailed.

2.3.1 EI

EI, also known as electron impact ionization, is widely used for MS and was one of the earliest sources to be introduced. It was first described by Dempster in 1921[24] and later redesigned by Bleakney[25] and Nier[26,27] as a source on which the current models are still based. In EI (Figure 2.3a), an electron beam, usually created by heating a tungsten or rhenium filament, passes through the ionization chamber where the sample molecules are in the vapour phase at a reduced pressure. A spiral path of electrons is generated by collimating the electron beam with a magnetic field and is then accelerated through the action of a potential difference. The electron beam ionizes the molecule by removing an electron, obtaining a radical cation $[M]^{\bullet+}$, leaving the molecule to extensive fragmentation. Depending on the fragment lost during ionization, we can define the following two types of fragments: odd-electron radical cations and even-electron cations. Commonly, odd-electron ions are formed by direct cleavage and are generated through charge retention with the ejection of neutral molecules. By contrast, the even-electron cations are obtained from charge migration, where the radical and charge part ways.[28] The elevated number of fragments are responsible for the named EI as a hard ionization source because the potential difference applied to accelerate the electron beam is conventionally set to 70 eV (adopted as a standard for analytical applications) even though the ionization (ionization potential is

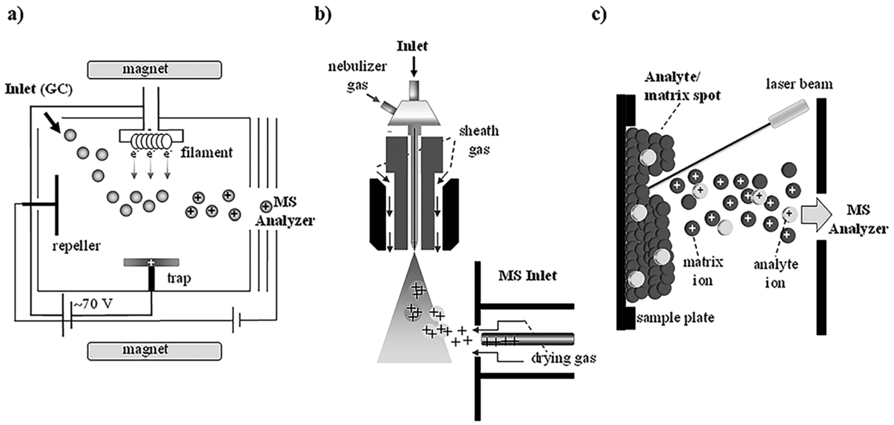

Figure 2.3 Ionization source representations for EI (a), ESI (b) and MALDI (c).

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

electron energy that will produce a molecular ion) only requires ~15 eV and bond cleavage only requires ~3–10 eV.[29] Note that decreasing the potential results in less fragmentation, but the number of ions that are formed will also decrease. Finally, a repeller electrode guides the electrons toward the mass analyser.

The sample to be ionized can be solid, liquid or gas; however, the sample must be volatile and can be introduced with heated batch inlet systems and heated direct probes and interfaces. The interface with the GC device is the most commonly used, but it has also been coupled with LC *via* a particle–beam interface.[7]

EI is considered to be a useful technique for analysing small molecules (<1000 Da), but the analytes must be neutral and volatile. To partially overcome these limitations, chemical derivatization was introduced to improve the volatility of some molecules. Another limitation of EI is that the fragmentation is so widespread that the molecular ion detection is usually inhibited. However, the spectra observed under standard conditions are reproducible and instrument independent; thus, large libraries have been created over the past 50 years. For these reasons, the compound spectrum can be considered to be a molecular fingerprint, which allows for compound identification by matching the obtained EI spectra against well-documented mass spectral libraries.

GC-(EI)MS has long been the established method for measuring volatile metabolites.[30,31] Nonetheless, GC-(EI)MS, using a previous chemical derivatization of analytes, has had and maintains a firm impact on targeted and untargeted metabolomics analysis, especially through the extensively available spectral libraries, which have facilitated the identification of diagnostic biomarkers and aids the subsequent mechanistic elucidation of the biological or pathological variations.[32]

2.3.2 CI

CI was developed as a technique to complement EI by Munson and Field in 1966,[33–35] and involves a two-step process that first covers the ionization of reagent gas molecules by an electron beam and then subsequent reactions occur between the reagent gas ions and analyte molecules. In this technique, the reagent gas should be at a pressure that is sufficient enough to obtain ionized plasma. Different reagent gases can be used depending on the analyte, such as butane, methane, water, methanol, ethanol or ammonia. There are various reactions involved in this type of ionization to control the formation of the principal molecular ion. The most common is a proton transfer reaction to produce protonated molecular ions $[M + H]^+$, which implies that the analyte has a higher proton affinity than the reagent gas.[23] As in EI, radical cation formation by the reaction of charged ions with gases with high ionization potential is rarely obtained; however, in CI, the resulting cation contains less energy than in EI. It is important to highlight that in this ionization the fragmentation is much lower than in EI and CI can be considered

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

to be soft ionization.[7] GC is the predominant sample introduction technique, as was described for EI. The advantage of CI is that by choosing the adequate gas reagent it is possible to selectively ionize a targeted compound class that is contained in a complex sample mixture. The main drawbacks of CI are that it is restricted to limited families of compounds and that the sensitivity is approximately one order of magnitude lower compared to EI due to different competing chemical reactions. The low sensitivity is the main reason why GC-(CI)MS is scarcely reported for biomolecular studies such as metabolomics.[32]

2.3.3 ESI

Over the last two decades, ESI has become one of the most important ionization techniques in the fields of chemistry, biochemistry and materials science. In 1968 Dole and co-workers first showed that ESI could be used to generate molecular ion beams from large molecules in polymer chemistry. In 1984, ESI was first reported as being coupled with MS by Yamashita and Fenn,[2] and, in 1988, Fenn[36] proved that it was possible to transfer large ionized molecules to the gas phase without breaking them using a multiple charging process, resulting in lower *m*/*z* values for the resulting ions, which were easily detected in the mass range of the most common mass analysers (*m*/*z* of less than 2000). These advances allowed for the systematic analysis of proteins using modern MSs and led to the emergence of proteomics, which is one of the fastest growing research areas in the chemical sciences. Its importance was recognized by the 2002 Nobel Prize in Chemistry given to John Fenn, who said that "A few years ago the idea of making proteins or polymers 'fly' by ESI seemed as improbable as a flying elephant, but today it is a standard part of modern MSs".[37]

Although, initially, ESI had a greater impact on the ability to analyse large multi-charged molecules, such as proteins and polymers, its application was later extended to a broad range of analytes that were preferably charged, and polar and ionisable molecules. Currently, ESI coupled with MS has become the most common technique for liquid samples because it is capable of ionizing both very small and extremely large biomolecules with a wide range of polarities in complex biological sample mixtures.

In ESI (Figure 2.3b), the sample, which is preferably solubilized in a polar solvent, is infused under atmospheric pressure into the source *via* stainless steel capillary. A high voltage, 1–5 kV, is applied to the extreme of the needle, and a positive or negative potential is applied between the needle and nozzle, causing the sample solution to disperse into an aerosol formed by highly charged droplets. The magnitude of the electric field is closely related to the capillary diameter and its distance from the sampling cone. The efficiency of nebulization is assisted by a concentrically applied gas, namely the sheath gas. The droplets are then driven electrically to the desolvation chamber in the atmospheric pressure interface where they shrink due to solvent evaporation, which is aided by a warm neutral gas

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

(commonly N_2). Finally, the droplets are introduced into the high vacuum region through a tiny orifice or skimmer. There are two major theories that are currently accepted to describe the ionization mechanism of electrospray: the ion evaporation model[38,39] and the charge residue model.[40] The first suggests that the droplet reaches a certain radius (the Rayleigh limit) where the charge repulsion is greater than the surface tension resulting in droplet disintegration.[41] The second proposes that ion formation comes from the combination of droplet evaporation and fission until the droplet contains only one or fewer analyte ions. According to the literature, the ion evaporation model is valid for ions with low m/z values, while the charge residue model would predominantly explain the behaviour of the ions with very high m/z values.[42]

Nanospray ionization arose to exploit the benefits of working at nano flow rates (1–500 nL min^{-1}) with electrospray sources. Nanoelectrospray ionization (nanoESI) provides high sensitivity due to the increased desolvation efficiencies that are achieved through the production of nanometre sized droplets.[43] Unfortunately, this high sensitivity is often accompanied by poor robustness and reproducibility, while conventional nanospray emitters/sources are prone to both clogging and spray irreproducibility, and sometimes require skilled operators. Similar to what happened at the beginning of the ESI era,[44] proteomics was the first chemical science to systematically operate and exploit the advantages of this technique.[45,46]

Sample introduction is a key point for ESI and it can be performed by direct flow injection, CE or LC and, as discussed above, the flow rate can vary from a few nL min^{-1} to 1 mL min^{-1}. Analyte solubility in a suitable solvent, rather than the volatility, becomes the most important factor in determining the likely success of the technique.[47] In addition, the concentrations and nature of the solvent and eluent additives can greatly affect the ionization. Neutral molecules that have at least one non-bonding pair of electrons (*i.e.*, alcohols, ketones, esters, amines, amides, carbohydrates, *etc.*) can be ionized taking a positive charge from the solution. Thus, in positive ionization, the analyte is sprayed at a low pH to favour protonation to obtain the ion $[M + H]^+$. By contrast, in negative ionization, the analysis is performed above the isoelectric point of the analytes to deprotonate them and obtain the $[M - H]^-$ ion. Normally, hydrogen is the most common adduct formed, but other chemicals, often in trace amounts, may form adducts too, such as sodium (Na^+), potassium (K^+) and ammonium ($NH_4{}^+$) in positive mode and chloride (Cl^-), formate (HCO_2H^-) and acetate ($CH_3CO_2H^-$) in negative mode. The adduct formation occurs by addition of additive in mobile phase such as, ammonium formate and acetate, ammonium acetate, formic and acetic acid, and alkali metal ions salts. Adducts can also be formed from mobile phase solvent, as the chloride adducts when chlorinated solvents such as chloroform are used. In addition, the low concentrations of sodium derived from glassware and storage bottles, or present as impurities even in analytical grade solvents, almost always lead to the presence of adducts in ESI studies.[48]

In summary, the benefits of ESI are the good ionization of polar, acid or basic molecules, the ability to perform multi-charged ionization of very large molecules, such as proteins or polymers, and the capacity to ionize labile biomolecules, and even non-covalent complexes, that can be detected without dissociation.[49] In addition, most polar solvents can be used, and it is an extremely gentle ionization method where the (de)protonated molecular ion is predominantly obtained. Nonetheless, ESI also presents some disadvantages because it is not useful for non-polar compounds (solved by APCI), its low fragmentation does not provide structural information, and the MS/MS spectral libraries are very poor compared to EI, and differ considerably among the instruments. Moreover, ESI is susceptible to ion suppression or enhancement caused by charge competition between the electrolytes and analytes,[7] which implies that quantification has to be performed carefully and accurately.[49]

Despite these limitations, ESI is a most versatile methodology, capable of separating and detecting numerous peptides and metabolites, as shown by the large number of LC-(ESI)MS methods developed over the last decade.[50,51]

2.3.4 APCI

APCI is based on theories of ionization in the gas phase and its hyphenation to MS was developed in 1973 by Horning[52] using a radioactive ^{63}Ni foil. APCI is gas-phase ionization in contrast to ESI, where the ionization occurs in the liquid phase. In APCI, the solubilized sample is infused into the source *via* a stainless steel capillary under atmospheric pressure, with a similar design to ESI. However, in this case the potential is not applied at the tip, but a combination of nebulizing gas and heating induces the formation of the spray while an auxiliary gas (N_2) minimizes the interactions of the analytes against the walls. Then, in the area of maximum aerosol formation, at the end of the heated region, a corona discharge electrode generates the chemical ionization. A combination of molecular collisions and charge transfer processes induces an ionized gas plasma of the solvent and then the analytes are ionized mainly *via* the transfer of protons resulting in $[M + H]^+$ or $[M - H]^-$ (de)protonated molecular ion formation, depending on the applied potential.[7] This ionization mode is very soft, and fragmentation is rarely observed. APCI is considered to be a complementary ionization technique to ESI because the molecules ionized by APCI are in the low-moderate mass range (less than 2000 Da) and have medium to low polarity.

Sample introduction can be accomplished *via* direct flow injection or interfacing with LC. In contrast to ESI, it does not work well at low flow rates ($<100\ \mu L\ min^{-1}$) being more suitable for high flow rates ($>750\ \mu L\ min^{-1}$).[43]

In APCI, the solvent choice plays an important role since the solvent affects the ionization process. However, compared to ESI, APCI results in less ion suppression or enhancement and allows for working at major volatile buffer concentrations.[49]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

2.3.5 MALDI

MALDI was first introduced in 1988 by Hillenkamp and Karas[53–55] and, simultaneously, Tanaka was able to ionize biological macromolecules, such as proteins, using the proper combination of the laser wavelength and matrix. For this work, he was awarded the Nobel Prize in Chemistry in 2002, shared with Fenn (see above). MALDI and ESI are the two soft ionization techniques most suitable for analysing high molecular weight and labile biomolecules. In contrast to ESI, where the ionization of molecules is produced directly from the liquid phase, in MALDI the biomolecules must co-crystallize with an ultraviolet-light-absorbing organic substance or matrix.

MALDI (Figure 2.3c) employs a laser to resonantly excite the molecules of the matrix and sample. The most commonly used lasers are ultraviolet lasers, such as a nitrogen laser (337 nm) and frequency-tripled and quadrupled Nd : YAG lasers (355 nm and 266 nm, respectively),[7,56] and, to a lesser extent, an infrared laser. With laser irradiation, the upper layer (~1 μm) of the matrix co-crystallized to the sample is ablated to produce a plume that subsequently forms protonated molecular ions $[M + H]^+$, although in some cases multicharged ions $[M + nH]^{n+}$ can also be obtained. Currently, the exact desorption mechanism of the analyte in MALDI is still unclear, and two approaches have been postulated. The thermal spike model suggests that ablation is the outcome of the sublimation of the matrix, and the pressure pulse theory involves the formation of a gradient of pressures on the sample surface. In addition, the mechanism associated with the proton-transfer process from the matrix to the samples also remains unclear, despite numerous studies reported in the literature.

The nature of the matrix plays a very important role and should have high molar absorptivity, small heat of sublimation, large proton affinity and gas-phase acidity. The matrix choice depends on the sample type, analyte class and the ionization polarity mode.[57,58] Organic acid matrices are the most commonly used, such as sinapinic acid, 2,5-dihydroxybenzoic acid, and α-cyano-4-hydroxycinnamic acid, being used for proteins, carbohydrates, biopolymers and peptides.[59] In addition, other matrices, such as ionic liquids,[60] proton stripping,[61] and inorganic substances,[62] have been used. Beyond matrix selection, another important factor for obtaining good MALDI mass spectra is the sample preparation process, which should accomplish a homogenous and reproducible matrix-analyte co-crystallization on the plate surface.

A relevant characteristic of MALDI is its capacity to support high concentrations of salts, buffers or other additives, useful for biological complex sample analysis, although it should be noted that it would be necessary to improve the signal-to-noise ratio of the spectra, spot homogeneity and reproducibility to become a comparable alternative to nanoESI-MS in the proteomics field. The mass-resolving power and ion signal intensities in MALDI are highly dependent on laser fluence, which is continuously adjusted during data acquisition to be close to the ion production threshold level.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

The standard MALDI source operates under high-vacuum conditions, but an atmospheric pressure ionization MALDI source (AP-MALDI) has also been developed. The advantages of working at ambient pressure are the ability to readily exchange with other atmospheric sources, the capacity to be coupled to different types of MSs, and that the replacement of samples becomes a much easier and faster process. However, the sensitivity of MALDI coupled with a linear time-of-flight (TOF) analyser is still better when MALDI is used under high-vacuum conditions.[63]

2.3.6 Other Ionization Techniques

Among the aforementioned desorption ionization techniques applied in biomolecules, there are other less extended possibilities such as fast atom bombardment ionization[64,65] and secondary ion MS.[66–68] In addition, molecules can also be ionized under atmospheric conditions using minimal or no sample preparation by recently developed ambient ionization sources, where the samples remain in their original state. With these sources, the ionization occurs on the surface of the sample enabling direct analysis of body fluids, plant materials, tissues or even single cells. Ambient ionization approaches can be classified according to their intrinsic desorption/ionization mechanisms,[69] such as desorption ESI,[70] direct analysis in real time,[71] desorption atmospheric pressure photoionization,[72] atmospheric solid analysis probe[73] or by laser ablation ESI.[74] Otherwise, they are barely used in proteomics and metabolomics fields due to their low sensitivity because of the lack of sample purification or enrichment steps.

2.4 Mass Analysers

Mass analysers can be considered as the heart of an MS instrument. Their function is to separate the gas-phase ions that are generated in the ion source according to their mass-to-charge ratio and then the detector measures the ion intensities. This section describes the most widely used mass analysers in chemical biology applications, the physical principles that separate the ions and how the measurement of the ion intensities is performed by the detectors. Furthermore, hybrid mass analysers and tandem MS are also explained.[7,75–78]

In MS there are two main parameters that need to be evaluated before carrying out a biochemical measurement, mass resolution and mass accuracy, which are dependent on the selected mass analyser. Mass resolution can be defined as the ability of the MS to separate ions with a very small difference in their *m*/*z* values and mass accuracy refers to the uncertainty of the measurement of a specific *m*/*z* value. These two parameters often go together and high-resolution instruments (>25 000 FWHM) also provide high mass accuracy (<5 ppm).[79,80]

The following major benefits of high resolution MS (HRMS) over low resolution (LRMS) are clear: fewer isobaric interferences that produce fewer false-positive identifications and accurate quantification.[81–84] However, sensitivity can be the limiting factor for HRMS if the molecule of interest is at very low levels in complex matrices. Thus, triple quadrupole LRMS instruments are the best option for trace level determination while TOF or Orbitrap-based instruments are more often used for screening or non-target experiments. Nonetheless, most of the biochemical analysis can be performed in a multitude of mass analysers with different quantification and identification performance.[85–90]

Table 2.1 lists the different mass analysers and a summary of their performance based on the mean resolution, accuracy, dynamic range and sensitivity.

2.4.1 TOF Mass Analysers

TOF mass analysers (Figure 2.4a) separate ions with different m/z by their dispersion over time as they fly into a tube to reach the detector. Then, m/z values are obtained as a function of the time 'of flight'. The concept of how a TOF analyser can provide m/z values is one of the easiest principles to understand for beginners in MS because it is based on the law of conservation of energy.[91–93] Thus, if an ion is accelerated in an electric field, its initial electrical potential energy (which only depends on the charge of the molecule and the electric field) is converted to kinetic energy (which depends on the mass of the molecule and the velocity). Then, assuming that all of the ions start at the same distance from the detector and the initial velocity is equal to 0, the resulting v is inversely proportional to the square root of the mass. Thus, molecules with a lower m/z ratio reach the detector faster than the bigger ones.[7]

Table 2.1 Comparison of the most outstanding features of the available mass analysers.

Analyser	Resolution	Mass accuracy	Maximum m/z	Sensitivity	Dynamic range	Speed
TOF	High (25 000–50 000)	High (<2 ppm)	>100 000	Medium	Medium	Fast
E/B[a]	Ultra-high (>100 000)	High (<5 ppm)	<20 000	Medium	High	Slow
Q	Low (1000–2000)	Low (>100 ppm)	<2000	High	High	Fast
IT	Medium (5000–10 000)	Medium (>50 ppm)	<4000	Medium	Low	Fast
FTICR	Ultra-high (>10 000 000)	Ultra-high (<0.1 ppm)	<20 000	Medium	Medium	Slow
Orbitrap	Ultra-high (>500 000)	High (<2 ppm)	<4000	High	Medium	Medium

[a] Electric sector (E)/magnetic sector (B).

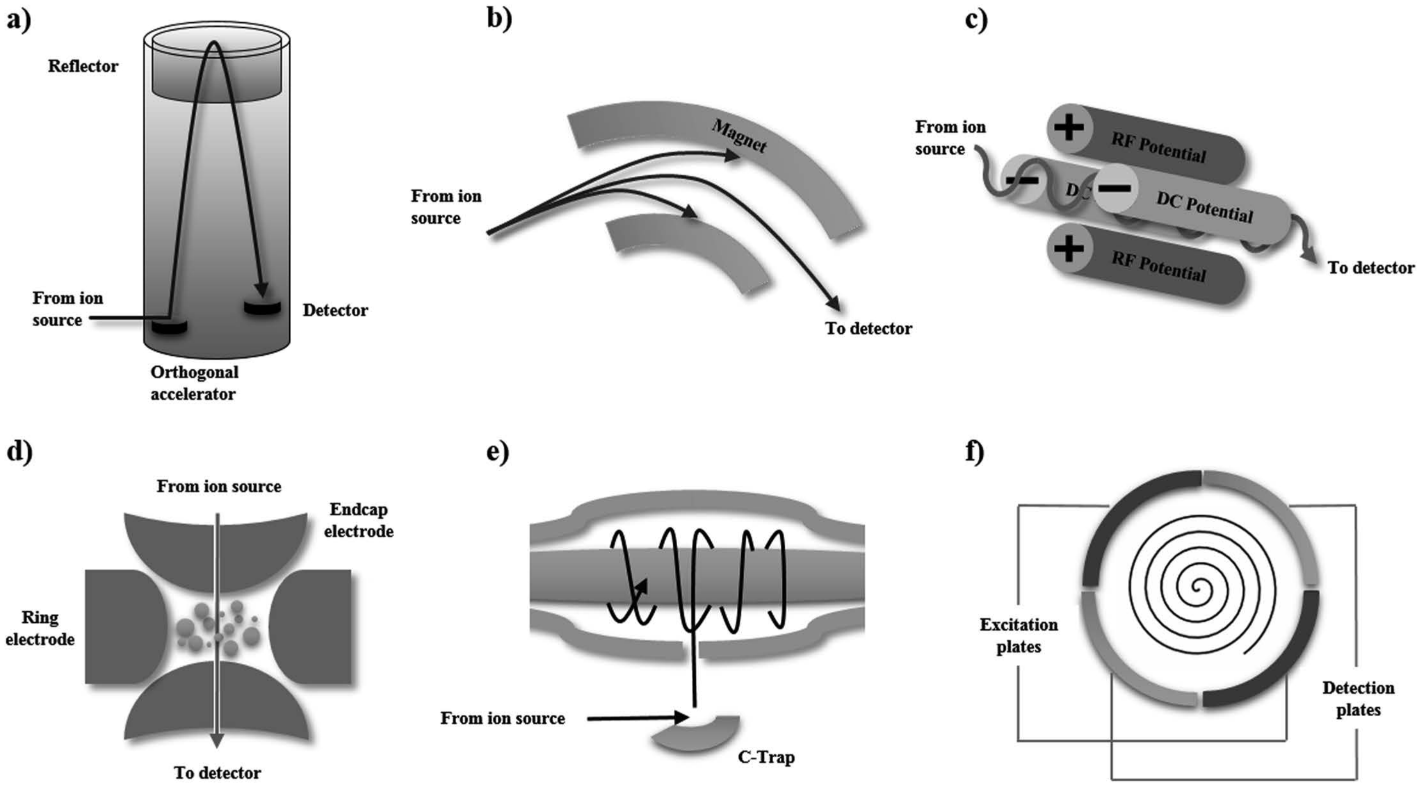

Figure 2.4 Mass analyser device diagrams for TOF (a), magnetic sector (b), quadrupole (c), ion trap (d), Orbitrap (e) and ICR cell (f).

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

The linear TOF analyser is the only mass analyser where neutral fragments cause a detector response because the metastable fragments formed during the flight conserve the velocity of the parent ion; therefore, they are detected at the same time as the precursor. These properties make the linear TOF analyser the ideal mass analyser for very labile molecules. However, the simple concept of a linear TOF analyser has two main drawbacks that limit their use. The first problem is the small difference in kinetic energy provided to the ions during the acceleration process, which strongly reduces the resolving power of the instrument. The other problem is the need for pulsed ionization to introduce a batch of ions into the mass analyser, which limits the TOF instruments to laser desorption based ionizations, such as MALDI.[92]

These disadvantages were overcome with reflectron TOF (ReTOF) and orthogonal acceleration TOF (oaTOF) instruments. The reflector was first proposed by Alikhanov[94] and was built in 1973 by Mamyrin;[95,96] currently, almost all TOF instruments have a reflector in their flight tube. ReTOF solves the poor resolution caused by different kinetic energies due to an ion mirror reflecting the ions, which also increases the flight path. To expand the TOF analysers to any ionization source (both pulsed and continuous) an orthogonal acceleration was proposed by pulsing the ions from a continuous ion beam coming from the ionization source. oaTOF can be linear or ReTOF; however, due to the better performance of ReTOF, most of them are oaReTOF instruments. However, the generic name TOF is used. Currently, most TOF analysers have a resolution of approximately 25 000 FWHM at 1000 *m*/*z*, and, in theory, they have an unlimited mass range; however, they usually work below 100 000 *m*/*z* for linear TOF and below 10 000 *m*/*z* for oaReTOF.

2.4.2 Magnetic Sector Mass Analysers

Magnetic sector mass analysers (Figure 2.4b) were the earliest instruments that were developed to separate ions with different masses. In the 1910s, Thomson used magnetic and electric fields for this purpose,[1] but it was not until the 1950s, with the development of reliable laboratory high vacuum and electronic technology, that these instruments started to be commercially available.[97] Magnetic sector analysers are based on forcing the ion beam from the source into a curved trajectory where high velocity ions pass through a perpendicular and strong magnetic field. Usually, an electric sector device is placed before the magnetic sector analyser (forward geometry) to enhance the resolution.[7,78,92,93,98]

Magnetic sector instruments couple excellently with all of the continuous ion sources (ESI, APCI, EI, CI, *etc.*) since the acquisition is also continuous but is not well suited to coupling with pulsed ion sources such as MALDI. Moreover, magnetic sector instruments provide high resolution, mass accuracy, and very high performance to separate two adjacent *m*/*z* signals to resolve the isobaric interferences when they are working in peak-matching mode. Nonetheless, these instruments are actually obsolete for biochemical applications because of their slow scan speed, high cost and large dimensions.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

They have been replaced by TOF, Orbitrap or Fourier transform ion cyclotron resonance (FTICR) mass analysers in many laboratories, depending on the required application.

2.4.3 Quadrupole

Quadrupole (Q) mass analysers act as a filter of masses that allow the passing of ions of a limited *m*/*z* range. Their development was published in 1953 by Morrison[99] and they consist of four parallel rods (cylindrical or hyperbolical) that are arranged symmetrically.[100] Then, a combination of direct current (DC) and radio frequency (RF) potentials is used to stabilize the travelling trajectory of the desired ion (*i.e.*, a given *m*/*z*) through the quadrupole analyser (Figure 2.4c). The ions that do not have a stable trajectory will collide with the rods and never reach the detector. The *m*/*z* range that the quadrupole allows to pass (*i.e.*, the resolution of the quadrupole analysers) depends on the ratio between the DC and RF potentials. Hence, the trajectory stabilization for heavy ions mainly responds to the DC potential of the field while for lighter ions their stabilization is mainly affected by the RF potential. These two components act as the lower cut-off and upper cut-off of the mass filter, respectively.[7,92,93]

The operation mode of a quadrupole is very simple. In scan mode, the DC and RF potentials are synchronized to stabilize a certain *m*/*z* value and the ions that reach the detector are counted, which gives the intensity. Since a quadrupole is a discrete mass analyser, its sensitivity in scan mode is limited by the scanning speed. However, when operating in selected ion monitoring, high sensitivity can be obtained for a single *m*/*z*.

A variation of the quadrupoles, which is present in the ion optics of all of the MSs, is the RF-only quadrupoles (without a DC component) that stabilize a wide *m*/*z* range and are then used as ion guides inside the spectrometer. Currently, the RF-only quadrupoles are actually hexapoles or octopoles because they have wider pass characteristics.

Moreover, they are ideal for continuous ion beam interfaces and for coupling with chromatographic devices. Currently, single-quadrupole analysers are limited to applications that demand lower performance and their use is limited to routine well-defined quantitative analysis. However, most of the tandem mass analysers have a quadrupole as the initial mass analyser because of their high performance and speed in filtering the ions before recording the MS/MS spectrum.

2.4.4 Ion Traps (ITs)

Linear quadrupole ion traps (LITs) or 2D ITs are RF-only quadrupoles in which two high-potential electrodes are placed at both ends of the quadrupole. These devices are capable of trapping the ions inside the quadrupole until they are selectively ejected to the detector in the order of ascending *m*/*z*.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

This type of IT is currently used to accumulate the ions before injecting them into another mass analyser such as an Orbitrap or FTICR.

Three-dimensional ITs are known as quadrupole ITs (QITs) or 'Paul's' traps, for which there was a shared Nobel Prize in Physics in 1989 for the invention.[100] The QIT is a type of QIT that uses static DC and RF to trap ions. This mass analyser consists of two 360° hyperbolic end-cap electrodes and a hyperbolic central ring electrode where the ions are trapped. To select an ion or scan over a *m*/*z* range, the RF frequencies are changed to eject the desired ions from the trap or eliminate the undesired ions by colliding them with the walls by destabilizing their trajectories (Figure 2.4d).[7,78,92,93,101]

One of the distinct features of ITs is that they are filled with gas (usually N_2) to decelerate the ions when entering the trap, also allowing the IT devices to perform a multi-stage fragmentation process (MS^n) in the same analyser. One important feature of the trapping device is that their performance is negatively affected by the presence of multicharged ions due to electrostatic repulsions, which strongly limits their use for the analysis of intact proteins.

2.4.5 Orbitrap

The Orbitrap was developed by Makarov in 1999 and is the last mass analyser to be conceived.[5] It is a modification of the Kingdom trap, which had an outer barrel-like electrode, a central spindle-like electrode along the axis and two end-cap electrodes. Between the central and outer electrode a logarithmic DC potential is applied, which triggers the ions to orbit around the central electrode. The orbital frequency is proportional to the inverse of the square root of *m*/*z* and ion detection is performed by image current detection and Fourier transform of the time-domain signal to a mass spectrum, similar to the FTICR mass analysers.[7]

The Orbitrap is a high-resolution mass analyser with better performance in terms of resolution (>100 000 FWHM at 200 *m*/*z*) and mass accuracy (<2 ppm) than ion traps, quadrupoles and time-of-flight instruments; Orbitrap is only surpassed by magnetic sectors and FTICR. Currently, the hybrid Orbitrap is one of the most powerful and versatile instruments available in many laboratories. Similarly to other ion trapping devices, it is a discontinuous analyser; however, it needs to be preceded by a C-shaped IT to store the ions coming from the ionization source before injecting them into the Orbitrap analyser. Moreover, this C-trap serves to accumulate the maximum quantity of ions to maximize the signal without losses in resolution and mass accuracy due to electrostatic repulsions in the Orbitrap. Orbitrap analysers have an *m*/*z* range between 50 and 4000, which is ideal for coupling to atmospheric pressure ionization sources to determine a wide range of biomolecules. However, is not useful for MALDI since the *m*/*z* range is too low for most peptides and proteins with single charged ions (Figure 2.4e).[5,102–104]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

2.4.6 Ion Cyclotron Resonance (ICR)

ICR analysers are based on the fact that a strong homogeneous magnetic field causes the ions to move in circular motion at a frequency that depends on their *m*/*z* ratio. Thus, ions are excited with an RF pulse to force them to move close to the detection plates, which induces a small current each time an ion passes by. Therefore, for a 'single ion', ICR can measure their circular frequency, which is unique and inversely proportional for each *m*/*z*.[7,92,93]

ICR instruments (Figure 2.4f) basically have two parts, a superconducting magnet and an ICR cell that is similar to a QIT since it traps the ions and places them inside the magnetic field (the superconducting magnet).

The principle of ICR was described in the early 1930s;[105] however, it was not until 1974, when Fourier transform (FT) was applied to ICR, that it became the top-performing MS with resolution and accuracy values that had never been observed before, and have not been achieved since for any other MSs. In non-FT instruments, the *m*/*z* is determined by scanning the excitation frequencies and counting the number of cycles necessary until the ions collide with the detection plate, which is also proportional to *m*/*z*. In FTICR, a wide amplitude RF frequency is applied to excite all of the ions in the cell and then a time domain image current is recorded and transformed to a mass spectrum *via* FT.

The FTICR mass analyser has the peculiarity that it is the only MS instrument in which the ion is not necessarily lost, since detecting the ions is not destructive and MS^n experiments can be performed by ejecting undesired ions from the cell.

2.5 Detectors

A detector on an MS is the device responsible for converting the energy of incoming particles from the mass analyser into a current signal that is recorded electronically, providing interpretable analytical information. In MS instruments, this means that the ion intensity can be interpreted as analyte abundance.

The overall procedure is that the charged molecules strike the detector and then the energy resulting from their impact generates the emission of secondary particles, usually electrons, which can be easily detected. Since the number of secondary particles that are emitted depends on the kinetic energy (velocity) of the impacting ions, these are often post-analyser accelerated to enhance the sensitivity.

As a general requirement, a detector should have high converting ion-electron efficiency, linear response, low noise and short recovery time. The MS detectors most currently used are described below.[7,92,93,106]

2.5.1 Faraday Cup

The Faraday cup is a metal cup placed in a vacuum system to intercept a beam of charged particles. The cup is an element of an electric circuit in which the measured current flow is directly proportional to the number of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

ions that have been collected. Although the Faraday cup is one of the oldest developed detectors, it is still in use in some isotope-ratio MSs allowing the precise measurement of mixtures of naturally occurring isotopes. Nonetheless, other detection devices based on electron multipliers (EMs) are preferred for biomolecule determination due to their better sensitivity.

2.5.2 EM Detectors

EM detectors are vacuum-tube structures that essentially multiply the number of incident charges *via* secondary emission on a metal or semiconductor material. Thus, the incident particles generate an emission of electrons that are strongly amplified using a successive series of secondary emission electrodes or dynodes, where the emitting electrons are accelerated to strike the next dynode, which produces a considerably increased number of electrons that reach the anode. This type of EM is known as a discrete dynode EM and it exists in numerous dynode geometries.

Another type is a channel EM, or channeltron, in which an electron cascade is produced in a continuous tube (single dynode) instead of multiples dynode. This type of EM is a very compact device that is also often used. Curved designs of channeltrons provide higher gains and an increased signal-to-noise ratio.

One of the most-used EM configurations in MS is microchannel plates (MCPs), which are a parallel array of millions of microdiameter linear channeltrons. To avoid ion losses due to the ions entering parallel to the microchannel surface they are slightly inclined. The major drawback of MCPs is the low gain compared to single EM devices because the electron cascade in one channel drains the adjacent channels, causing a saturation effect that results in a non-linear detector response. To obtain more gain, two or three MCPs are staked together with opposing microchannel angles.

Currently, EM detectors emit many secondary electrons per primary incoming particle, have a linear gain for high currents and have low thermionic emission, which results in low electric noise. These excellent features make EM the most common detector in MS instruments, making MCPs the detectors that are commonly used for TOF analysers and a single EM the most frequently used in triple instruments (QqQ).

2.5.3 Scintillator Detectors

Scintillation detectors are essentially a photomultiplier coupled with a scintillator material. Usually, this material is a phosphorous screen that releases photons as a response to the impact of the electrons originating in a previous dynode. The generated photons then pass into the multiplier where amplification occurs in a cascade mode, similar to EM devices. The main advantages of scintillator detectors are their long lifetime compared to EMs and their fast recovery time. Currently, scintillation detectors are also common detectors in MS instruments.

2.5.4 FT

FT is not exactly a detector device since it is a mathematical operation that decomposes as a function of time into the frequencies that make it up and provides valuable chemical information. FT measurements record an image current frequency, making them the only non-destructive detection method for MS that is currently used on FTICR and Orbitrap MSs.

2.6 Tandem MS (MS/MS)

MS/MS is essentially a technique that provides structural information through the fragmentation of intact parent ions that were previously isolated. Thus, multistage MS/MS (also referred to as MS^n) can be performed by successive isolation-activation processes for the desired precursor and fragment ions. While ion trapping devices (QIT, LIT and ICR) can perform multistage fragmentation, the other mass analysers cannot perform MS^n experiments. Hence, a wide variety of mass analysers have been coupled together, known as hybrid instruments, to perform enhanced acquisition modes and take advantage of the best ability of each mass analyser. The most common acquisition modes in tandem or hybrid instruments are selected reaction monitoring (SRM), product ion scan, precursor ion scan and neutral loss scan. These varieties of acquisition modes provide different analytical information and performance capabilities based on structural information to unequivocally identify unknown molecules (product ion scan), look for structure-similar compounds (precursor and neutral loss scan), or determine low detection levels for a known molecule (SRM).

There are two predominant categories of MS/MS, in-space and in-time. Tandem in-space MSs consist of at least two non-trapping mass analysers where, usually, the first one is used to isolate the precursor ion in a narrow *m*/*z* ratio and the second one is used to detect the full MS/MS spectrum (product ion scan) or to isolate a specific product ion (SRM) depending on the selected mass analyser. In tandem in-time MS, the ions are trapped, isolated, fragmented and scanned in the same physical analyser, which can only be performed using LIT, QIT and FTICR instruments. Note that the Orbitrap analyser is not a tandem in-time instrument, despite the fact that it is a trapping instrument, since fragmentation is performed outside the Orbitrap, which is the analyser itself.[7,92,93]

Over the last few years, some strategies to obtain MS/MS data without a precursor selection stage have been proposed for untargeted LC-MS-based metabolomics to improve metabolite identification capabilities. Hence, all ion fragmentation[107] and sequential window acquisition for all of the theoretical fragment ion spectra (SWATH) strategies are the most common examples.[108] In the former strategy, at least three different scans were performed for all of the ions entering the MS. One scan without fragmentation (MS level) and two more scans applying low and high collision energy to obtain MS/MS data. After this, advanced processing software deconvolutes the obtained raw files and provides a semi-empirical MS/MS spectrum that can be used for database metabolite matching. The SWATH strategy is similar to all ion fragmentation; however, in this case, a quadrupole mass filter makes many small precursor segments that have a wide mass band (25 m/z window, approximately) before they are fragmented and obtain MS/MS spectra for all of the ions present in the MS spectra. These types of fragmentation strategies for LC-MS-based metabolomics intend to emulate GC-(EI)MS where compound identification by spectral database matching is a well-resolved issue from a long time ago.

2.6.1 Hybrid Instruments

As mentioned before, hybrid instruments are MSs with at least two mass analysers between the ion source and the detector for performing enhanced acquisition modes. Table 2.2 lists the different possible combinations of two analysers and whether the combination exists in the market. Usually, the nomenclature of hybrid instruments is a combination of the acronyms of each mass analyser, for example, Q-TOF or Q-Orbitrap. However, there is an important feature of hybrid instrument nomenclature that has to be noted, the difference between ‘q’ and ‘Q’. The term ‘q’ refers to an RF-only quadrupole and is not used as a mass analyser since it only acts as a focusing device or as a part of the activation cell. Thus, QqQs are in fact a two quadrupole mass analyser hybrid instrument with an RF-quadrupole (q)-based collision cell between both Q analysers. The ‘q’ term is sometimes missed in the acronym, such as in Q-TOF, which is truly a QqTOF instrument.

In this section the principal hybrid instruments available for biomolecule determination are described. For most of them the first analyser is a filtering device such as a quadrupole (mainly) or an IT, while the second one is a high resolution MS to provide the best quality information for biomolecule identification and characterization. The exception is QqQ since it is based on obtaining better sensitivity to accurately quantify the biomolecules.

Table 2.2 Mass analyser combinations in hybrid MS instruments.[a]

MS2/MS1	E/B	Q	TOF	IT	FTICR	Orbitrap	IM
E/B[b]	✓	?	?	?	✗	✗	✗
Q	?	✓	✗	✗	✗	✗	?
TOF	?	✓	✓	✓	?	?	✓
IT	?	✓	✗	✓	✗	✗	?
FTICR	✗	?	✗	✓	?	✗	?
Orbitrap	✗	✓	?	✓	✗	?	?

[a]✓ available on the market; ? only described in the literature; ✗ not reported.
[b]Electric sector (E)/magnetic sector (B).

2.6.1.1 QqQ

QqQ is the combination of two quadrupoles mass analysers with a quadrupole-based collision cell between them. This combination of analysers provides the best sensitivity to determine known small biomolecules, such as metabolites and peptides, when it operates in SRM or multiple reaction monitoring (MRM) modes, which is a combination of multiple SRMs in the same analysis. This acquisition mode consists of selecting a known precursor ion in the first-quadrupole, fragmenting it in the 'second' quadrupole and finally selecting a specific *m*/*z* fragment in the third quadrupole, where the ion reaches the detector. This mode allows for obtaining a very selective signal, which permits the determination of molecules below ppt (part-per-trillion) levels in complex matrices. The main advantages of QqQ when it operates in MRM are their rapid duty cycle (10–50 ms) and high dynamic range (up to six orders of magnitude), which allows for the determination of multiple compounds in a wide range of concentrations and high-throughput analysis at ultra-low concentration levels.

The QqQ instrument can also operate as a single quadrupole instrument, which is necessary during MRM optimization using scan and product ion scan modes. Nonetheless, in these acquisition modes the sensitivity is much lower than in MRM. Another interesting acquisition mode exclusively for QqQ is the precursor ion scan, where the first quadrupole operates in scan mode while the third operates as a mass filter looking for a characteristic *m*/*z* fragment of a compound family, which is very useful for lipidomics.[6]

2.6.1.2 Quadrupole Time-Of-Flight (QqTOF) Analyser

The QqTOF analyser is the combination of a quadrupole mass filter with an oaTOF. Between Q and TOF there is a quadrupole-based collision cell. Thus, QqTOF instruments are similar in design to a QqQ instrument, but the third Q is replaced by a TOF analyser. This combination provides much better sensitivity when it works in product ion scan mode compared to QqQ instruments. However, as a tool to quantify, in terms of sensitivity and dynamic range, it is very limited compared to QqQ operating in MRM mode. These instruments are very useful for unknown analyses such as untargeted metabolomics, lipidomics and proteomics since it combines the high-resolution measurement of precursor ions for molecular formula determination, and high resolution MS/MS spectra for molecule identification. Another advantage of QqTOF instruments is that previous detailed knowledge of interesting molecules in a biological sample is not necessary because all of the ions are detected simultaneously and then the same acquired spectra can be reprocessed for use in future studies. As previously mentioned, QqTOF can perform all ion fragmentation and SWATH acquisition to enhance the throughput in untargeted experiments.

Recently, ion mobility MS (IM-MS) has been incorporated into many top-class instruments and adds a new level to the analyses of biomolecules, distinguishing them by 3D conformation.[109–112] Some researchers consider IM-MS to be an extra mass analyser where the ions first migrate through the IM drift tube

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

in a time that depends on the ion's size and shape, which determines the collision cross section, before these ions are separated by the *m*/*z* ratio in the mass analyser. The IM-MS is usually mounted in QqTOF instruments due to its high scanning rate compatibility with IM drift times and its main application fields are untargeted metabolomics and lipidomics. There are two different hardware configurations that exist for IM-MS instruments, including the IM cell just before the filtering quadrupole (IM-QqTOF) or after the collision cell (Qq-IM-TOF). The former configuration provides an IM drift for an intact molecule, while the later provides an IM drift both for the precursor and product ions.

2.6.1.3 *Tandem TOF (TOF–TOF)*

TOF–TOF instruments consist of the combination of a short linear TOF followed by a collision cell and a linear reflector TOF. This hybrid instrument needs to be combined with a pulsed ion source, usually MALDI, and the first TOF serves as a precursor selector based on the flight time of the precursor. Afterwards, the selected ion (based on time) is fragmented with a moderated energy and the resulting fragment ions are further accelerated before entering the reflector TOF analyser. This instrument configuration is very popular in proteomics, and, for many years, MALDI-TOF and MALDI-TOF-TOF were the gold-standard instruments for MS-based proteomics.

2.6.1.4 *Quadrupole Ion Trap (QqIT)*

The QqIT is a hybrid MS formed by the combination of a quadrupole mass filter and LIT. This type of hybrid MS is similar to QqTOF instruments, but it also has the high sensitivity of QqQ instruments since the linear trap can operate as a simple quadrupole, providing the same performance as a conventional triple quadrupole instrument with the same acquisition modes. When the QqIT operates as LIT, the full product ion scan can be recorded with medium–high sensitivity and with medium resolution performance, which makes it useful for some screening analyses but without the superior performance of QqTOF in this area of analysis. QqIT is also known commercially as the QTRAP® (AB Sciex).

2.6.1.5 *Orbitrap-based Hybrid Analysers*

Orbitrap analysers were introduced in 2005 with the first LTQ-Orbitrap, where an LIT (referred to by Thermo™ as an LTQ) was hyphenated with an Orbitrap mass analyser. In this hybrid instrument, an LIT was hyphenated with an Orbitrap mass analyser. The LTQ serves both as a mass filter and collision cell to perform MS^n at ultra-high resolution in the Orbitrap analyser. Due to the excellent capabilities of Orbitrap analysers, this underwent exponential development by combining different types of mass analysers and increasing the resolution (up to 500 000 FMHW), sensitivity (femtomoles) and scan speed (20 Hz), which has made Orbitrap-based instruments one of the most valuable instruments on the market today. Either with an LTQ or with a simple quadrupole before the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

Orbitrap, this instrument has become the new gold standard for proteomic analysis due to its excellent sensitivity and high resolution. Very recently, Thermo launched the first commercial trihybrid MS (Orbitrap Fusion®), combining in a single instrument a quadrupole, an LTQ and an Orbitrap mass analyser, which can perform at almost every MS acquisition modality. This feature, combined with multiple ionization and fragmentation methods, are also available, and makes the Orbitrap trihybrid MS the most versatile MS instrument to date.

2.6.1.6 Other Hybrid Mass Analysers

There are other hybrid mass analyser combinations; however, their use in biomolecular analysis is limited since they are more focused on an elemental analysis or their price is extremely elevated. Some examples are four-sector instruments, which combine two electric and two magnetic sector analysers in different configurations that are used principally for isotope and radioisotope ion detection. Another example is the QLIT-FTICR with high sensitivity, ultra-high accuracy and ultra-high resolution for MS^n experiments, which has been used over the years to solve very complex analyses. Nonetheless, new hybrid Orbitrap instruments have been demonstrated to be efficient enough for most complex bioanalytical measurements, becoming a more affordable option than QLIT-FTICR.

2.6.2 Fragmentation Devices

MS/MS requires a fragmentation device, where product ions originate depending on the ion activation method and the physicochemical properties of the molecule.

Additionally, some molecule fragmentation may occur without an activation device, as with in-source and post-source decay phenomena. In-source decay relies on an increase in the internal energy during the ionization process, resulting in some fragmentation of molecules before the ions enter the high vacuum region of the MS. Post-source decay is specific for MALDI-TOF instruments and refers to when fragmentation occurs after the acceleration region due to the collision of ions with residual gas inside the flight tube or the fragmentation of metastable ions that are stable enough to leave the source but contain excess energy, which results in fragmentation before reaching the detector.

The most common fragmentation methods currently used for biomolecule analysis are described below.

2.6.2.1 Collision Induced Dissociation (CID)

CID, also known as collision activated dissociation, is based on forcing ions to collide with gas atoms or molecules, usually N_2, Ar or He. This collision converts part of the ion kinetic energy into vibrational/rotational internal energy, which causes precursor ion fragmentation.[7,92,93,102] Essentially, two types of CID exist, low-energy CID (<100 eV), which is the common device in quadrupoles and IT, and high-energy CID (>1 keV), which is the common

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

fragmentation device in TOF–TOF and four sector mass analysers. This is not to be confused with higher-energy collisional dissociation (HCD), which is a low-energy CID technique specific to the Orbitrap MS in which fragmentation takes place externally to the trap. The term higher energy in HCD refers to the higher radiofrequency voltage applied in the C-trap to retain the fragment ions instead of the energy applied to induce fragmentation.

In high-energy CID (HE-CID) the product ion spectra are complex, presenting abundant low masses and internal fragment ions, while in low-energy CID (LE-CID) the spectra are dominated by low-energy fragments, which are more easily interpretable. This is one of the main reasons why LE-CID methods are the most used in MS instruments for biomolecule applications. Nonetheless, HE-CID has some advantages over LE-CID because the type of gas, pressure and temperature do not significantly change the obtained product ion spectra.

2.6.2.2 Electron Capture Dissociation (ECD)

The ECD process occurs when the precursor ion captures a low-energy electron (<1 eV) forming an excited radical species that rapidly dissociates and generates a fragmentation spectrum. One of the characteristics of ECD is that labile molecule groups are not fragmented and are useful for structure characterization since these labile groups dominate the MS/MS spectra in CID fragmentation. For example, phosphorylation or glycosylation post-translational modifications (PTMs) are retained using ECD, while peptide sequencing can be successfully performed by the amide-bond cleavage along the peptide sequence. However, this type of fragmentation is only available for FTICR instruments.[7,92,93,102]

2.6.2.3 Electron Transfer Dissociation (ETD)

The dissociation principle for ETD is almost the same as for ECD. Nonetheless, in ETD the electrons involved in the fragmentation process come from an anion molecule created by the CI of a polycyclic aromatic hydrocarbon, with fluoranthene the preferred compound. This configuration is currently available in most of the top-class MSs used for biochemical analysis, especially in hybrid TOFs and Orbitrap instruments for proteomics applications.

2.6.2.4 Photodissociation (PD)

The PD process takes place when precursor ions are activated by photons. There are two main variants of PD depending on the wavelength of incident light, ultra-violet PD (UVPD) and infrared multiphoton dissociation (IRMPD).[113] The main difference between UVPD and IRMPD, beyond irradiation energy, is the origin of the photons. In UVPD the photons come from a UV-lamp, while in IRMPD a CO_2 laser is used. The fragmentation pathways are similar to ETD and they have been proposed as alternatives for the characterization of molecules with labile groups and their appearance in

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

the analysis of protein PTM, glycans and glycolipids is expected. One of the advantages of PD is the fact that both the precursor and product ions may undergo photoactivation, resulting in a very rich spectrum.

2.7 Application of MS to Chemical Biology

MS is considered a fundamental technique for analysing biological samples and has evolved into an indispensable tool for detecting and identifying sample components by offering an exceptional assessment of structural elucidation, high sensitivity, reproducibility and a wide dynamic range. Nevertheless, to be totally accessible to the routine laboratory, powerful and easy to use bioinformatics tools are essential for processing the large volume of produced data.

2.7.1 Types of Biomolecules Analysed by MS

MSs are extensively used in industry and academia for both routine and research purposes. As a result of their accurate molecular mass measurements and high throughput capabilities, the major fields and applications for MS are the analysis of drugs in pharmaceutical companies (drug discovery, pharmacokinetics, drug metabolism), clinical screenings, forensics, biotechnology applications (proteomics, lipidomics, metabolomics, genomics), and environmental and safe-quality analysis (oil composition, air and water quality, food contamination).

2.7.1.1 Analysis of Oligonucleotides: Genomics

Oligonucleotides comprise deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and their analogues, which are linear polymeric sequences of nucleotides. These molecules contain a nitrogenous base, a ribose sugar and a phosphate group, and they commonly include some natural covalent modifications. Furthermore, a variety of DNA and RNA structures exist, including G-quadruplexes (G4s), that have been demonstrated to modulate gene expression levels both at the transcriptional and translational levels. Awareness of the structure–function relationships in nucleic acids requires advanced biotechnology instruments, including MS.

However, MS analysis is not commonly used for oligonucleotide characterization because of the buffers and ion-pairing agents that are traditionally used in the mobile phase for the chromatographic separation of these analytes leading ion suppression. Nonetheless, some LC-MS methods have been developed in the last few years to circumvent this apparent mobile phase incompatibility ranging from the use of relatively volatile or low levels of ion-pairing agents such as triethylamine and hexafluoroisopropanol,[114–116] to alternative chromatography such as HILIC,[117–119] weak anion-exchange

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

chromatography[120–122] or the use of porous graphitic carbon columns[123] that have also been proposed as alternatives, along with CE-MS[124–126] and MALDI-TOF.[127,128]

MS for nucleic acid research usually operates in negative ESI and plays an important role in the quality control of synthetic oligonucleotides, genotyping single nucleotide polymorphisms and short tandem repeats, characterization of modified DNA and RNA molecules, and the study of non-covalent interactions among nucleic acids as well as interactions with drugs and proteins to determine their structure and position in the oligonucleotide.[129–131] Moreover, native MS and IM-MS are employed to monitor nucleic acid assembly, study the interactions with other ligands, and characterize the affinity, specificity, and ligand binding mode.[132,133]

2.7.1.2 Analysis of Proteins: Proteomics

Proteins are linear polymers that results from combinations of the 20 amino acids connected by amide bonds. The term proteomics refers to the large-scale study of proteins and aims to characterize the whole proteome (the set of all proteins that are expressed by an organism). Proteomics provides the analysis not only for the expression, but also for the localization, function, and interaction of proteins.

NanoLC-MS allows for the accurate determination of the molecular mass of peptides and their sequences by performing MS/MS experiments and also provides quantitative information about their abundance, becoming the method of choice for proteomics experiments. In the last decade, the development of an Orbitrap HRMS and new dissociation methods such as ETD have promoted proteomics advances for the accurate detection of PTMs, determination of the number of disulphide bridges, monitoring the H/D exchange for protein folding and structural studies by protein–protein interactions, among others.

Depending on the existing knowledge of the evaluated proteins, proteomics can be performed in both targeted and untargeted ways (shotgun proteomics).[134] Nevertheless, proteomic strategies can be classified as bottom-up, middle and top-down, depending on whether the analyses are on proteolytic peptide mixtures, longer peptides or intact proteins, respectively.

Generally, the changes in protein abundances are responsible for functional alterations in a biology system and consequently quantitative methods, both absolute and relative, are becoming a principal MS application.[135–137] However, for a more complete characterization of functional proteomes, including isoforms, interaction partners and PTMs, a combination of different analytical strategies is recommended. Quantitative affinity purification MS is used for protein–protein interaction studies. A pre-enrichment step is helpful for identifying PTMs such as phosphorylation and glycosylation.[138–140] Additionally, the usage of stable isotope labelling strategies has allowed dynamic changes in the measurements for protein expression,

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

interaction and modification, which transforms MS into a more representative technique for biological studies.[141]

2.7.1.3 Analysis of Metabolites: Metabolomics

MS-based metabolome profiling is emerging as a method of choice for enhancing the understanding of metabolomics pathways in complex biological samples. Metabolomics focuses on the study of small molecule levels in a biological system and the dynamics of metabolic networks. Metabolomics may provide valuable information that supplements genomics, transcriptomics and proteomics, to which it is intimately related. Metabolites are the end products of the catalytic process regulated by the genome and proteome, and comprise the total low weight molecules that cellular activity leaves behind. Metabolites are more sensitive to biological perturbations and fluctuations occur rapidly making them good biomarkers for clinical diagnosis.[142,143]

Metabolomic strategies are also divided into two distinct approaches, targeted and untargeted. The latter accomplishes the measurement of complex metabolic profiles in biological samples;[144,145] it offers an opportunity for obtaining better biochemical pathways and cellular mechanism insight, and for the discovery of novel biomarkers. Conversely, targeted metabolomics analyses predefined groups of metabolites to accurately quantify them.[146]

The metabolome analysis is hampered by chemical diversity, variation in concentration range by several orders of magnitude, and some difficulties in the identification and characterization of metabolites. Consequently, no single analytical platform is able to measure all of the existing metabolites. Therefore, multiple approaches using different analytical platforms, such as LC-MS, GC-MS and nuclear magnetic resonance spectroscopy, are required to extensively cover the metabolome profile. Nevertheless, substantial improvements have been achieved recently, using HRMS instruments, such as Q-TOF and Orbitrap coupled with a wide-range of chromatographic systems.[85,147]

2.7.1.4 Analysis of Lipids: Lipidomics

Metabolomics methods typically involve the analysis of water-soluble metabolites, while lipidomics studies focus on analysing the lipid profile, although some overlap of the molecular species is observed. Lipids are an extensive family of small molecules that are soluble in organic solvents. Within recent years, increasing interest in lipids has been observed in biomedicine research, not only as energy storage compounds but also as key regulators in numerous cellular physiological and pathological processes such as cancer, diabetes and neurodegenerative diseases.

Lipidomics methods are much more well-established than metabolomics since lipids are biosynthesized from a combination of a defined set of head groups and fatty acid species, which makes their properties more uniform, and consequently their analysis was easier compared to aqueous metabolites.[148–155]

Using MS, it is possible to determine the molecular mass, elemental composition, and the nature and branching position of the substituents in the lipid molecule. HRMS allows for the identification of lipids by their exact mass, combined with lipid class-specific neutral loss or product ions. Moreover, GC and LC retention time data yields an extra level of specificity. The determination of fatty acids methyl esters (FAMEs) by GC-MS is among the most common analyses in lipid research. Triglycerides of fatty acids cannot be analysed directly by GC; they must first be hydrolysed and derivatized. The ester bonds are hydrolysed and the free fatty acids that are formed in the process are converted to the corresponding FAMEs. FAMEs are moderately apolar and sufficiently volatile to be determined by GC or GC/MS.[156–158] Other popular and powerful MS-based lipidomics technologies for identifying and quantifying individual lipid species comprise ESI sources and MS/MS, using QqQ analysers for targeted analysis and Q-TOF, FTICR or Orbitrap-based mass analysers for untargeted studies.[147]

2.7.1.5 Analysis of Glycans: Glycomics

Glycans or oligosaccharides are molecules that are formed by the association of several monosaccharides connected through glycosidic bonds. The characterization and elucidation of glycan structures is a difficult task that comprises monosaccharide sequence identification, and the branching pattern, isomer position and anomeric configuration of glycosidic bonds.

MS applied in glycomics and glycobiology is broadly used to analyse the glycan part of a large biomolecule previously cleaved either enzymatically or chemically. In the case of glycolipids, they can be analyzed directly without cleaving the lipid component.

Protein glycosylation is one of the most prominent PTMs, with an important role in numerous biological processes ranging from fertilization and immune response to cell–cell recognition and inflammation.[159–162] Nonetheless, proteomics studies are rarely focused on intact glycopeptides characterization; therefore, valuable information regarding the glycan structure and the glycosylation site is lost, limiting the complete understanding of the actual biological function of protein glycosylation.

The MS/MS analysis of glycans allows for their structure elucidation, paying special attention to the N-linked, O-linked, ganglioside and glycosaminoglycan compound classes. Recently, significant improvements have been made in the characterization of intact glycopeptides from enrichment and separation methods to MS detection. For example, intact glycans may be directly detected by ESI-MS[163–166] and MALDI-MS,[167,168] and glycoconjugate analysis can be performed on a wide range of MS instruments using either direct sample injection or coupled online to separation methods such as LC[169,170] and CE.[19,171,172]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

2.7.2 Other MS Applications: Imaging MS and Microorganism Identification

The identification and *in situ* localization of specific biomolecular species involved in pathological situations are still challenging for immunohistochemistry methods. Nowadays, emerging imaging MS (IMS)-based techniques seem to solve these issues.[6,173–177] IMS allows for the acquisition of *m*/*z* data and the visualization of the spatial distribution of biomolecules without extraction, purification, separation or labelling. IMS is applied for peptides, proteins, amino acids, carbohydrates, metabolites and lipids directly from tissue sections.[167,178–180] Importantly, IMS can be accomplished without prior knowledge of the tissue composition and the use of antibodies. IMS techniques have emerged for analysing biological samples, mainly tissue samples, and its versatility has opened up new opportunities in several fields, such as medicine, agriculture, biology, pharmacology and pathology.

A different MS application is the ability to identify microorganisms using MALDI-TOF instruments that provide high-speed and high-confidence identification, and taxonomical classification of bacteria, yeasts and fungi.[181–184] This fast process can determine the unique molecular fingerprint of a microorganism by measuring highly abundant proteins, such as ribosomal proteins, that are found in all microorganisms. The characteristic spectrum pattern of this molecular fingerprint is used to reliably and accurately identify a particular microorganism by matching it against a library. Biotyper and SARAMIS are commercially available software solutions that help researchers with routine clinical microbial identification, environmental and pharmaceutical analysis, taxonomical research, food and consumer product processing, quality control, and veterinary microbiology.

Abbreviations

APCI	Atmospheric pressure chemical ionization
C18	Octadecyl carbon chain
CE	Capillary electrophoresis
CI	Chemical ionization
CID	Collision induced dissociation
DC	Direct current
DNA	Deoxyribonucleic acid
ECD	Electron capture dissociation
EI	Electronic ionization
EM	Electron multiplier
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FAME	Fatty acid methyl ester
FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

G4s	G-quadruplexes
GC	Gas chromatography
HCD	Higher-energy collisional dissociation
HE	High energy
HILIC	Hydrophilic interaction liquid chromatography
HPLC	High-performance liquid chromatography
ICR	Ion cyclotron resonance
IM	Ion mobility
IMS	Imaging mass spectrometry
IRMPD	Infrared multiphoton dissociation
IT	Ion trap
LC	Liquid chromatography
LE	Low energy
LIT	Linear quadrupole ion trap
MALDI	Matrix-assisted laser desorption ionization
MCP	Microchannel plate
MRM	Multiple reaction monitoring
MS	Mass spectrometry/mass spectrometer
MS/MS	Tandem mass spectrometry
nanoESI	Nanoelectrospray ionization
NPLC	Normal-phase liquid chromatography
oaTOF	Orthogonal acceleration time-of-flight
PD	Photodissociation
PLOT	Porous layer open tubular
PTM	Post-translational modification
Q	Quadrupole
QIT	Quadrupole ion trap
QqQ	Triple quadrupole
QqTOF	Quadrupole time-of-flight
ReTOF	Reflectron time-of-flight
RF	Radio frequency
RNA	Ribonucleic acid
RPLC	Reversed-phase liquid chromatography
SRM	Selected reaction monitoring
SWATH	Sequential window acquisition of all theoretical fragment ion spectra
TOF	Time-of-flight
TOF–TOF	Tandem time-of-flight
UVPD	Ultra-violet photodissociation
WCOT	Wall-coated open tubular

Acknowledgements

We gratefully thank Miguel Ángel Rodríguez-Gómez for his reading and valuable comments, which helped to improve this chapter.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

References

1. J. J. Thomson, *Proc. R. Soc. London, Ser. A*, 1913, **89**, 1–20.
2. M. Yamashita and J. B. Fenn, *J. Phys. Chem.*, 1984, **88**, 4451–4459.
3. J. B. Fenn, *J. Biomol. Tech.*, 2006, **13**, 101–118.
4. K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida and T. Matsuo, *Rapid Commun. Mass Spectrom.*, 1988, **2**, 151–153.
5. A. Makarov, *Anal. Chem.*, 2000, **72**, 1156–1162.
6. N. Canela, P. Herrero, S. Mariné, P. Nadal, M. R. Ras, M. Á. Rodríguez and L. Arola, *J. Chromatogr. A*, 2015, **1428**, 16–38.
7. R. Ekman, J. Silberring, A. M. Westman-Brinkmalm, A. Kraj, D. M. Desiderio and N. M. Nibbering, *Mass Spectrom.*, 2015, **1**, 111–113.
8. R. Malviya, V. Bansal, O. Pal and P. Sharma, *J. Glob. Pharma Technol.*, 2010, **2**, 22–26.
9. N. Yandamuri, K. R. S. Nagabattula, S. S. Kurra, S. Batthula, L. P. S. N. Allada and P. Bandam, *Int. J. Pharm. Sci. Rev. Res.*, 2013, **23**, 52–57.
10. J. Hernández-Borges, Z. Aturki, A. Rocco and S. Fanali, *J. Sep. Sci.*, 2007, **30**, 1589–1610.
11. D. R. Jones, Z. Wu, D. Chauhan, K. C. Anderson and J. Peng, *Anal. Chem.*, 2014, **86**, 3667–3675.
12. M. R. Gama, C. H. Collins and C. B. G. Bottoli, *J. Chromatogr. Sci.*, 2013, **51**, 694–703.
13. N. Yandamuri and S. K. Dinakaran, *World J. Pharm. Res.*, 2015, **4**, 1355–1367.
14. M. Scherer, K. Leuthäuser-Jaschinski, J. Ecker, G. Schmitz and G. Liebisch, *J. Lipid Res.*, 2010, **51**, 2001–2011.
15. B. Buszewski and S. Noga, *Anal. Bioanal. Chem.*, 2012, **402**, 231–247.
16. C. Petucci, A. Zelenin, J. A. Culver, M. Gabriel, K. Kirkbride, T. T. Christison and S. J. Gardell, *Anal. Chem.*, 2016, **88**, 11799–11803.
17. G. A. Eiceman, H. H. Hill and J. Gardea-torresdey, *Anal. Chem.*, 2000, **72**, 137–144.
18. A. Týčová, V. Ledvina and K. Klepárník, *Electrophoresis*, 2017, **38**, 115–134.
19. R. G. Jayo, M. Thaysen-Andersen, P. W. Lindenburg, R. Haselberg, T. Hankemeier, R. Ramautar and D. D. Y. Chen, *Anal. Chem.*, 2014, **86**, 6479–6486.
20. H. Mischak, J. J. Coon, J. Novak, E. M. Weissinger, J. Schanstra and A. F. Dominiczak, *Mass Spectrom. Rev.*, 2009, **28**, 703–724.
21. D. C. Simpson and R. D. Smith, *Electrophoresis*, 2005, **26**, 1291–1305.
22. I. Kohler, J. Schappler and S. Rudaz, *Anal. Chim. Acta*, 2013, **780**, 101–109.
23. M. L. Vestal, *Chem. Rev.*, 2001, **101**, 361–375.
24. A. J. Dempster, *Phys. Rev.*, 1918, **11**, 316–325.
25. W. Bleakney, *Phys. Rev.*, 1929, **34**, 157–160.
26. A. O. Nier, *Rev. Sci. Instrum.*, 1940, **11**, 212.
27. A. O. Nier, *Rev. Sci. Instrum.*, 1947, **18**, 398.
28. J. T. Watson and O. D. Sparkman, *Introduction to Mass Spectrometry: Instrumentation, Applications, and Strategies for Data Interpretation*, John Wiley & Sons, 2007.

29. R. M. Silverstein, F. X. Webster, D. J. Kiemle and D. L. Bryce, *Spectrometric Identification of Organic Compounds*, John Wiley & Sons, 2014.
30. A. Zlatkis, W. Bertsch, H. A. Lichtenstein, A. Tishbee, F. Shunbo, H. M. Liebich, A. M. Coscia and N. Fleischer, *Anal. Chem.*, 1973, **45**, 763–767.
31. Z. Zhang and G. Li, *Microchem. J.*, 2010, **95**, 127–139.
32. K. K. Pasikanti, P. C. Ho and E. C. Y. Chan, *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, 2008, **871**, 202–211.
33. A. Muñoz-Garcia, J. Ro, J. C. Brown and J. B. Williams, *J. Chromatogr. A*, 2006, **1133**, 58–68.
34. M. S. B. Munson and F.-H. Field, *J. Am. Chem. Soc.*, 1966, **88**, 2621–2630.
35. F. H. Field, *J. Am. Soc. Mass Spectrom.*, 1990, **1**, 277–283.
36. C. K. Meng, M. Mann and J. B. Fenn, *Proceedings of the 36th ASMS Conference on Mass Spectrometry and Allied Topics*, 1988, pp. 5–10.
37. J. B. Fenn, *Angew. Chem. Int. Ed. Engl.*, 2003, **42**, 3871–3894.
38. S. Nguyen and J. B. Fenn, *Proc. Natl. Acad. Sci.*, 2007, **104**, 1111–1117.
39. J. V. Iribarne and B. A. Thomson, *J. Chem. Phys.*, 1976, **64**, 2287–2294.
40. M. Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson and M. B. Alice, *J. Chem. Phys.*, 1968, **49**, 2240–2249.
41. M. Wilm, *Mol. Cell. Proteomics*, 2011, **10**, M111–M9407.
42. M. E. Monge, G. A. Harris, P. Dwivedi and F. M. Fernández, *Chem. Rev.*, 2013, **113**, 2269–2308.
43. T. R. Covey, B. A. Thomson and B. B. Schneider, *Mass Spectrom. Rev.*, 2009, **28**, 870–897.
44. B. Domon and R. Aebersold, *Science*, 2006, **312**, 212–217.
45. L. Hu, M. Ye, X. Jiang, S. Feng and H. Zou, *Anal. Chim. Acta*, 2007, **598**, 193–204.
46. Y. Shen, N. Tolic, C. Masselon, L. Paša-Tolic, D. G. Camp, K. K. Hixson, R. Zhao, G. A. Anderson and R. D. Smith, *Anal. Chem.*, 2004, **76**, 144–154.
47. W. Henderson, *Chem. N. Z.*, 2015, **79**, 128–131.
48. N. B. Cech and C. G. Enke, *Mass Spectrom. Rev.*, 2001, **20**, 362–387.
49. Chromacademy, *Mass Spectrometry Fundamental LC-MS Electrospray Ionisation –Instrumentation*, 2010.
50. W. Lu, B. D. Bennett and J. D. Rabinowitz, *J. Chromatogr. B*, 2008, **871**, 236–242.
51. G. Theodoridis, H. G. Gika and I. D. Wilson, *TrAC, Trends Anal. Chem.*, 2008, **27**, 251–260.
52. E. C. Horning, M. G. Horning, D. I. Carroll, I. Dzidic and R. N. Stillwell, *Anal. Chem.*, 1973, **45**, 936–943.
53. M. Karas, D. Bachmann, U. el Bahr and F. Hillenkamp, *Int. J. Mass Spectrom. Ion Processes*, 1987, **78**, 53–68.
54. M. Karas and F. Hillenkamp, *Anal. Chem.*, 1988, **60**, 2299–2301.
55. M. Karas, U. Bahr and F. Hillenkamp, *Int. J. Mass Spectrom. Ion Processes*, 1989, **92**, 231–242.
56. M. W. Little, J.-K. Kim and K. K. Murray, *J. Mass Spectrom.*, 2003, **38**, 772–777.

57. M. Kussmann, E. Nordhoff, H. Rahbek-Nielsen, S. Haebel, M. Rossel-Larsen, L. Jakobsen, J. Gobom, E. Mirgorodskaya, A. Kroll-Kristensen, L. Palm and others, *J. Mass Spectrom.*, 1997, **32**, 593–601.
58. M. B. O'Rourke, S. P. Djordjevic and M. P. Padula, *Mass Spectrom. Rev.*, 2016, DOI: 10.1002/mas.21515.
59. Y. Fukuyama, *Mass Spectrom.*, 2015, **4**, A0037.
60. J. Liu, G. Jiang and J. Å. Jönsson, *TrAC, Trends Anal. Chem.*, 2005, **24**, 20–27.
61. T. Nakanishi, I. Ohtsu, M. Furuta, E. Ando and O. Nishimura, *J. Proteome Res.*, 2005, **4**, 743–747.
62. J. Rappsilber, Y. Ishihama and M. Mann, *Anal. Chem.*, 2003, **75**, 663–670.
63. J. K. Lewis, J. Wei and G. Siuzdak, *Encycl. Anal. Chem.*, 2000.
64. M. A. Moseley, L. J. Deterding, K. B. Tomer and J. W. Jorgenson, *Anal. Chem.*, 1991, **63**, 1467–1473.
65. R. M. Caprioli, T. Fan and J. S. Cottrell, *Anal. Chem.*, 1986, **58**, 2949–2954.
66. D. S. McPhail, *J. Mater. Sci.*, 2006, **41**, 873–903.
67. R. N. S. Sodhi, *Analyst*, 2004, **129**, 483–487.
68. S. Hofmann, *Philos. Trans. R. Soc. London, Ser. A*, 2004, **362**, 55–75.
69. R. G. Cooks, Z. Ouyang, Z. Takats and J. M. Wiseman, *Science*, 2006, **311**, 1566–1570.
70. Z. Takats, J. M. Wiseman, B. Gologan and R. G. Cooks, *Science*, 2004, **306**, 471–473.
71. R. B. Cody, J. A. Laramée and H. D. Durst, *Anal. Chem.*, 2005, **77**, 2297–2302.
72. M. Haapala, J. Pól, V. Saarela, V. Arvola, T. Kotiaho, R. A. Ketola, S. Franssila, T. J. Kauppila and R. Kostiainen, *Anal. Chem.*, 2007, **79**, 7867–7872.
73. C. N. McEwen, R. G. McKay and B. S. Larsen, *Anal. Chem.*, 2005, **77**, 7826–7831.
74. P. Nemes and A. Vertes, *Anal. Chem.*, 2007, **79**, 8098–8106.
75. C. Brunnée, *Int. J. Mass Spectrom. Ion Processes*, 1987, **76**, 125–237.
76. J. H. Beynon, *Mass Spectrometry and its Applications to Organic Chemistry*, Elsevier, 1960.
77. C. Brunnée, *Rapid Commun. Mass Spectrom.*, 1997, **11**, 694–707.
78. S. A. McLuckey, *Instrumentation for Mass Spectrometry: 1997*, 1997.
79. M. P. Balogh, *Spectroscopy*, 2004, **19**, 10.
80. 96/23/Ec Commission Decision, *96/23/Ec Comm. Decis.*, 2002, p. 29.
81. H. Gallart-Ayala, O. Nuñez, E. Moyano, M. T. Galceran and C. P. B. Martins, *Rapid Commun. Mass Spectrom.*, 2011, **25**, 3161–3166.
82. A. Kaufmann, P. Butcher, K. Maden, S. Walker and M. Widmer, *Anal. Chim. Acta*, 2010, **673**, 60–72.
83. A. Kaufmann, P. Butcher, K. Maden, S. Walker and M. Widmer, *Anal. Chim. Acta*, 2011, **700**, 86–94.
84. F. Hernández, Ó. J. Pozo, J. V. Sancho, F. J. López, J. M. Marín and M. Ibáñez, *TrAC, Trends Anal. Chem.*, 2005, **24**, 596–612.
85. P. Herrero, N. Cortés-Francisco, F. Borrull, J. Caixach, E. Pocurull and R. M. Marcé, *J. Mass Spectrom.*, 2014, **49**, 585–596.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

86. H. Henry, H. R. Sobhi, O. Scheibner, M. Bromirski, S. B. Nimkar and B. Rochat, *Rapid Commun. Mass Spectrom.*, 2012, **26**, 499–509.
87. S. J. Bruce, B. Rochat, A. Béguin, B. Pesse, I. Guessous, O. Boulat and H. Henry, *Rapid Commun. Mass Spectrom.*, 2013, **27**, 200–206.
88. A. Kaufmann, P. Butcher, K. Maden, S. Walker and M. Widmer, *Talanta*, 2011, **85**, 991–1000.
89. A. Kaufmann, P. Butcher, K. Maden, S. Walker and M. Widmer, *Rapid Commun. Mass Spectrom.*, 2011, **25**, 979–992.
90. S. De Baere, A. Osselaere, M. Devreese, L. Vanhaecke, P. De Backer and S. Croubels, *Anal. Chim. Acta*, 2012, **756**, 37–48.
91. W. E. Stephens, *Phys. Rev.*, 1946, **69**, 691.
92. J. H. Gross, *Mass Spectrometry*, Springer Verlag, Berlin Heidelberg, 2004.
93. E. De Hoffmann and V. Stroobant, *Mass Spectrometry - Principles and Applications*, 2007, vol. 29.
94. S. G. Alikhanov, *J. Exp. Theor. Phys.*, 1957, **4,** 452–453.
95. B. A. Mamyrin, *Int. J. Mass Spectrom.*, 2001, **206**, 251–266.
96. B. A. Mamyrin, V. I. Karataev, D. V Shmikk and V. A. Zagulin, *J. Exp. Theor. Phys.*, 1973, **37**, 45.
97. E. G. Johnson and A. O. Nier, *Phys. Rev.*, 1953, **91**, 10.
98. T. W. Burgoyne and G. M. Hieftje, *Mass Spectrom. Rev.*, 1996, **15**, 241–259.
99. J. D. Morrison, *Org. Mass Spectrom.*, 1991, **26**, 183–194.
100. W. Paul and H. Steinwedel, *Z. Naturforsch., A*, 1953, **8**, 448–450.
101. J. C. Schwartz, M. W. Senko and J. E. P. Syka, *J. Am. Soc. Mass Spectrom.*, 2002, **13**, 659–669.
102. Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman and R. Graham Cooks, *J. Mass Spectrom.*, 2005, **40**, 430–443.
103. A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat and S. Horning, *Anal. Chem.*, 2006, **78**, 2113–2120.
104. A. Makarov and M. Scigelova, *J. Chromatogr. A*, 2010, **1217**, 3938–3945.
105. E. O. Lawrence and M. S. Livingston, *Phys. Rev.*, 1932, **40**, 19.
106. D. W. Koppenaal, C. J. Barinaga, M. B. Denton, R. P. Sperline, G. M. Hieftje, G. D. Schilling, F. J. Andrade and J. H. Barnes IV, *Anal. Chem.*, 2005, **77**, 418A–427A.
107. T. Geiger, J. Cox and M. Mann, *Mol. Cell. Proteomics*, 2010, **9**, 2252–2261.
108. L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner and R. Aebersold, *Mol. Cell. Proteomics*, 2012, **11**, O111.016717.
109. R. W. Purves, R. Guevremont, S. Day, C. W. Pipich and M. S. Matyjaszczyk, *Rev. Sci. Instrum.*, 1998, **69**, 4094–4105.
110. A. B. Kanu, P. Dwivedi, M. Tam, L. Matz and H. H. Hill, *J. Mass Spectrom.*, 2008, **43**, 1–22.
111. H. Borsdorf and G. A. Eiceman, *Appl. Spectrosc. Rev.*, 2006, **41**, 323–375.
112. A. López, T. Tarragó, M. Vilaseca and E. Giralt, *New J. Chem.*, 2013, **37**, 1283–1289.
113. J. S. Brodbelt, *Chem. Soc. Rev.*, 2014, **43**, 2757–2783.
114. M. Gilar, K. J. Fountain, Y. Budman, U. D. Neue, K. R. Yardley, P. D. Rainville, R. J. Russell and J. C. Gebler, *J. Chromatogr. A*, 2002, **958**, 167–182.

115. B. Chen, S. F. Mason and M. G. Bartlett, *J. Am. Soc. Mass Spectrom.*, 2013, **24**, 257–264.
116. R. Erb and H. Oberacher, *Electrophoresis*, 2014, **35**, 1226–1235.
117. G. S. Philibert and S. V. Olesik, *J. Chromatogr. A*, 2011, **1218**, 8222–8230.
118. H. Qiu, E. Wanigasekara, Y. Zhang, T. Tran and D. W. Armstrong, *J. Chromatogr. A*, 2011, **1218**, 8075–8082.
119. E. Rodríguez-Gonzalo, D. García-Gómez and R. Carabias-Martínez, *J. Chromatogr. A*, 2011, **1218**, 3994–4001.
120. G. Shi, J. Wu, Y. Li, R. Geleziunas, K. Gallagher, T. Emm, T. Olah and S. Unger, *Rapid Commun. Mass Spectrom.*, 2002, **16**, 1092–1099.
121. S. Cohen, M. Megherbi, L. P. Jordheim, I. Lefebvre, C. Perigaud, C. Dumontet and J. Guitton, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2009, **877**, 3831–3840.
122. A. Zimmermann, R. Greco, I. Walker, J. Horak, A. Cavazzini and M. Lämmerhofer, *J. Chromatogr. A*, 2014, **1354**, 43–55.
123. J. Xing, A. Apedo, A. Tymiak and N. Zhao, *Rapid Commun. Mass Spectrom.*, 2004, **18**, 1599–1606.
124. H.-T. Feng, N. Wong, S. Wee and M. M. Lee, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2008, **870**, 131–134.
125. J. Chen, Q. Shi, Y. Wang, Z. Li and S. Wang, *Molecules*, 2015, **20**, 5423–5437.
126. J.-X. Liu, J. T. Aerts, S. S. Rubakhin, X.-X. Zhang and J. V Sweedler, *Analyst*, 2014, **139**, 5835–5842.
127. J. Donfack and A. Wiley, *Forensic Sci. Int.: Genet.*, 2015, **16**, 112–120.
128. L. J. Kobrynski, G. K. Yazdanpanah, D. Koontz, F. K. Lee and R. F. Vogt, *Clin. Chem.*, 2016, **62**, 287–292.
129. F. Chen, B. Gülbakan, S. Weidmann, S. R. Fagerer, A. J. Ibáñez and R. Zenobi, *Mass Spectrom. Rev.*, 2016, **35**, 48–70.
130. X. Gao, B.-H. Tan, R. J. Sugrue and K. Tang, *Top. Curr. Chem.*, 2013, **331**, 55–77.
131. K. Meyer and P. M. Ueland, *Adv. Clin. Chem.*, 2011, **53**, 1–29.
132. M. C. Monti, S. X. Cohen, A. Fish, H. H. K. Winterwerp, A. Barendregt, P. Friedhoff, A. Perrakis, A. J. R. Heck, T. K. Sixma, R. H. H. van den Heuvel and J. H. G. Lebbink, *Nucleic Acids Res.*, 2011, **39**, 8052–8064.
133. J. R. Enders and J. A. McLean, *Chirality*, 2009, **21**(suppl. 1), E253–E264.
134. T. Shi, E. Song, S. Nie, K. D. Rodland, T. Liu, W.-J. Qian and R. D. Smith, *Proteomics*, 2016, **16**, 2160–2182.
135. A. Bensimon, A. J. R. Heck and R. Aebersold, *Annu. Rev. Biochem.*, 2012, **81**, 379–405.
136. S. Tate, B. Larsen, R. Bonner and A.-C. Gingras, *J. Proteomics*, 2013, **81**, 91–101.
137. Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek and J. R. Yates, *Chem. Rev.*, 2013, **113**, 2343–2394.
138. A. Leitner, *Methods Mol. Biol.*, 2016, **1355**, 105–121.
139. J. Huang, F. Wang, M. Ye and H. Zou, *J. Chromatogr. A*, 2014, **1372**, 1–17.
140. H.-C. Liang, E. Lahert, I. Pike and M. Ward, *Bioanalysis*, 2015, **7**, 383–400.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

141. O. Chahrour, D. Cobice and J. Malone, *J. Pharm. Biomed. Anal.*, 2015, **113**, 2–20.
142. I. Tzoulaki, T. M. D. Ebbels, A. Valdes, P. Elliott and J. P. A. Ioannidis, *Am. J. Epidemiol.*, 2014, **180**, 129–139.
143. T. Fuhrer and N. Zamboni, *Curr. Opin. Biotechnol.*, 2015, **31**, 73–78.
144. I. Aretz and D. Meierhofer, *Int. J. Mol. Sci.*, 2016, **17**, 632.
145. H. G. Gika, G. A. Theodoridis, R. S. Plumb and I. D. Wilson, *J. Pharm. Biomed. Anal.*, 2014, **87**, 12–25.
146. T. Cajka and O. Fiehn, *Anal. Chem.*, 2016, **88**, 524–545.
147. M. Ghaste, R. Mistrik and V. Shulaev, *Int. J. Mol. Sci.*, 2016, **17**, 816.
148. M. Li, L. Yang, Y. Bai and H. Liu, *Anal. Chem.*, 2014, **86**, 161–175.
149. S. M. Lam and G. Shui, *J. Genet. Genomics*, 2013, **40**, 375–390.
150. B. Brügger, *Annu. Rev. Biochem.*, 2014, **83**, 79–98.
151. T. Cajka and O. Fiehn, *TrAC, Trends Anal. Chem.*, 2014, **61**, 192–206.
152. J. Cimino, D. Calligaris, J. Far, D. Debois, S. Blacher, N. Sounni, A. Noel and E. De Pauw, *Int. J. Mol. Sci.*, 2013, **14**, 24560–24580.
153. M. Haag, A. Schmidt, T. Sachsenheimer and B. Brügger, *Metabolites*, 2012, **2**, 57–76.
154. H. C. Köfeler, A. Fauland, G. N. Rechberger and M. Trötzmüller, *Metabolites*, 2012, **2**, 19–38.
155. K. Yang and X. Han, *Metabolites*, 2011, **1**, 21–40.
156. E. D. Dodds, M. R. McCoy, L. D. Rea and J. M. Kennish, *Lipids*, 2005, **40**, 419–428.
157. G. C. Burdge, P. Wright, A. E. Jones, S. A. Wootton, T. G. Bernhardt, P. A. Cannistraro, D. A. Bird, K. M. Doyle, M. Laposata, P. A. Caesar, S. J. Wilson, C. S. Normand, A. D. Postle, A. S. Fosbrooke, I. Tamir, A. C. v. Houwelingen, M. M. H. P. F.-v. Drongefen, U. Nicolini, K. H. Nicolaides, M. D. M. Al, A. D. M. Kester, G. Hornstra, E. B. Hoving, G. Jansen, M. Volmer, J. J. van Doormaal, F. A. J. Muskiet, R. G. Jensen, R. M. Clark, S. A. Gerrior, M. B. Fey and A. M. Gotto, *Br. J. Nutr.*, 2000, **84**, 781–787.
158. P. Henry, O. Owopetu, D. Adisa, T. Nguyen, K. Anthony, D. Ijoni-Animadu, S. Jamadar, F. Abdel-Rahman and M. A. Saleh, *J. Environ. Sci. Health, Part B*, 2016, **51**, 546–552.
159. N. Taniguchi and Y. Kizuka, *Adv. Cancer Res.*, 2015, **126**, 11–51.
160. E. Maverakis, K. Kim, M. Shimoda, M. E. Gershwin, F. Patel, R. Wilken, S. Raychaudhuri, L. R. Ruhaak and C. B. Lebrilla, *J. Autoimmun.*, 2015, **57**, 1–13.
161. G. Lauc, M. Pezer, I. Rudan and H. Campbell, *Biochim. Biophys. Acta*, 2016, **1860**, 1574–1582.
162. S. S. Ferreira, C. P. Passos, P. Madureira, M. Vilanova and M. A. Coimbra, *Carbohydr. Polym.*, 2015, **132**, 378–396.
163. L. Cao, Y. Qu, Z. Zhang, Z. Wang, I. Prytkova and S. Wu, 2016, **13**, 513–522.
164. M. Wuhrer, *Glycoconj. J.*, 2013, **30**, 11–22.
165. V. Kolli, K. N. Schumacher and E. D. Dodds, *Bioanalysis*, 2015, **7**, 113–131.

166. H. Hu, K. Khatri, J. Klein, N. Leymarie and J. Zaia, *Glycoconj. J.*, 2016, **33**, 285–296.
167. T. Powers, S. Holst, M. Wuhrer, A. Mehta and R. Drake, *Biomolecules*, 2015, **5**, 2554–2572.
168. A. Kilár, Á. Dörnyei and B. Kocsis, *Mass Spectrom. Rev.*, 2013, **32**, 90–117.
169. J. Nilsson, *Glycoconj. J.*, 2016, **33**, 261–272.
170. M. Thaysen-Andersen and N. H. Packer, *Biochim. Biophys. Acta, Proteins Proteomics*, 2014, **1844**, 1437–1452.
171. J. Zaia, *Methods Mol. Biol.*, 2013, **984**, 13–25.
172. X. Sun, L. Lin, X. Liu, F. Zhang, L. Chi, Q. Xia and R. J. Linhardt, *Anal. Chem.*, 2016, **88**, 1937–1943.
173. N. Canela, M. Á. Rodíguez, I. Baiges, P. Nadal and L. Arola, *Electrophoresis*, 2016, **37**, 1748–1767.
174. P. J. Trim and M. F. Snel, *Methods*, 2016, **104**, 127–141.
175. Y. Dong, B. Li and A. Aharoni, *Trends Plant Sci.*, 2016, **21**, 686–698.
176. G. Hochart, G. Hamm and J. Stauber, *Bioanalysis*, 2014, **6**, 2775–2788.
177. L. Minerva, A. Ceulemans, G. Baggerman and L. Arckens, *Proteomics: Clin. Appl.*, 2012, **6**, 581–595.
178. J. Mach, *Plant Cell*, 2012, **24**, 371.
179. E. H. Seeley and R. M. Caprioli, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 18126–18131.
180. S. Uzbekova, S. Elis, A.-P. Teixeira-Gomes, A. Desmarchais, V. Maillard and V. Labas, *Biology*, 2015, **4**, 216–236.
181. A. Karger, *Proteomics: Clin. Appl.*, 2016, **10**, 982–993.
182. F. Nomura, *Biochim. Biophys. Acta*, 2015, **1854**, 528–537.
183. A. Dierig, R. Frei and A. Egli, *Pediatr. Infect. Dis. J.*, 2015, **34**, 97–99.
184. E. Nagy, *Future Microbiol.*, 2014, **9**, 217–233.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00017

CHAPTER 3

Metabolomics

RICARDO R. DA SILVA[a,c], NORBERTO PEPORINE LOPES*[a] AND DENISE BRENTAN SILVA[a,b]

[a]Núcleo de Pesquisa em Produtos Naturais e Sintéticos (NPPNS), Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil; [b]Laboratório de Produtos Naturais e Espectrometria de Massas (LAPNEM), Faculdade de Ciências Farmacêuticas, Alimentos e Nutrição (FACFAN), Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande, MS, Brazil; [c]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
*E-mail: npelopes@fcfrp.usp.br

3.1 Introduction

The rise of "omics sciences", with high-throughput measurements of cellular macromolecules DNA, RNA and proteins, has also opened up avenues to the measurement of cellular small organic molecules, which is the foundation of metabolomics. Mass spectrometry (MS) technology, one of the main platforms used for metabolomics, is older than DNA sequencing technologies, and has been used on high-throughput measurements of organic molecules for over 30 years,[1] in the fields of flavor and fragrance analysis, applying electron ionization (EI). Other pioneer studies of complex biological extracts of small organic molecules were also carried out long before large omics sequencing development, such as the analysis of human feces extracts in the 1960s.[2] However,

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

only after the rise of genomics did small organic molecule measurements start to be contextualized as part of an omics approach in the late 1990s.[3]

After the introduction of the terms metabolome and metabolomics, at the beginning of the 21st century, the terms started to be extensively used, and the new field of metabolomics started to be delineated. The metabolome is defined as the complete set of small organic molecules (generally arbitrarily defined as <1500 Da) produced by a given cell in a given time and space. The definition of time (stage of development) and space (from environmental conditions to the tissue in which the cells are present) is extremely important as the metabolome will be strictly regulated by these variables. Metabolomics is therefore defined as the set of analytical techniques used to measure a large subset of the metabolome.[4,5] In this chapter we will focus on the MS platforms applied to metabolomics, under the assumption that no single analytical platform is capable of measuring all of the metabolome.

Although metabolomics has some similar principles to dereplication, such as analyses on a large scale, database search and online compound identification, these techniques cannot be considered to be synonyms for each other. The coverage of metabolomics is higher, producing a larger amount of data, which requires extensive application of statistics and computer science to facilitate the data analyses and interpretation.[6–8] The number of studies involving metabolomics has increased significantly. Untargeted metabolomics applications are mainly in the areas of health and disease (42%), nutrition/lifestyle (18%), plant metabolism (11%), pharmacology (7%), microbial interactions (6%), functional genomics (6%), biotechnology (4%), metabolic regulation (3%) and others (4%).[9,10]

The metabolomics scientific community advocates that metabolites integrate information from higher cellular levels, encoded in the genome and further processed by the proteome, in response to environmental clues. Therefore, the metabolome is regarded as a direct measurement of the phenotype. As the precursor to the omics sciences, the field of genomics has greatly shaped the systematic study of genes and genomes. Genomes from a large number of organisms have been characterized, giving a broad view of genome size, organization and regulation.[11] Despite all the advances in genomics since the early 1990s, the function of at least 30% of the genes of the smallest known genome remains unknown.[12] The fields of transcriptomics and proteomics inherited the “blueprint” structure from the genome, deriving from central dogma the set of transcripts and proteins that could be possibly codified from the genome. From those blueprints a great number of post-transcriptional and post-translational, as well as more complex challenges to the central dogma, were later discovered.[13]

Recently, a combination of proteomics and metabolomics was applied to correlate protein levels and metabolites from strains for the optimization of isopentenol production.[14] The same idea has been applied to classical functional genomics, named metabolomics genome-wide association studies

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

(mGWAS), to correlate genetic variation with the metabolic changes. These studies, for example, have been applied to human samples (blood and urine) and microorganisms to determine potential enzymatic activities, and to discover unknown enzymatic functions.[9,15,16] These innovative combinations are not an easy task and have not yet been explored routinely, but research has evolved in this direction since the results allow the understanding of biological systems as a whole.[9]

Therefore, to this day, there are no systematic approaches to defining the metabolome and to create direct links to the genome. Current assessments of organism specific metabolome sizes are largely incomplete,[17] and a direct link to the genome is still complicated by a series of factors such as incomplete genome annotations and enzymatic promiscuity,[18] hampering the direct association between genes and genomes.

The largely incomplete knowledge on the set of all possible organic structures, the so-called "chemical space",[18,19] the lack of comprehensive ontology/hierarchical classification of molecular classes[20] for the main metabolite databases,[21] and a historical gap on community data sharing (see Chapter 10) have limited metabolomics development. In this chapter we try to contextualize the main MS-based metabolomics approaches to molecular classes and metabolic partition, and identify the best analytical workflow, including the most recent analytical and computational resources, for the desired subset of the metabolome. In this way, we seek to enable metabolomics practitioners to correctly design experiments, based on specific biological questions, and keep in mind which metabolites are being sampled and which workflow is best suited to the study goal.

3.2 Experimental Design

The great diversity of physicochemical properties from small organic molecules prevents the metabolome from being analyzed by a single analytical technique.[22] However, few attempts have been made to formalize the extent of those limitations on the classes of molecules and partitions of the metabolism that can be accessed by a given metabolomics experiment. To design a metabolomics experiment, one of the first questions a biologist should ask is: given the physicochemical properties, such as polarity, melting and boiling points, and solubility, what is the best setup for analyzing a given metabolic pathway? In Figure 3.1 we outline the steps, starting from the target molecules in a given metabolism partition to the main steps to be considered in developing a metabolomics protocol.

In a recent editorial, Choi and Verpoorte[23] called attention to the fact that most researchers choose a metabolomics analytical method without any previous assessment of the quality or coverage of the target metabolome. The authors also point to several factors to be considered when setting up a protocol. The central idea of experimental design is to take into consideration all the possible factors that could influence the analysis, and explore all opportunities to determine the best design.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

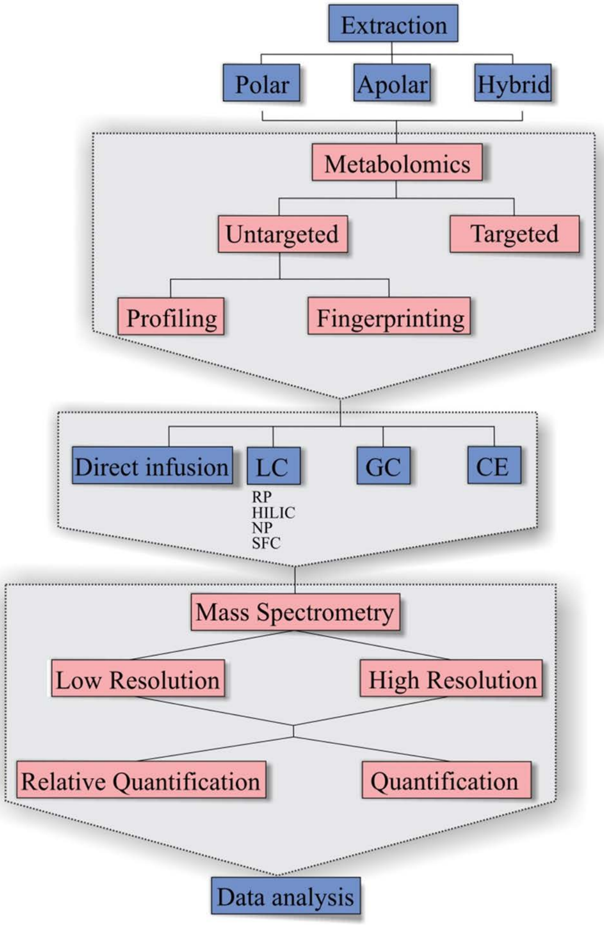

Figure 3.1 Conceptual representation of design-guided extraction of a specific metabolite class. The extraction protocol, separation technique, ionization source, and detector should be selected based on the physicochemical properties of the metabolites being targeted. According to the same concept, the metabolism partition targeted by the study goal has to be contained in the molecular classes measured by the specific experimental setup. LC: liquid chromatography; GC: gas chromatography, CE: capillary electrophoresis, RP: reverse phase; HILIC: hydrophilic interaction chromatography; NP: normal phase; SFC: supercritical fluid chromatography.

The experimental design begins with the establishment of rigorous procedures to control experimental variation, to detect the best analytical parameters and reagents, and to establish the metabolic partition and molecular classes being targeted. Many studies have shown that the use of quality control (QC) samples, with internal standards and blanks, is a simple and powerful tool that helps to detect contamination, carryover, batch effects and false molecular features.[24–26] Hughey and collaborators[26] developed a QC system to monitor short- and long-term retention time reproducibility, instrument response and mass accuracy for untargeted metabolomics. The authors called attention to the fact that, in targeted quantification, calibration curves are routinely used, but similar control measures are not adopted in untargeted analyses. The realization of this importance has made the U. S. Food and Drug Administration propose bioanalytical validation criteria using QC samples and calculate the relative standard deviation from compounds present in QC.[27] Duan and coworkers[25] devised a workflow to improve the discrimination of biological from non-biological signals on untargeted LC-MS-based metabolomics, using the above-cited tools. First the authors used blank samples to tune the instrument, next, a QC mix was analyzed with technical replicates to identify peaks that could not be reproducibly detected, and, finally, a QC mix dilution series was analyzed to help with quantification. The authors reported improved detection of true signals when validating an artificial mixture of 20 standard compounds. The standard peak extraction and alignment found 1342 molecular features, which, after the described filtering procedure, were reduced to 102 features, in which the authors stated that all standards were present.[25] In other words, from a sample of only 20 standards, it is possible to detect up to a thousand peaks, which, after accounting for multiple adducts, in-source fragments, isotopic peaks and other possible sources of transformations from original peaks, can be attributed to contamination and noise signal.

In an ideal scenario, pilot experiments should be conducted to determine the best experimental setup. These pilots should be conducted using statistical tools in order to quantitatively determine the best experimental conditions, as well as minimum required number of samples. Some concepts of experimental planning, such as replication, blocking and randomization are essential for characterizing the source of the observed experimental variability.[28] Without these controlled experimental procedures it is not possible to distinguish random variation and systematic experimental variance from true biological differences.[29] Therefore, important issues should be evaluated beforehand, avoiding the acquisition of biased measurements due to artifact formation, instrumental variations, and errors in processing and sampling, including the number of samples, to ensure statistical significance, and that the correct sampling methods, sample handling, decisions on analytical methods, the parameters to data treatment, *etc.* are chosen.[9,30] In addition, the pilot experiment could be applied to determine several experimental parameters and to identify variants, such as the stability of samples, the determination of contaminant presence, and specific issues relating to sample preparation, such as the production of emulsion and temperature changes, which can affect the reproducibility.[10,30]

In this context, statistical design of experiments (DOE) is a useful tool for optimizing the method using the responses between the combinations of the variables in a system (Figure 3.2). Thus, it can be used to determine the best set of parameters to be used to obtain the best outcome; in the case of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

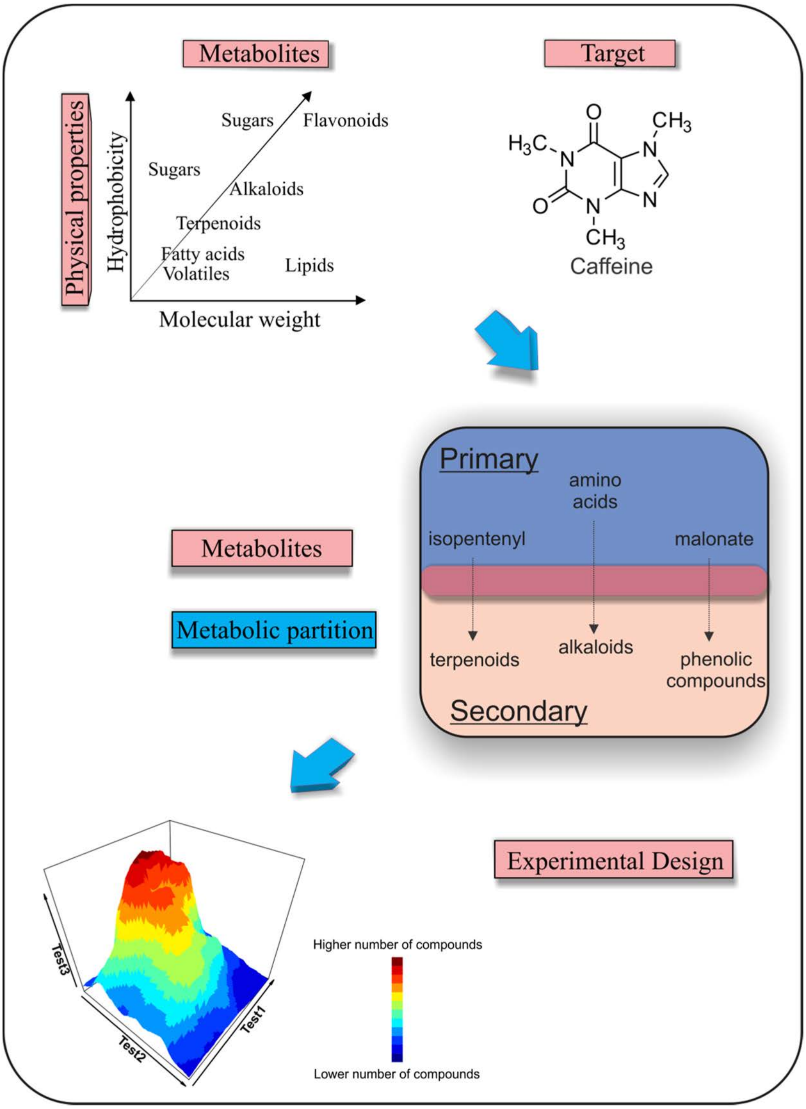

Figure 3.2 General metabolomics workflow. The concept of a general workflow shows the options available for targeting specific classes of compounds, which should be matched with the study goal.

metabolomics, the highest number of molecular features, for example. The outcome is measured as a multidimensional surface, over the simultaneous change in multiple parameters, as depicted in Figure 3.2. The DOE approach accounts for variable interaction, being preferred over single-variable analysis.[29] A detailed tutorial on experimental design has already been published, describing the types of available design and the statistical tools applicable.[31] However, another review article recently described different and modern

DOE, and demonstrated case studies and comparisons between methods, such as central composite design, full factorial, fractional factorial, Taguchi array and others.[32] For DOE, the following steps can be listed: selection of significant factors, elaboration of design strategy, model building and application of the optimized parameters.[33]

The concept of power analysis is used to address the question of the minimum number of samples required for an experiment to detect statistically significant differences for the desired statistical analysis.[34] As pilot experiments may not be possible due to resource limitations, the classical methods of power analysis are not suitable for metabolomics. Efficient methods for power analysis using large simulated datasets are becoming available in the omics literature and more recently for metabolomics.[34,35] These methods are becoming critical, for example in sensitive clinical trials, and to decrease the cost of studies in large populations.[36]

After definition of the strategy for optimizing the experimental conditions and controlling for variation, it is necessary to choose the analytical workflow based on the project's final goal. Metabolomics is commonly divided into untargeted and targeted metabolomics (Figure 3.1). In the target approach the metabolites are previously selected and the methods are developed for these metabolites, while the untargeted approach aims to record all the detectable metabolites in a sample, including unknown metabolites. This category is further sub-classified into fingerprinting analyses, which use minimal extraction protocols and faster spectrometric measurements without mandatory identification, and profiling, which also includes the quantification of the metabolites.[37] In the next sections we will show how these methods are connected with non-selective sample pretreatment and hypothesis-driven targeted separation, according to the workflow presented on Figure 3.2. It is important to note that these classifications are not discrete, and a combination of approaches is possible.

3.3 Sample Preparation

The isolation of the target analytes from a matrix is one of the most important steps in metabolomics. Also involved are the many steps of the clean-up and concentration of the analytes, which can represent a problem for quantitative analyses.[38,39] Thus the extraction protocol is a point of experimental design that deserves total attention. The choice of this protocol has to take into account the experimental design principles outlined above, and has to be combined with the metabolomics analysis category and spectrometry platform. Villas-Boas[40] enumerates three main principles for developing an efficient method of extraction of intracellular metabolites, which are: the knowledge of the biological matrix, in particular the cell envelope structure; the chemical nature of the metabolites (*i.e.*, physicochemical properties); and the sources of losses (photo-degradation, thermo-degradation, solvent exclusion, *etc.*). According to the author, the first step of the extraction process is the rapid inactivation of metabolism enzymatic activities, usually achieved

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

by placing the biological sample in contact with a cold (<−40 °C) or hot (>80 °C) solution or with an acidic (pH < 2.0) or alkaline (pH > 10) solution,[40] or, for plant material, drying of the samples (by lyophilization or circulating air) is the most important.[41]

Therefore, an ideal extraction protocol must be non-selective (untargeted analysis), simple, fast, adequate for quenching representative metabolites, reproducible, and must avoid degradation or losses.[38,42] Beyond the extraction, other steps involving sample preparation are equally important, such as sample collection, transportation, storage, and clean-up of samples.[43] The extraction and clean-up processes will be described in more detail later.

Metabolites are generally classified according to physicochemical properties related to molecular weight, polarity and volatility. These properties define the choice of extraction solvent, clean-up method, analytical separation and MS platform. The most common polar solvents applied are methanol, methanol–water mixtures, or ethanol with or without organic acids, and the common non-polar solvents are chloroform, ethyl acetate, or hexane, for extracting lipophilic compounds.[37,41,44] According to Cajka and Fiehn[24] the extraction efficiency can be optimized with separate extraction steps and/or the use of solvent mixtures as opposed to a single step/solvent extraction. Yuliana and collaborators[45] proposed the use of a gradient of extraction solvents. The metabolites extracted were clustered in three groups, classified as lipophilic (*n*-hexane or ethyl acetate), medium polar (methanol) and polar or hydrophilic (water). One of the drawbacks of this approach is increased complexity, time and cost, and it should be carefully considered for large-scale metabolomics studies. The 25th edition of the journal *Phytochemical Analysis* published articles with discussions on the optimization of extraction protocols.[23] There is also a vast amount of literature covering sample specificities that have to be taken in account during the extraction process, such as plant, tissue, cell cultures, biofluids, *etc.*[38,39,41–43,46,47]

Thus, one of the challenges in metabolomics is to capture metabolites distributed over a huge dynamic range, in which metabolites are present in biological matrices.[22] In this context, ultrasound has recently been applied as an extraction tool, demonstrating the main advantages of shortening the extraction time and increased efficiency, in addition to possible automation. However, degradation can occur, and the frequency should be evaluated, since the authors normally use a fixed frequency (20 or 40 kHz). It is not often applied in metabolomics and lipidomics, but ultrasound-assisted extraction has also been described for different strategies such as assisting in the steps for polymer matrices, liquid–liquid extraction (LLE) and (bio) chemical reactions.[48]

After extraction, other steps can be applied before the analyses, such as concentration and clean-up. Concentration is required for some cases, and impacts specific metabolite classes being targeted. One of the ways of addressing this issue is evaporation in order to concentrate the metabolites, followed by re-suspension on solvents compatible with the analytes and the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

chromatographic system; however, loss of metabolites can occur due to their volatilization or degradation.[49]

The removal of interferents is a crucial stage that can provide better results due to lower matrix effects, and preservation of the LC system, as well as the MS ionization source.[30,39] The main methods for this purpose are protein precipitation, solid phase extraction and LLE. In addition to these methods, other recent methodologies for sample preparation have been applied, such as supported liquid extraction, phospholipid removal plates, magnetic beads, Turboflow, monolithic spin column extraction, microextraction by packed sorbent, carbon nanotubes, restricted access materials, immunosorbents, molecular imprinted polymers and aptamers (Figure 3.3).[39,50,51] Many studies have been carried out to evaluate different sample preparation methods and to establish their effects on the metabolic profile, but fast and minimal handling, reproducibility, and metabolism quenching were the goals of them all. From issues in sample preparation, there is limited information in the literature, for example, on the minimization of exogenous interference addition, adsorptive losses, recovery of different metabolite classes, storage effects (long and short term), comparison of metabolite profiles (*in vivo* and *ex vivo*), *etc.*[38,52]

The preferred sample biofluid is plasma, but serum has shown interesting results in analyzing small molecules, mainly due to easier handling. Two large metabolomics projects (the Human Metabolome Project and HUSERMET), based on serum metabolomics, have stimulated its application.[53] From these projects and other published studies, a common sample preparation method is protein precipitation by organic solvent (addition of 3 : 1, v/v), using methanol or acetonitrile and its subsequent centrifugation.[39,54] Recently, important review articles have been published describing sample preparation methods, for plasma and serum, for metabolomics approaches.[38,55]

In addition to metabolomics studies on tissue and cultured cells, single cell and subcellular metabolomics studies have recently been carried out, producing new information on metabolomes of specific cells, or its compartments, and elucidating open questions, mainly for plants.[56,57] The sampling for single cells and single-cell types can be achieved by different methods, including laser microdissection (LMD), LMD and pressure catapulting, laser capture microdissection, fluorescence activated cell sorting, protoplasting, and cell suspension culture.[57] A subcellular metabolomics study was carried out using vacuoles from mesophyll cell protoplasts of *Hordeum vulgare*, allowing the identification of 259 metabolites by GC-MS and ultra-performance LC Fourier transform (UPLC-FT) MS.[58] Recently, a protocol for live single-cell MS was developed, and thousands of metabolites from a single plant cell were detected, but this methodology could also be applied to animal cells. A cell is selected by an optical microscope and a metal-coated nanospray microcapillary tip is introduced in the cell to extract its metabolites. After addition of the ionization solvent, it is trapped in the MS and analyzed by MS and MS/MS, and the isomers separated by ion mobility. However, this methodology was demonstrated to be inefficient for protein analysis.[59,60]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

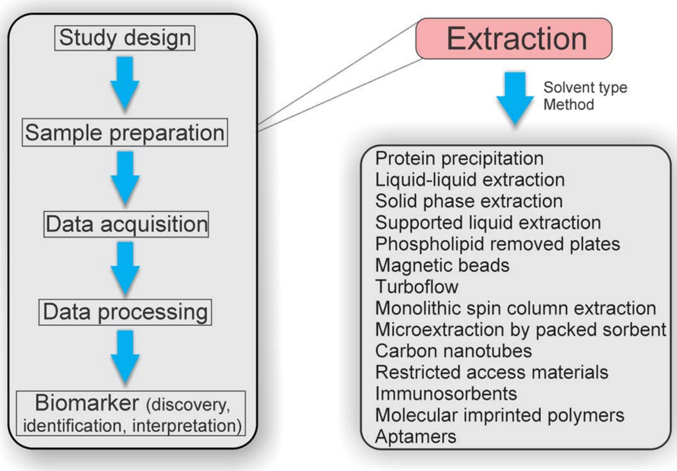
Study design
Sample preparation
Data acquisition
Data processing
Biomarker (discovery, identification, interpretation)
Extraction
Solvent type
Method
Protein precipitation
Liquid-liquid extraction
Solid phase extraction
Supported liquid extraction
Phospholipid removed plates
Magnetic beads
Turboflow
Monolithic spin column extraction
Microextraction by packed sorbent
Carbon nanotubes
Restricted access materials
Immunosorbents
Molecular imprinted polymers
Aptamers

Figure 3.3 General workflow for a metabolomics study highlighting the extraction process, and some strategies applied for interference removal.

3.4 Analytical Platforms—Hyphenated Methods

The large range of physical and chemical differences found among metabolites, such as polarity, volatility, *etc.*, makes the distinction of different metabolite classes from each other, and from matrix components, challenging.[61] Therefore, these properties will influence the choice of analytical methods, including the sample preparation, chromatographic separation method, MS parameters, and the choice of the ionization source and mass analyzer. In metabolomics, the analysis can be performed using a chromatographic system or by direct infusion (DI), injecting the compound mixture directly in the ionization source of the spectrometer.[37] Although DI in the electrospray ionization (ESI) source is a fast technique, some disadvantages have been described, such as ion suppression, which produces fewer ions compared to LC-ESI-MS.[10,37] Other important aspects are described in Chapter 2.

Thus, a chromatographic system can be coupled to an MS in the metabolomics analysis, increasing the amount of chemical information from the samples, and also making up for the finite mass resolution of spectrometers, to reduce the inference of the analytes and the background.[62] The most-used techniques are GC and LC. The first, GC, is widely used to analyze non-polar and small molecules (low molecular weight), or derived small polar compounds such as sugars and amino acids.[10,38,44] However, this chapter describes the main issues related to the LC technique, since it has recently been referred to in many reviews due to its versatile applications.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

The usual LC method takes 30 to 60 min, representing a low number of samples analyzed per day, but the development of new stationary phases has reduced the analysis time and the co-elution of the analytes by the use of small particles and new phases.[10,62] Sub-2 μm particles have been used for this purpose, and they reveal high sensitivity, high peak capacity and high reproducibility. In addition LC analysis with elevated temperatures allows for higher flow rates and lower analysis time with high chromatographic resolution.[63] Recently, superficially porous particles (fused core) have also been applied, revealing high separation efficiency and reasonable pressures (particles 2.5 to 2.7 μm), however, a low loading capacity has been described for them. These particles were applied as a stationary phase in several studies on biofluids, food, plants, fungus, bile acids, *etc.*[24,64–66]

Another innovative stationary phase is the monolithic column, which is composed of a single porous polymer from a network of copolymers (polymethacrylate, polystyrene or bonded silica). It has mesopores and macropores that produce lower resistance of solvent flow and lower pressures. This increased porosity and flow-through pore size results in enhanced separation speeds and efficiency.[67,68] The bonding of alkyl chains and endcapping reactions are possible for monolithic columns using the same methods used for conventional silica particles of high-pressure LC (HPLC), so it is possible to compare similar selectivity for different columns in the development of analytical methods. Monolithic columns have demonstrated lower separation capacity and efficiency compared to sub-2 μm particles, but enhanced properties in relation to 5 μm particles (conventional HPLC) are evident.[67,69]

Advances in chromatographic systems have also been evident, such as SFC, ultra-HPLC (UHPLC), nano/capillary LC and multiple dimensions in LC.[63] Nano/capillary LC applies flows lower than 1 μL min^{-1}, showing higher sensitivity. However, it has some disadvantages related to lower robustness, peak enlargement due to the high dead-volume of systems, sensitivity to sample preparation and higher re-equilibration time.[70,71] Recently, this technique was used with matrix-assisted laser desorption/ionization (MALDI)-MS equipment, and despite its wide application in peptide and protein analyses, this coupling could give more chemical information on the composition.[72,73]

Although the multiple dimensions in LC are not widely used in metabolomics, it is a combination of stationary phases with different separation/selectivity mechanisms that promotes different physicochemical interactions to increase the separation of the compounds, improving the problems related to ion suppression in the ESI source.[74] Some applications carried out by offline analyses combining a normal phase column, another with reverse phase and, recently, HILIC, an innovative stationary phase, have also been used to gather information on polar compounds. In addition, ion mobility spectrometry (IMS) is regarded as another orthogonal dimension, which separates isomeric ions and thus, LC x IMS can be extremely useful for complex mixture analyses, since the first dimension separates the compounds by stationary interactions and the IMS can afford the separation of co-eluted compounds with the same (isobar ions) or different *m*/*z*.[63,75]

For LC analyses, different stationary phases have been applied, such as alkyl (C18, C8, and C30), cyano, a polar group embedded alkyl chain, aromatic-based, polar end-capped alkyl, fluorinated stationary phase and HILIC.[63] There are many reviews in the literature reporting the characteristics of each one, and articles comparing them.[63,76–78] The most applied is the C18 column, but in recent applications, HILIC has been reported as being complementary, since the polar compounds non-separated by C18 can be analyzed, increasing the chemical information due to the lower ion suppression in ESI. It is composed of silica chemically modified by, for example, diol, amide, zwitterionic and aminopropyl groups, or polymers with these polar groups.[79] Generally, the mobile phase is composed of aprotic solvent and water, and is used to separate polar compounds (peptides, oligosaccharides, amino sugars, amino acids, phosphorylated compounds, sugar nucleotides and others); non-polar compounds (lipids) have also shown efficiency. However, HILIC has shown some disadvantages that deserve attention, including the lack of reproducibility and slow equilibration of the column, mainly when buffers are used as mobile phases.[63,79,80]

Another strategy to enlarge the metabolome coverage is the combination of multiple orthogonal and complementary analytical analyses, such as the combination of GC-MS and LC-MS to obtain chemical information on non- and polar constituents, respectively.[81] Dunn and collaborators[53] reported serum analyses using the combination of GC and UHPLC-MS, increasing the detection and identification of the metabolites, where 3000 metabolites peaks were putatively annotated. These techniques are complementary and each technique is best suited for specific metabolite classes. Büscher and coworkers[82] compared six separation methods using a mixture of ninety-one metabolites and found that liquid-phase separation systems can handle large polar metabolites, but suggest that complementing with GC can achieve the best detection of the complete mix.

3.5 Data Acquisition

As presented in Chapter 2, a wide array of MS technologies are available (ionization sources and analyzers), and their advantages and disadvantages have been discussed, as well as their applications, to answer different questions in metabolomics. Thus, a better mass spectrometer instrument should also be selected for each analyte class, also carefully selecting the analyzer and ionization sources, such as EI, chemical ionization (CI), ESI, MALDI, atmospheric pressure CI, desorption ESI (DESI), laser ablation ESI, *etc.*[37,83] (see Chapter 2). The selection of ionization source should be performed based on the nature of the analytes, with respect to polarity, solubility, molecular mass and thermal stability, while for analyzer selection other issues should be considered, such as mass range, scan speed, mass resolution, dynamic range, *etc.*, as described in Chapter 2.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

For metabolomics analyses, the data can be acquired directly by MS without chromatographic separation, or after the separation using GC or LC.[37,84–86] For direct injection MS (DIMS), the analyte mixtures are inserted directly into the ionization source without separation of the constituents. Several studies have been reported that have applied DIMS as a fast metabolomics tool,[37] for example for fungi, plants, *etc.*[87,88] DIMS analysis is less informative than LC-MS, and normally ion suppression or enhancement can occur in ESI and the inability to distinguish isomers is also drawback of this technique.[37,88] The infusion of the ESI source can be done by continuous flow injection or by loop injection,[88] which is important in the use of high resolution analyzers because ions with same nominal mass and different exact mass can be separated, and the calculation of empirical formula can also be acquired from accurate mass, though low-resolution instruments have also been employed.[37,84,85] Beyond high-resolution analyzers, direct injection has also been performed by IMS to supply the separation of ions, increasing the chemical information from data.[86,89,90]

Previously, the first plant metabolic fingerprinting protocol using MALDI-TOF MS was described, demonstrating lower ion suppression than ESI.[91] In addition, polar and non-polar compounds can be analyzed using the same instrument, increasing the chemical information obtained and the possibility of more holistic untargeted metabolomics approaches. In this study, the taxonomic classification based on plant metabolic fingerprints showed 92% similarity to the literature classification, which highlights the huge potential of MALDI for metabolomics.[91,92]

The most applicable method in metabolomics operates MS with chromatography to separate the metabolites, such as GC-MS and LC-MS. GC-MS was the first approach used in metabolomics studies, and there are several protocols and reviews that describe the detailed methodology.[93–96] Recently, procedures have been described by integrated data from GC-MS and LC-MS, including analyses from serum and plasma.[53,97]

LC-MS with ESI is an approach widely used in metabolomics for semi-polar to polar compounds and the derivation is not necessary before the analyses. This methodology shows advantages, compared to DIMS, such as the reduction of matrix effects and ion suppression, separation of isomers, more precise information of quantification and additional data (retention time).[85] Cajka and Fiehn[98] reviewed the TOF and orbital ion trap, and quadrupole (Q)-TOF dominates untargeted metabolomics. The authors highlighted that recent improvements allow better detection limits and increasing mass-resolving power improves co-eluting isobar compounds in complex samples.

The experimental parameters must be carefully developed and revised. In this chapter, only the issues relating to MS were explored and the most important considerations discussed. Thus, the data acquisition is normally performed by the centroid mode, even with the loss of information relating to mass peak shape and purity, because the raw data files are reduced to a useful size after processing.[10,99]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

Before the data acquisition, the MS should be stabilized and calibrated, mainly for the high resolution analyzer Q-TOF, as well as the method needing to be optimized in relation to resolution and accuracy. The required resolution for Q-TOF should be higher than 8500 at full width at half maximum and the mass range is m/z 80–1200. The MS mode is preferred for the metabolomics data, since the alignment of mass peaks is easily performed by software, such as XCMS.[100] The MS/MS data can be posteriorly acquired for the chemical identification of metabolites.[10,85,99]

The injection of samples into the system should be in randomized order, avoiding time-dependent changes. For this reason, quality control (QC) can be used to evaluate the technical reproducibility and applied if corrections of variations are necessary. The QC is normally prepared from a homogeneous pooled material of samples or from a blank matrix spiked with analytes, and it is analyzed throughout the analytical runs to check the performance of the method and to align the dataset.[101,102]

The instrumental and experimental effects in artifact production are other relevant points in metabolomics studies that may distort the results. Some artifacts can be controlled by good experimental and sample-handling methods, but there are peculiar artifacts of the analytical techniques that need to be detected, and thus allow appropriate correction before the statistical analysis. Basically, in LC-MS analysis there are two artifact types: chemical noise (variations related to the instrument, such as ion suppression or enhancement) and introduced sample differences (by experimental method or data processing). The artifacts can occur due to the following effects: carry over and sample decomposition in the autosampler, background peaks, built-up or washed-up contamination, sensitivity changes, saturation (in MS), and ion suppression or enhancement.[102,103]

3.6 Data Processing

Data analysis is becoming central in omics studies. Data analysis begins with the experimental design, having in mind the study goals and the metabolite classes, data structure and sources of noise that can arise from a specific experimental setup (Figure 3.2). Great advances are being made in standard analytical methods, data standards, and data sharing and reproducibility.

All the steps of metabolomics analysis workflow have been greatly improved since its early description.[104] Systematic data analysis is gradually helping to delimitate the metabolome,[37] showing the extension of metabolites sampled by a given analytical workflow[105] and linking it to phenotypes of interest. The main challenges that still remain are the standards for reporting findings in metabolomics studies, data publication and the confidence in identification.[106,107]

Metabolomics experiments result in large datasets (Figure 3.4-L1), generally with a larger number of features. The main steps for a data analysis preprocessing workflow are peak picking and filtering, grouping peaks throughout samples and retention time correction (Figure 3.4-L2).

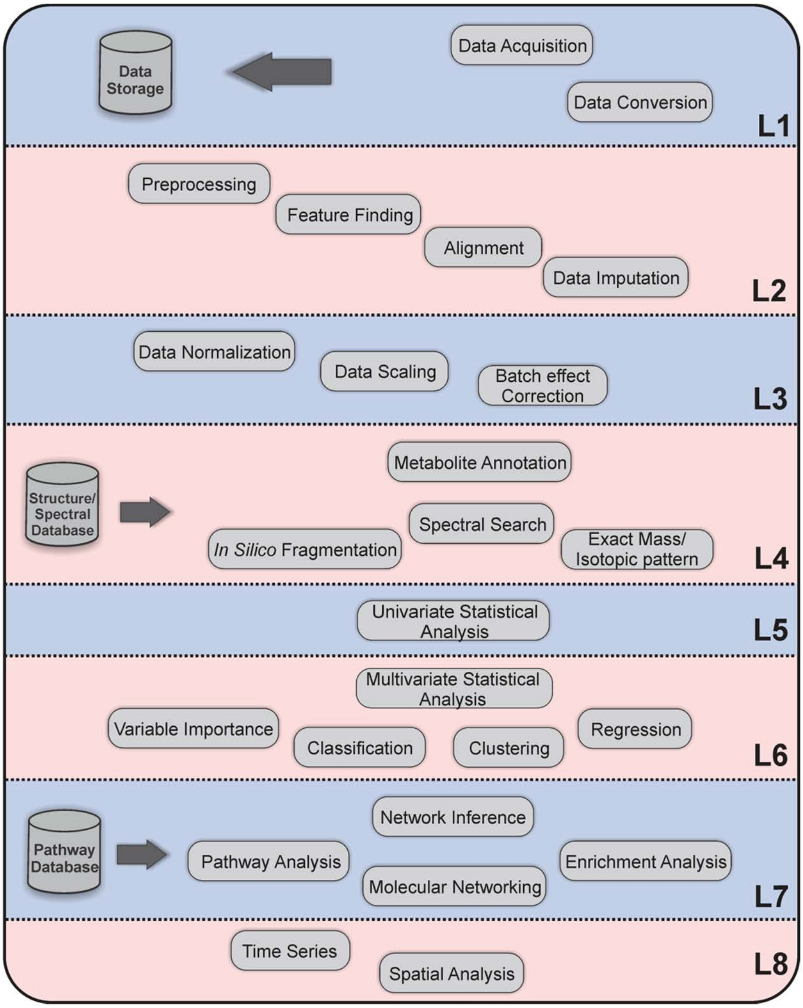

Figure 3.4 Layers of metabolomics data analysis (L1–L8). Each layer shows an increasing level of complexity that is often dependent on the previous layer. See the discussion in the text and the details in Table 3.1.

Most commercial MSs provide preprocessing software such as MarkerLynx™, Mass Profiler Professional (Agilent Technologies), and MassHunter (Agilent Technologies). The scientific community has also developed extensive open source platforms, for example OpenMS, XCMS, MZmine 2, MetAlign and IDEOM.[108] Several stand-alone packages, for processing specific analytical steps, are also available in large scientific programming communities, such as MatLab, R and python.[109]

Preprocessing analysis is under constant development and a comparative assessment of methods is being carried out by the bioinformatics community.[110] For the general peak-picking analytical step, one of the most critical preprocessing steps, it is generally difficult to optimize the parameters to find peaks in all regions of the chromatogram, or even inside a single spectrum, in the case of direct injection by ESI or MALDI.[111] Some assessments point to the combined use of multiple peak-picking tools.[112,113] Another important aspect is the programmatic optimization of peak-picking parameters and standard/QC-aided feature extraction.[25,26,114] Hughey and collaborators[26] reported on how the use of QC statistics can improve data processing, including molecule feature extraction and alignment across samples.

Other important preprocessing is the grouping of isotopes, fragments and adducts.[115–117] After preprocessing a series of multivariate data analyses are required to relate metabolome changes to phenotypes of interest. A series of steps, to extract noise, correct for experimental bias (Figure 3.4-L4) and produce unbiased statistical summaries (Figure 3.4-L5,L6), are required. Table 3.1 gives a list of visualization strategies that can be used to summarize the data, detect quantitative changes between groups of samples as well as to detect the statistical significance of these changes. Depending on the required result, different techniques can be applied to visualize global metabolomics (Table 3.1), such as heatmaps, scatter plots, sores and loading plots, volcano plots and designed cloud plots.[118] The heatmaps, for example, can be used to elucidate biological properties, correlating chemical and biological information, which assists in the identification of synergism between the components, determination of pro-drugs and possible active substances.[119] As highlighted above, one of the great challenges in the field is metabolite annotation (Figure 3.4-L4). Table 3.1 also presents important visualization tools to aid metabolite annotation.[120]

New tools are emerging in the metabolomics field with the new requirements from the scientific community for reproducibility and data sharing practices. Online analytical platforms, integrating multiple data analysis tools such as GNPS, Metaboanalyst and XCMS online[100,121,122] provide a large array of processing, multivariate statistical analyses and metabolite annotation options (Figure 3.4-L7,L8). Open source and open development infrastructures such as Galaxy and KNIME environments[123,124] enable developers to readily integrate new modules. These platforms also provide the provenance of all tools used in each data analysis step, allowing higher reproducibility of data analysis results. In the metabolomics field the projects Workflow4Metabolomics and OpenMS[125,126] provide growing platforms in these environments. On the data sharing side (see Chapter 10) emerging tools are allowing users to share raw experimental data, in addition to processed data, fostering reproducibility, and allowing data scientists to learn from data and to expand inter-experiment comparison to increase the number of annotated compounds. The emerging concepts presented by the GNPS platform (http://gnps.ucsd.edu/), with a community of collaborators increasingly contributing with new fragmentation spectra and the new paradigm of “living data”, in which raw datasets are routinely re-analyzed and the results are communicated to subscribed users, are expected to have a huge impact on metabolomics applications. Another interesting functionality is the possibility of user

Table 3.1 Common visualization methods used for metabolomics data analysis.

Aim	Tools/methods	Advantage	Drawback	Visualization	Ref.
Convert data to standard formats	msConvert	User interface for multiple instrument vendors	Need proprietary software library/Windows dependent	Total ion chromatogram, base peak chromatogram	Chambers *et al.*[129]
Detect and quantify ions and align along samples	xcms, MZmine, MetSign, OpenMS, Optimus, AMDIS, TargetSearch, TagFinder	Data reduction, retention time correction, alignment along samples, missing value replacement	Highly dependent manual setup of parameters, very sensitive to noise detection, split and merge features	Heatmap, extracted ion chromatogram	Lange *et al.*;[110] Chen *et al.*;[112] Coble and Fraga[113]
Correct for experimental bias/transform data to downstream analysis	Standard normalization, TIC normalization, quantile normalization	Extract experimental effect, improve statistical analysis	Dependent on study and target analysis, use of wrong method can give misleading answer to statistical analysis	Heatmap	Alonso *et al.*;[115] Kuhl *et al.*;[116] Treutler and Neumann[117]
Assign putative identity to *m*/*z*	MassBank, Metlin, GnPS, MetFrag, Sirius, CSI-FingerID	Allows association to the studied biological model and generation of hypothesis for follow-up experiments	Low level of confidence for annotations based on MS1 (accurate mass, isotopic pattern, retention time), and restricted fragmentation libraries EI and MSMS	Mirror plot, fragmentation tree	Böcker and Dührkop;[130] Wang *et al.*[122]
Detection of statistically significant alterations	MetaboAnalyst, XCMS online, MZmine, R, MatLab, QIIME/ ANOVA, Kruskal-Wallis, rank product	Clear resolution at single feature level	Increased false discovery rate with multiple tests	Heatmap, volcano plot, cloud plot, box plot, extracted ion chromatogram	Tautenhahn *et al.*;[100] Vinaixa *et al.*;[131] Barnes *et al.*[132]

(*continued*)

Table 3.1 (*continued*)

Aim	Tools/methods	Advantage	Drawback	Visualization	Ref.
Detect patterns with data, internal structure (unsupervised learning) or train models to learn from data labels (supervised learning)	MetaboAnalyst, XCMS online, Mzmine, R, MatLab, QIIME/ hierarchical clustering, PCA, PCoA, PLS-DA, Random forest	High dimension summaries, extract linear/non-linear weighted contribution of all variables	No clear resolution at single feature level, subject to constraints of model derivation that don't fit biological model	Sore plots, loading plots, latent variable plots, dendrograms	Harvey *et al.*;[119] Xia *et al.*,[121]
Detect relationship between samples/ features	KEGG, WikiPathways, MetaCyc, GnPS, Cytoscape/correlation, MS2 cosine, partial correlation, structural similarity, enzymatic reaction	Detect single relationships and global patterns in a hierarchical fashion in data structure that better reflects the model of metabolism	Pathways are convoluted, hard to define biologically/chemically meaningful thresholds	Pathway maps, weighted networks (directed, undirected), bipartite graphs	Tautenhahn *et al.*;[100] Xia, *et al.*;[121] Wang *et al.*;[112]
Integration of time and space scales to upstream workflow	illi, MetaboAnalyst	Association of space and time defined events to biological model	Metabolic changes happen at much lower time (fraction of seconds) and Space (nm) scales than current experimental resolution	Heatmap overlaid on 2d, 3d pictures or models, line and bar plots over time	Tautenhahn *et al.*;[100] Xia, *et al.*;[121] Bouslimani *et al.*;[133] Wang *et al.*;[112]

ranking of community deposited spectra. These improvements are likely to decrease the gap between public data and tools compared to the genomics field.[122,127]

The main goal of metabolomics is the understanding of phenotypical changes through unbiased data analysis interpretation. To achieve this goal, an integrated approach from experimental design to data analysis is needed. With the advances in the best experimental and data sharing practices, the metabolomics community will be ready for more complex challenges, integrating multiomics analysis into system biology interpretation of metabolic networks.[128]

References

1. J. Nielsen, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, M. A. E. Hansen and U. Roessner-Tunali, John Wiley & Sons, 2007, vol. 24, pp. 3–17.
2. P. Eneroth, B. Gordon, R. Ryhage and J. Sjovall, *J. Lipid Res.*, 1966, **7**, 511–523.
3. S. Oliver, *Trends Biotechnol.*, 1998, **16**, 373–378.
4. R. D. Hall, *New Phytol.*, 2006, **169**, 453–468.
5. J. Smedsgaard, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, M. A. E. Hansen and U. Roessner-Tunali, John Wiley & Sons, 2007, vol. 24, pp. 83–145.
6. S. Moco, J. Vervoort, S. Moco, R. J. Bino, R. C. H. De Vos and R. Bino, *TrAC, Trends Anal. Chem.*, 2007, **26**, 855–866.
7. M. M. W. B. Hendriks, F. A. van Eeuwijk, R. H. Jellema, J. A. Westerhuis, T. H. Reijmers, H. C. J. Hoefsloot and A. K. Smilde, *TrAC, Trends Anal. Chem.*, 2011, **30**, 1685–1698.
8. F. Allen, R. Greiner and D. Wishart, *Metabolomics*, 2014, **11**, 98–110.
9. D. C. Sévin, A. Kuehne, N. Zamboni and U. Sauer, *Curr. Opin. Biotechnol.*, 2015, **34**, 1–8.
10. T. F. Jorge, J. A. Rodrigues, C. Caldana, R. Schmidt, J. T. van Dongen, J. Thomas-Oates and C. António, *Mass Spectrom. Rev.*, 2016, **35**, 620–649.
11. J. Fraser, I. Williamson, W. A. Bickmore and J. Dostie, *Microbiol. Mol. Biol. Rev.*, 2015, **79**, 347–372.
12. C. A. Hutchison, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson and J. C. Venter, *Science*, 2016, **351**, aad6253.
13. E. V. Koonin, *Biol. Direct*, 2012, **7**, 1–7.
14. K. W. George, A. Chen, A. Jain, T. S. Batth, E. E. K. Baidoo, G. Wang, P. D. Adams, C. J. Petzold, J. D. Keasling and T. S. Lee, *Biotechnol. Bioeng.*, 2014, **111**, 1648–1658.

15. S.-Y. Shin, E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector and N. Soranzo, *Nat. Genet.*, 2014, **46**, 543–550.
16. G. Larrouy-Maumus, T. Biswas, D. M. Hunt, G. Kelly, O. V. Tsodikov and L. P. S. de Carvalho, *Proc. Natl. Acad. Sci.*, 2013, **110**, 11320–11325.
17. T. Kind, M. Scholz and O. Fiehn, *PLoS One*, 2009, **4**, e5440.
18. T. Kind and O. Fiehn, *Bioanal. Rev.*, 2010, **2**, 23–60.
19. J. E. Peironcely, T. Reijmers, L. Coulier, A. Bender and T. Hankemeier, *PLoS One*, 2011, **6**, e28966.
20. J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes and C. Steinbeck, *Nucleic Acids Res.*, 2015, gkv1031.
21. A. Mohamed, C. H. Nguyen and H. Mamitsuka, *Briefings Bioinf.*, 2015, **17**, 309–321.
22. U. Roessner, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, M. A. E. Hansen and U. Roessner-Tunali, John Wiley & Sons, 2007, vol. 24, pp. 15–38.
23. Y. H. Choi and R. Verpoorte, *Phytochem. Anal.*, 2014, **25**, 289–290.
24. T. Cajka and O. Fiehn, *Anal. Chem.*, 2015, **88**, 524–545.
25. L. Duan, I. Molnár, J. H. Snyder, G.-A. Shen, X. Qi, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro and Y. Li, *et al.*, *Mol. Plant*, 2016, **9**, 1217–1220.
26. C. A. Hughey, C. M. McMinn and J. Phung, *Metabolomics*, 2015, **12**, 11.
27. FDA, *Dep. Health Hum. Serv. Food Drug Adm.*, 2001.
28. J. S. Morris, K. A. Baggerly, H. B. Gutstein and K. R. Coombes, *Methods Mol. Biol.*, 2010, **641**, 143–166.
29. L. S. Riter, O. Vitek, K. M. Gooding, B. D. Hodge and R. K. Julian, *J. Mass Spectrom.*, 2005, **40**, 565–579.
30. H. G. Gika, G. A. Theodoridis, R. S. Plumb and I. D. Wilson, *J. Pharm. Biomed. Anal.*, 2014, **87**, 12–25.
31. D. B. Hibbert, *J. Chromatogr. B*, 2012, **910**, 2–13.
32. E. S. Hecht, A. L. Oberg and D. C. Muddiman, *J. Am. Soc. Mass Spectrom.*, 2016, **27**, 767–785.
33. A. Tebani, I. Schmitz-Afonso, D. N. Rutledge, B. J. Gonzalez, S. Bekri and C. Afonso, *Anal. Chim. Acta*, 2016, **913**, 55–62.
34. G. Nyamundanda, I. C. Gormley, Y. Fan, W. M. Gallagher and L. Brennan, *BMC Bioinf.*, 2013, **14**, 338.
35. B. J. Blaise, *Anal. Chem.*, 2013, **85**, 8943–8950.
36. B. J. Blaise, G. Correia, A. Tin, J. H. Young, A.-C. Vergnaud, M. Lewis, J. T. M. Pearce, P. Elliott, J. K. Nicholson, E. Holmes and T. M. D. Ebbels, *Anal. Chem.*, 2016, **88**, 5179–5188.
37. M. Ernst, D. B. Silva, R. R. Silva, R. Z. N. Vêncio and N. P. Lopes, *Nat. Prod. Rep.*, 2014, **31**, 784–806.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

38. D. Vuckovic, *Anal. Bioanal. Chem.*, 2012, **403**, 1523–1548.
39. C. Bylda, R. Thiele and U. Kobold, *et al.*, *Analyst*, 2014, **139**, 2265.
40. S. G. Villas-Boas, in *Metabolome Analysis: An Introduction*, ed. S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, M. A. E. Hansen and U. Roessner-Tunali, John Wiley & Sons, 2007, vol. 24, pp. 39–82.
41. H. K. Kim, Y. H. Choi and R. Verpoorte, *Nat. Protoc.*, 2010, **5**, 536–549.
42. N. Li, Y. peng Song, H. Tang and Y. Wang, *Arch. Biochem. Biophys.*, 2016, **589**, 4–9.
43. P. Yin, R. Lehmann and G. Xu, *Anal. Bioanal. Chem.*, 2015, **407**, 4879–4892.
44. Y. Wang, S. Liu and Y. Hu, *et al.*, *Anal. Chem.*, 2011, **83**, 6902–6906.
45. N. D. Yuliana, A. Khatib, R. Verpoorte and Y. H. Choi, *Anal. Chem.*, 2011, **83**, 6902–6906.
46. O. Deda, H. G. Gika, I. D. Wilson and G. A. Theodoridis, *J. Pharm. Biomed. Anal.*, 2015, **113**, 137–150.
47. Z. Ser, X. Liu, N. N. Tang and J. W. Locasale, *Anal. Biochem.*, 2015, **475**, 22–28.
48. M. D. Luque de Castro and M. M. Delgado-Povedano, *Anal. Chim. Acta*, 2014, **806**, 74–84.
49. L. Whiley, J. Godzien, F. J. Ruperez, C. Legido-Quigley and C. Barbas, *Anal. Chem.*, 2012, **84**, 5992–5999.
50. B. Bojko, N. Reyes-Garcés, V. Bessonneau, K. Goryński, F. Mousavi, E. A. Souza Silva and J. Pawliszyn, *TrAC, Trends Anal. Chem.*, 2014, **61**, 168–180.
51. E. Boyacı, Á. Rodríguez-Lafuente, K. Gorynski, F. Mirnaghi, É. A. Souza-Silva, D. Hein and J. Pawliszyn, *Anal. Chim. Acta*, 2015, **873**, 14–30.
52. S. Naz, M. Vallejo, A. García and C. Barbas, *J. Chromatogr. A*, 2014, **1353**, 99–105.
53. W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall, J. N. Haselden, A. W. Nicholls, I. D. Wilson, D. B. Kell, R. Goodacre and T. H. S. M. (HUSERMET) Consortium, *Nat. Protoc.*, 2011, **6**, 1060–1083.
54. N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend and D. S. Wishart, *PLoS One*, 2011, **6**, e16957.
55. R. E. Patterson, A. J. Ducrocq, D. J. McDougall, T. J. Garrett and R. A. Yost, *J. Chromatogr. B*, 2015, **1002**, 260–266.
56. R. Zenobi, *Science*, 2013, **342**, 1243259.
57. B. B. Misra, S. M. Assmann and S. Chen, *Trends Plant Sci.*, 2014, **19**, 637–646.
58. T. Tohge, M. S. Ramos, A. Nunes-Nesi, M. Mutwil, P. Giavalisco, D. Steinhauser, M. Schellenberg, L. Willmitzer, S. Persson, E. Martinoia and A. R. Fernie, *Plant Physiol.*, 2011, **157**, 1469–1482.

59. T. Fujii, S. Matsuda, M. L. Tejedor, T. Esaki, I. Sakane, H. Mizuno, N. Tsuyama and T. Masujima, *Nat. Protoc.*, 2015, **10**, 1445–1456.
60. M. Lorenzo Tejedor, H. Mizuno, N. Tsuyama, T. Harada and T. Masujima, *Anal. Chem.*, 2012, **84**, 5221–5228.
61. Y. Iwasaki, T. Sawada, K. Hatayama, A. Ohyagi, Y. Tsukuda, K. Namekawa, R. Ito, K. Saito and H. Nakazawa, *Metabolites*, 2012, **2**, 496–515.
62. T. Fuhrer and N. Zamboni, *Curr. Opin. Biotechnol.*, 2015, **31**, 73–78.
63. J.-L. Wolfender, G. Marti, A. Thomas and S. Bertrand, *J. Chromatogr. A*, 2015, **1382**, 136–164.
64. S. Fekete, J. Schappler and J.-L. Veuthey, *TrAC, Trends Anal. Chem.*, 2014, **63**, 2–13.
65. E. M. Borges, M. A. Rostagno and M. A. A. Meireles, *et al.*, *RSC Adv.*, 2014, **4**, 22875.
66. R. Preti, *Int. J. Anal. Chem.*, 2016, **2016**, 3189724.
67. N. Ishizuka, H. Kobayashi, H. Minakuchi, K. Nakanishi, K. Hirao, K. Hosoya, T. Ikegami and N. Tanaka, *J. Chromatogr. A*, 2002, **960**, 85–96.
68. M. Motokawa, H. Kobayashi, N. Ishizuka, H. Minakuchi, K. Nakanishi, H. Jinnai, K. Hosoya, T. Ikegami and N. Tanaka, *J. Chromatogr. A*, 2002, **961**, 53–63.
69. G. Guiochon, *J. Chromatogr. A*, 2007, **1168**, 101–168.
70. M. Wilm and M. Mann, *Anal. Chem.*, 1996, **68**, 1–8.
71. J. R. Yates, C. I. Ruse and A. Nakorchevsky, *Annu. Rev. Biomed. Eng.*, 2009, **11**, 49–79.
72. F. Pereira, X. Niu and A. J. deMello, *et al.*, *PLoS One*, 2013, **8**, e63087.
73. E. Mirgorodskaya, C. Braeuer, P. Fucini, H. Lehrach and J. Gobom, *Proteomics*, 2005, **5**, 399–408.
74. T. Zhang, D. G. Watson and D. Ryan, *et al.*, *Analyst*, 2015, **140**, 2907–2915.
75. G. B. Gonzales, K. Raes, S. Coelus, K. Struijs, G. Smagghe and J. Van Camp, *J. Chromatogr. A*, 2014, **1323**, 39–48.
76. L. Nováková and H. Vlčková, *Anal. Chim. Acta*, 2009, **656**, 8–35.
77. T. L. Chester, *Anal. Chem.*, 2013, **85**, 579–589.
78. J. J. Pesek, R. I. Boysen and M. T. W. Hearn, *et al.*, *Anal. Methods*, 2014, **6**, 4496.
79. P. Jandera, *Anal. Chim. Acta*, 2011, **692**, 1–25.
80. Y. Guo and S. Gaiki, *J. Chromatogr. A*, 2011, **1218**, 5920–5938.
81. W. B. Dunn, N. J. C. Bailey and H. E. Johnson, *Analyst*, 2005, **130**, 606–625.
82. J. M. Bu̇scher, D. Czernik, J. C. Ewald, U. Sauer and N. Zamboni, *Anal. Chem.*, 2009, **81**, 2135–2143.
83. M. de Raad, C. R. Fischer and T. R. Northen, *Curr. Opin. Chem. Biol.*, 2016, **30**, 7–13.
84. K. Dettmer, P. A. Aronov and B. D. Hammock, *Mass Spectrom. Rev.*, 2007, **26**, 51–78.
85. Z. Lei, D. V. Huhman and L. W. Sumner, *J. Biol. Chem.*, 2011, **286**, 25435–25442.

86. A. Mastrangelo, A. Ferrarini, F. Rey-Stolle, A. García and C. Barbas, *Anal. Chim. Acta*, 2015, **900**, 21–35.
87. J. I. Castrillo, A. Hayes, S. Mohammed, S. J. Gaskell and S. G. Oliver, *Phytochemistry*, 2003, **62**, 929–937.
88. A. Koulman, B. A. Tapper, K. Fraser, M. Cao, G. A. Lane and S. Rasmussen, *Rapid Commun. Mass Spectrom.*, 2007, **21**, 421–428.
89. P. Dwivedi, P. Wu, S. J. Klopsch, G. J. Puzon, L. Xun and H. H. Hill, *Metabolomics*, 2008, **4**, 63–80.
90. R. Cumeras, E. Figueras and C. E. Davis, *et al.*, *Analyst*, 2015, **140**, 1391–1410.
91. M. Ernst, D. B. Silva, R. Silva, M. Monge, J. Semir, R. Z. N. Vêncio and N. P. Lopes, *Anal. Chim. Acta*, 2015, **859**, 46–58.
92. R. Silva, N. P. Lopes and D. B. Silva, *Planta Med.*, 2016, **82**, 671–689.
93. O. Fiehn, *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2016, pp. 30.4.1–30.4.32.
94. H. Tsugawa, Y. Tsujimoto, K. Sugitate, N. Sakui, S. Nishiumi, T. Bamba and E. Fukusaki, *J. Biosci. Bioeng.*, 2014, **117**, 122–128.
95. B. S. Mitrevski, K. A. Kouremenos and P. J. Marriott, *Bioanalysis*, 2009, **1**, 367–391.
96. H. H. Kanani and M. I. Klapa, *Metab. Eng.*, 2007, **9**, 39–51.
97. A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. J. C. van der Werff-van der Vat and R. H. Jellema, *Anal. Chem.*, 2005, **77**, 6729–6736.
98. T. Cajka and O. Fiehn, *Anal. Chem.*, 2015, **88**, 524–545.
99. R. C. H. De Vos, S. Moco, A. Lommen, J. J. B. Keurentjes, R. J. Bino and R. D. Hall, *Nat. Protoc.*, 2007, **2**, 778–791.
100. R. Tautenhahn, G. J. Patti, D. Rinehart and G. E. Siuzdak, *Anal. Chem.*, 2012, **84**, 5035–5039.
101. J. Godzien, V. Alonso-Herranz, C. Barbas and E. G. Armitage, *Metabolomics*, 2015, **11**, 518–528.
102. T. Sangster, H. Major, R. Plumb, A. J. Wilson and I. D. Wilson, *Analyst*, 2006, **131**, 1075.
103. L. Burton, G. Ivosev, S. Tate, G. Impey, J. Wingate and R. Bonner, *J. Chromatogr. B*, 2008, **871**, 227–235.
104. M. Brown, W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasić and D. B. Kell, *Metabolomics*, 2005, **1**, 39–51.
105. D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett and K. E. V. Burgess, *Anal. Chem.*, 2011, **83**, 8703–8710.
106. D. J. Creek, W. B. Dunn, O. Fiehn, J. L. Griffin, R. D. Hall, Z. Lei, R. Mistrik, S. Neumann, E. L. Schymanski, L. W. Sumner, R. Trengove and J.-L. Wolfender, *Metabolomics*, 2014, **10**, 350–353.
107. W. B. Dunn, A. Erban, R. J. M. Weber, D. J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka and M. R. Viant, *Metabolomics*, 2013, **9**, 44–66.
108. B. B. Misra and J. J. J. van der Hooft, *Electrophoresis*, 2016, **37**, 86–110.
109. R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, New York, 2005, vol. 746718470.

110. E. Lange, R. Tautenhahn, S. Neumann and C. Gröpl, *BMC Bioinf.*, 2008, **9**, 375.
111. S. Gibb and K. Strimmer, *Bioinformatics*, 2012, **28**, 2270–2271.
112. Y. Chen, J. Xu, R. Zhang, G. Shen, Y. Song, J. Sun, J. He, Q. Zhan and Z. Abliz, *Analyst*, 2013, **138**, 2669–2677.
113. J. B. Coble and C. G. Fraga, *J. Chromatogr. A*, 2014, **1358**, 155–164.
114. G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber and C. Magnes, *BMC Bioinf.*, 2015, **16**, 118.
115. A. Alonso, A. Julià, A. Beltran, M. Vinaixa, M. Díaz, L. Ibañez, X. Correig and S. Marsal, *Bioinformatics*, 2011, **27**, 1339–1340.
116. C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson and S. Neumann, *Anal. Chem.*, 2011, **84**, 283–289.
117. H. Treutler and S. Neumann, *Metabolites*, 2016, **6**, 37.
118. J. Ivanisevic, H. P. Benton, D. Rinehart, A. Epstein, M. E. Kurczy, M. D. Boska, H. E. Gendelman and G. Siuzdak, *Metabolomics*, 2014, **11**, 1029–1034.
119. A. L. Harvey, R. Edrada-Ebel and R. J. Quinn, *Nat. Rev. Drug Discovery*, 2015, **14**, 111–129.
120. R. R. da Silva, P. C. Dorrestein and R. A. Quinn, *Proc. Natl. Acad. Sci.*, 2015, 12549–12550.
121. J. Xia, I. V. Sinelnikov, B. Han and D. S. Wishart, *Nucleic Acids Res.*, 2015, **43**, W251–W257.
122. M. Wang, J. J. Carver and V. V. Phelan, *et al.*, *Nat. Biotechnol.*, 2016, **34**, 828–837.
123. E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko and J. Goecks, *Nucleic Acids Res.*, 2016, **44**, W3–W10.
124. S. Aiche, T. Sachsenberg, E. Kenar, M. Walzer, B. Wiswedel, T. Kristl, M. Boyles, A. Duschl, C. G. Huber, M. R. Berthold, K. Reinert and O. Kohlbacher, *Proteomics*, 2015, **15**, 1443–1447.
125. F. Giacomoni, G. Le Corguillé, M. Monsoor, M. Landi, P. Pericard, M. Pétéra, C. Duperier, M. Tremblay-Franco, J.-F. Martin, D. Jacob, S. Goulitquer, E. A. Thévenot and C. Caron, *Bioinformatics*, 2015, **31**, 1493–1495.
126. H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert and O. Kohlbacher, *Nat. Methods*, 2016, **13**, 741–748.
127. J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1743–E1752.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

128. C. H. Johnson, J. Ivanisevic and G. Siuzdak, *Nat. Rev. Mol. Cell Biol.*, 2016, **17**, 451–459.
129. M. C. Chambers, B. Maclean and R. Burke, *et al.*, *Nat. Biotechnol.*, 2012, **30**, 918–920.
130. S. Böcker and K. Dührkop, *J. Cheminf.*, 2016, **8**, 5.
131. M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart and O. Yanes, *Metabolites*, 2012, **2**, 775–795.
132. S. Barnes, H. P. Benton, K. Casazza, S. J. Cooper, X. Cui, X. Du, J. Engler, J. H. Kabarowski, S. Li, W. Pathmasiri, J. K. Prasain, M. B. Renfrow and H. K. Tiwari, *J. Mass Spectrom.*, 2016, **51**, 535–548.
133. A. Bouslimani, C. Porto, C. M. Rath, M. Wang, Y. Guo, A. Gonzalez, D. Berg-Lyon, G. Ackermann, G. J. Moeller Christensen, T. Nakatsuji, L. Zhang, A. W. Borkowski, M. J. Meehan, K. Dorrestein, R. L. Gallo, N. Bandeira, R. Knight, T. Alexandrov and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E2120–E2129.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00057

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

CHAPTER 4

Proteomics

SWATI VARSHNEY[a,b], TRAYAMBAK BASAK[c], MAINAK DUTTA[a,d] AND SHANTANU SENGUPTA*[a,b]

[a]Genomics and Molecular Medicine unit, CSIR-Institute of Genomics and Integrative Biology, Sukhdev Vihar, Mathura Road, New Delhi 110 020, India; [b]Academy of Scientific & Innovative Research (AcSIR), CSIR-IGIB South Campus, New Delhi 110 020, India; [c]Department of Medicine, Division of Nephrology and Hypertension, Vanderbilt University Medical Center, Nashville, Tennessee 37232, United States; [d]Birla Institute of Technology Pilani, Dubai Campus, Dubai International Academic City, P. O. Box No. 345055, Dubai, United Arab Emirates
*E-mail: shantanus@igib.res.in

4.1 Introduction to Proteomics

In 1838, Jacob Berzelius, a Swedish chemist, coined the term "protein",[1] which was derived from the Greek word "proteios" meaning "first rank" or primary, highlighting the importance of this macromolecule. Proteins are the actual functional molecules in a cell that perform diverse functions, and also receive signals from outside and activate appropriate intracellular response. In contrast to DNA, which mostly remains static in a cell and does not change in response to external and internal stimuli, mRNA and protein expression remain in a dynamic flux, and vary both qualitatively (post-translational/transcriptional modifications) and/or quantitatively (expression) with varying cellular conditions. This adds an additional level of complexity in studies involving mRNA and proteins. The variation at the mRNA level can be generated

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

due to alternate splicing, different promoter usage and RNA editing, while variation at the protein level can be due to unconventional use of start and stop codons, or post translational modifications. Thus, a gene can give rise to multiple mRNAs and proteins, rendering the study of the entire repertoire of proteins in a cell difficult. Most of the studies in the early '90s were hypothesis driven, where a hypothesis was put forward and, based on the "candidate" genes and their products (mRNA or protein), were studied in-depth. However, the advent of high-throughput technologies in the last couple of decades has enabled the study of genes and proteins at a global scale, and with this started what can be termed as an "omics" era where the focus was to create comprehensive data sets using a hypothesis free approach, thus removing study bias. Although the term "ome" was used in the initial half of the 20th century, it became popular when molecular biologists started using the term for studies that collectively considered all the constituents of a particular topic. For instance, while the term "genome" was known as early as 1926, "genomics" was coined by Tom Roderick in 1986 for mapping the human genome.[2] Almost a decade later, in 1994, Marc Wilkins coined the term "proteome" while the term "proteomics" was coined in 1997 [3–5] and was referred to as "the study of protein properties (expression level, post-translational modification, interactions) to obtain a global integrated view of disease processes, cellular processes and networks at the protein level".[6]

The first protein studies that could be termed as "proteomics" began with the introduction of two-dimensional (2D) gel electrophoresis (GE) in 1975 when O'Farrell, Klose and Scheele independently reported the mapping of proteins from *E. coli*, a mouse and a guinea pig, respectively.[7–9] However, during that time the proteins could not be identified although they could be separated and visualized. The first major technology to emerge for the identification of proteins was sequencing using Edman degradation, which was automated in 1967 by Edman and Beggs.[10] However, Edman degradation had its own limitations as it proceeds from the N-terminus of the protein and hence does not work for proteins whose N-terminal is blocked (or modified such as in acetylation). Further, it requires a large concentration of protein for sequencing purposes. One of the most important discoveries in the field of proteomics was mass spectrometry (MS) which enabled high-throughput identification of proteins in a complex mixture. MSs ionize the sample under investigation and separates them based on their mass-to-charge ratio, which are then detected in proportion to their abundance and a plot of abundance *versus* mass-to-charge ratio is provided. The sensitivity of MSs to identify the proteins at femtomole to attomole levels[11] and accuracy of less than 5 ppm[12–15] (parts per million) makes this the instrument of choice in proteomics. Proteomics is considered to be an important facet in the context of systems biology since it represents the actual functional molecule and gives us an idea of what is actually happening in a cell/tissue or organism, unlike the transcriptome. The abundance of a transcript does not necessarily reflect the abundance of the protein since the formation of proteins from mRNA involves a sequence of events such as post-transcriptional control and

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

post-translational modifications. Furthermore, the activity of the protein may entirely depend on its localization in the cell, which is not dependent on the transcript. In addition, protein–protein interactions, which are fundamental to many important biological processes, can be understood only from proteomics studies. Most importantly, biological fluids that are frequently used to identify disease markers or study disease progression, such as urine, plasma, serum, and cerebrospinal fluid, do not contain nucleic acid or RNA. Thus, proteomics is expected to bridge the gap between our understanding of the genome sequence and cellular behavior/function.

Proteomics studies (identification and analysis of the entire protein content) can be broadly classified into: structural proteomics, expression proteomics, and functional proteomics. Identifying proteins and understanding the structure and interactions of protein complexes or proteins present in a specific cellular organelle is an integral part of structural or "cell map" proteomics. Structural proteomics can help to understand the functions of newly discovered genes and identify residues where a drug interacts with proteins or proteins interact with each other. X-ray crystallography and nuclear magnetic resonance spectroscopy are often employed in structural proteomics. Expression proteomics is a term for studies that compare the expression of proteins between two or more samples. Protein expression profiling allows us to qualitatively and quantitatively differentiate between a normal and a variable state. Initially, expression proteomics was done using 2D GE (2-DGE). However, with the availability of highly sensitive and accurate MSs, these are being routinely used coupled with liquid chromatography (LC) systems for expression proteomics. This approach could help in the identification of novel proteins involved in signal transduction pathways or to identify disease biomarkers. The identification and characterization of protein–protein, protein–DNA, protein–RNA interactions that affect the function are termed functional proteomics. This can be used to study how protein complexes are formed and determine protein functions. This approach provides information about protein signaling, protein–drug interactions or understanding disease mechanisms.

In the last decade there has been tremendous progress in the field of proteomics, mainly due to tremendous improvements in MSs in terms of sensitivity, accuracy and resolution. However, despite these improvements, there are several challenges and limitations in the field of proteomics. A proteomics experiment involves the use of a range of technologies for large-scale protein studies and no single method is suitable for all kinds of applications. Each of these technologies has their strengths and weaknesses, and their use depends on the questions to be addressed. One of the major challenges in proteomics is to identify proteins with low abundance. The dynamic range of proteins can vary in order of magnitude between six in cells and tissues to ten in biological fluids such as plasma and serum.[16–19] The unavailability of an amplification technique akin to the polymerase chain reaction in genomics thus limits the detection of very-low-abundance proteins. There are other challenges associated with sample preparation, such as extraction of membrane proteins or protein degradation, that could be an issue in gel-based

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

proteomics. Similarly, separation of proteins with extreme pI values could be a challenge in the gel-based separation of proteins. In the area of bioinformatics one of the challenges is to identify proteins whose sequence is unavailable in the database.

A typical proteomics workflow involves various steps: separation and isolation of proteins from cells, tissues or organisms; proteolysis of the protein sample using a single or combination of proteolytic enzymes; separation of the peptide mixture using various chromatographic techniques to reduce the complexity of the sample; mass spectrometric analysis of the sample, which includes MS and MS/MS scans to generate a “peptide mass fingerprint”;[20] database search using bioinformatics tools to identify the proteins (Figure 4.1).

4.2 Sample Preparation for Proteomics Studies and Proteolysis

Sample preparation is the most critical and rate-limiting step in proteomics studies as the results of a proteomics experiment are greatly influenced by the quality of the sample. In general, it needs to be ensured that the sample preparation method entails minimal protein loss, which is achieved by using a minimal number of steps, and that protein modifications/degradation is prevented by performing all the steps at a cold temperature. Further, the proper storage of samples is important, especially in the case of biological fluids. For instance, it has been reported that various serum proteins degrade if stored for a long time even at −80 °C.[21] Sample preparation for proteomics studies depends on the type of experimental model and the complexity of the proteome, especially in terms of the dynamic range of proteins

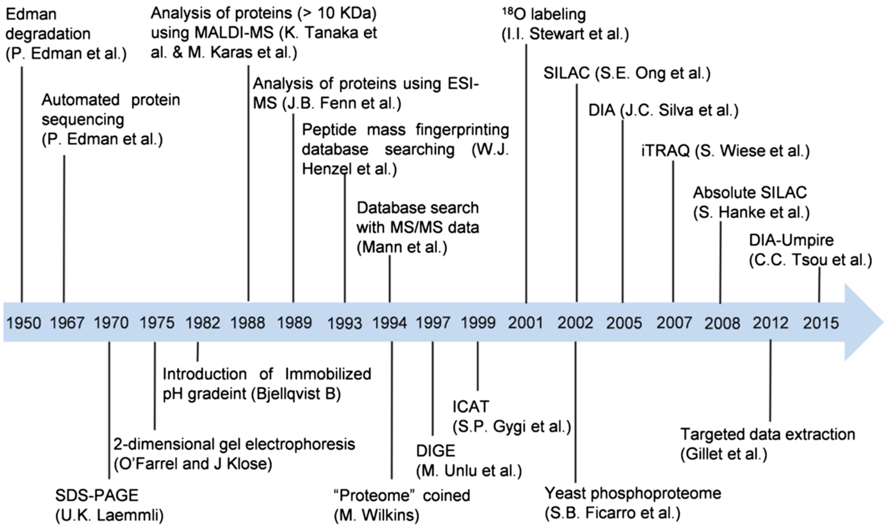
Edman degradation (P. Edman et al.)
Automated protein sequencing (P. Edman et al.)
Analysis of proteins (> 10 KDa) using MALDI-MS (K. Tanaka et al. & M. Karas et al.)
Analysis of proteins using ESI-MS (J.B. Fenn et al.)
Peptide mass fingerprinting database searching (W.J. Henzel et al.)
Database search with MS/MS data (Mann et al.)
¹⁸O labeling (I.I. Stewart et al.)
SILAC (S.E. Ong et al.)
DIA (J.C. Silva et al.)
iTRAQ (S. Wiese et al.)
Absolute SILAC (S. Hanke et al.)
DIA-Umpire (C.C. Tsou et al.)
1950
1967
1970
1975
1982
1988
1989
1993
1994
1997
1999
2001
2002
2005
2007
2008
2012
2015
Introduction of Immobilized pH gradeint (Bjellqvist B)
2-dimensional gel electrophoresis (O’Farrel and J Klose)
SDS-PAGE (U.K. Laemmli)
“Proteome” coined (M. Wilkins)
DIGE (M. Unlu et al.)
ICAT (S.P. Gygi et al.)
Yeast phosphoproteome (S.B. Ficarro et al.)
Targeted data extraction (Gillet et al.)

Figure 4.1 Timeline representing the major events in the development of proteomics.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

in the sample. The protocols followed for the extraction of proteins depends on the sample source. Animal tissues proteins are usually first extracted by properly mincing the tissues or powdering using snap freezing. The minced or powdered tissues are then lysed in a buffer containing chaotropes (*e.g.* urea, thiourea), detergents (*e.g.* CHAPS (3-[(3-cholamidopropyl)-dimethylammonio]-1-propanesulfonate hydrate), triton X-100, tween 80, nonidet P-40 (NP 40), *etc.*), reductants (*e.g.* dithiothreitol, tributylphosphine, triscarboxyethylphosphine, β-mercaptoethanol) and protease inhibitors (*e.g.* aprotinin, leupeptin, *etc.*). The lysed cells are then centrifuged and the supernatant containing proteins is stored at −80 °C. Proteins from the cell culture are isolated by mechanically or enzymatically dispersing cells followed by lysis using buffers detergents and protease inhibitors. In biological fluids such as plasma and serum only a handful of proteins (*e.g.* albumin, IgG, antitrypsin, haptoglobin, fibrinogen, apolipoprotein, *etc.*) account for more than 90% of the total protein, and hence identification of less abundant proteins becomes difficult in their presence. Thus, protein enrichment from these biological fluids often entails depletion of these highly abundant proteins using affinity depletion columns. Protein isolation from plant samples is relatively difficult due to the presence of a tough cell wall and the other non-proteinaceus compounds (polyphenols, lipids, and pigments, *etc.*), which interfere with the sample preparation procedure. In general, the sample is first ground and precipitated with 10% trichloroacetic acid in acetone and dissolved in Urea-CHAPS buffer. Sometimes chemical homogenization, followed by boiling using urea, thiourea, CHAPS, 2-mercaptoethanol and polyvinylopolypyrrolidone, followed by hexane precipitation, is done to remove lipids.[22] The bacterial peptidoglycan cell wall is usually disrupted by mechanical disruption, sonication, osmotic shocks, and detergent, pressure and freeze thaw cycles. The viral membrane, on the other hand, is disrupted using the detergents. The protein samples extracted are further processed to remove salts, detergents, lipids, nucleic acids, polysaccharides, *etc.*, and enriched using precipitation, centrifugation, electrophoresis, chromatographic techniques, pre-fractionation using isoelectric focusing (IEF) or solid phase protein enrichments techniques.

The proteins, isolated from various sources, are either separated using GE (1D or 2D), digested and subjected to mass spectrometric analysis, or are digested to peptides, which are then separated using various methods and then subjected to mass spectrometric analysis. In a proteomics workflow, the digestion of proteins to peptides is usually carried out using one or more proteases. A good proteolytic enzyme should have high specificity, good digestion efficiency (low missed cleavages), should be able to generate peptides of optimal length for MS analysis and should withstand hard buffer conditions provided by the reducing agents. The most commonly used protease is trypsin as it is has well-defined specificity and works at an optimum temperature of 37 °C. Trypsin cleaves the polypeptide chain at the carboxyl end of lysine (Lys) and arginine (Arg) residues. Both in-gel (protein extracted from the 2D or 1D gels) and in-solution digestion of proteins is usually done after reduction and alkylation to reduce the cysteine bonds and linearize the

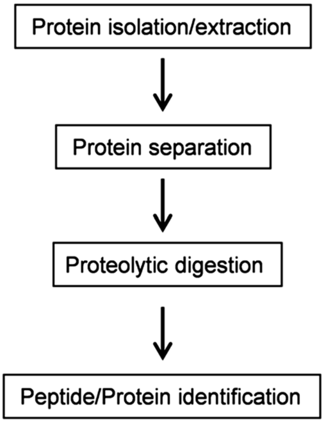

Figure 4.2 A flowchart representing the basic steps involved in a proteomics experiment.

protein before digestion. Reduction is carried out using beta-mercaptoethanol or dithiothreitol. This is followed by alkylation using iodoacetic acid or iodoacteamide to prevent disulfide bond formation. Although trypsin is the protease of choice it is unable to withstand harsh buffer conditions (like 8M urea), which are often used to study membrane proteomes and hence these are digested using LysC and LysN. To obtain a better coverage of the proteins, a combination of two or more proteases (such as chymotrypsin, LysC, LysN, AspN, GluC and ArgC) are increasingly being used.[23] These peptides are then separated using various chromatographic techniques to minimize the complexity of the samples. Figure 4.2 shows a schematic representation of the sample preparation.

4.3 Approaches for Protein Separation

A typical proteomics study aims to identify a large number of proteins in a given sample, which necessitates separation of these proteins or peptides. Various strategies (gel based or gel free) are used for this purpose depending on the objective of the study.

4.3.1 Gel-based Proteomics Approaches

In gel-based proteomics studies, either 1-DGE or 2-DGE techniques are used to separate the proteins.

4.3.1.1 1-DGE

Separation of proteins using 1D sodium dodecyl sulfate polyacrylamide GE (1D-SDS-PAGE) is based on their molecular weight. Following the separation of proteins using SDS-PAGE, the gel pieces are excised into a number of pieces

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

depending on the complexity of the sample. These excised gel pieces are then subjected to "in gel" digestion where the proteins are digested within the gel using proteolytic enzymes. The peptides extracted from the gel pieces can then be identified using MS. The 1-DGE-LC-MS/MS approaches have been shown to provide improved proteome coverage.[24–26] However, the incomplete digestion and extraction of peptides involved in this method make it unsuitable for quantitative proteomics owing to non-reproducibility.

4.3.1.2 2-DGE

This form of GE involves the separation of proteins in two dimensions. In the first dimension proteins are separated based on their isoelectric point (pI). A gel with a pH gradient is prepared and an electric potential is applied across this gel. Proteins in an applied electric field migrate towards the positive or the negative end, depending on their net charge, and stop at their pI (where the net charge is zero). This is therefore also known as IEF. The IEF gel with a pH gradient can be prepared either with the help of ampholytes or immobilized pH gradient gel (IPG) strips. The ampholytes are synthetic amphoteric molecules that exist as zwitterions at a particular pH representing a particular pI for a protein. When the electric field is applied to diffuse the mixture of ampholytes, they move according to their acidic and basic nature towards the anode or cathode, respectively. Since proteins are large molecules they move slowly and stabilize according to their pI. Originally, ampholytes with a pH gradient were made in soft tube gels, which lacked stability. Furthermore, a major disadvantage of using ampholytes is that pH gradients beyond 7-8 cannot be maintained, due to cathodic drift.[27–29] This drift results in the migration of ampholytes due to electro-osmotic flow of solvent towards the cathode wavering the pH gradient. This results in a loss of basic proteins. One of the major improvements in 2-DGE was the introduction of IPG strips.[27,30] In IPG strips, the acrylamide matrix itself has covalently bound weak acid and base buffering compounds, which maintain the pH gradient over wide ranges.[31–33] Further, the availability of pre-cast IPG strips has improved the reproducibility. These precast IPG strips are available both in narrow and broad pH ranges and, depending on the objective of the study, a particular pH range can be selected. In the second dimension, the proteins are separated on the basis of their mass using SDS-PAGE. The IPG strips containing proteins are separated on the basis of pI in the first dimension, and are equilibrated with SDS-PAGE gels by placing the IPG strips over the SDS-PAGE gels. The protein spots can then be visualized using Coomassie Brilliant Blue (detection limit ~10 ng),[34,35] SYPRO Ruby (detection limit ~1 ng)[36,37] or silver stain (detection limit <1 ng).[38,39] Separation based on IEF and SDS-PAGE results in a uniform distribution of protein spots across a 2D gel. Currently, 2-DGE can routinely identify ~2500 protein spots.[40–42] The number of proteins that can be identified is limited by the size of the gel and the complexity of the proteome since several proteins having similar pI and mass may not be completely resolved. The resolution can be improved by increasing the length of the IPG strips to more than 30 cm and large gels in the second dimension, which may help in the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

detection and analysis of 10 000 spots.[41,43,44] To improve proteome coverage, the concept of using multiple IPG strips in a narrow pH range and stitching of the separate images into one using computational tools was introduced[27,45,46] Resolution can be further improved by pre-fractionation of the complex protein sample using various chromatographic techniques or sucrose gradient centrifugation followed by subjecting each fraction to 2-DGE. Identification of proteins from 2D gels is accomplished by cutting the protein spots, digesting them with proteolytic enzymes, and analyzing them in an MS. The entire process of spot picking and downstream analysis is now automated. One of the major advantages of 2-DGE is its ability to identify protein isoforms. However, 2-DGE has limitations in terms of sensitivity of detection, resolution, masking effect by high abundant proteins and error in quantification, mainly due to gel to gel variations. Further, proteins in the extreme pI range (<4 or >9) and molecular weight <15 and >200 kDa are not well resolved. Thus, for quantitative analysis using 2-DGE, multiplexing techniques with fluorescent probes are generally used to circumvent the above-mentioned disadvantages. This technique is discussed in Section 4.8.1.

4.3.1.3 *Electrophoretic Separation of Native Proteins*

This method was developed to study protein complexes, in particular respiratory chain complexes.[47] In this method, native protein complexes are separated in non-denaturing conditions followed by separation under denaturing conditions. For separation under the non-denaturing conditions Coomassie Brilliant Blue G-250 is added to the electrophoresis buffer. The dye sticks to all the proteins and acts as a charge-shifting agent by conferring uniform charge to all the proteins. Thus, proteins are separated in native condition, without unfolding, based on molecular weight and also partially influenced by shape. The proteins are resolved in denaturing conditions of SDS-PAGE by placing the native gel lane perpendicular to the initial axis of separation. This technique is also known as Blue Native-PAGE (BN-PAGE). The technique has been widely used to explore the protein complexes of mitochondria, plastids and chloroplasts for plant proteomics studies[48,49] BN-PAGE can separate protein complexes in the range 20–1300 kDa.[50] However, small molecular protein (<100 kDa) complexes can co-migrate and thus results in false annotation. Further, the dye causes insolubilization of protein complexes resulting in a trailing effect, which makes the technique difficult to optimize.[48] This technique of BN-PAGE was modified by separating the protein in one and two dimensions using native gel, and a third dimension of SDS-PAGE, and hence is also known as 3D electrophoresis (3DE).[51]

4.3.2 Gel-free Proteomics Approaches

The use of GE in the field of proteomics is declining with the advent of high-end MSs, which are coupled to various LC systems. One of the important steps in gel-free chromatography is the separation of proteins using suitable LC techniques. LC is a widely accepted technique for separation of both peptides and proteins. Differences in various properties such as size, mass, hydrophobicity, charge and affinity are exploited for the separation of peptides or proteins. Based on the objective, LC can be classified as: affinity chromatography; size exclusion chromatography; reverse phase LC (RPLC); hydrophobic interaction LC (HILIC); ion exchange chromatography.

4.3.2.1 *Affinity Chromatography*

Affinity chromatography separates proteins or peptides based on their specific interactions with ligands attached to the solid phase, which have high specificity for binding certain proteins or peptides. This method thus helps either in the enrichment or depletion of certain classes of proteins and peptides before analysis. When a solution containing a complex mixture of peptides or proteins is passed over a column containing the stationary phase (with a bound ligand), the non-binding proteins or peptide will pass through the column, while those that interact remain bound with the stationary phase. Altering the buffer conditions where the binding interaction no longer takes place then elutes these bound proteins. This technique is used for the isolation of specific proteins using antibodies, to isolate fusion proteins having affinity tags such as glutathione-*S*-transferase affinity tagged proteins using glutathione coated beads or to isolate specific structural or functional classes of proteins such as phosphoproteins or negatively charged proteins using immobilized metal affinity chromatography using metals such as an Fe^{3+} or Ga^{3+}, *etc.* coated solid phase matrix.[52–54]

4.3.2.2 *Gel Filtration Chromatography*

Gel filtration chromatography is a separation technique based on the size of the protein. The solid phase is composed of inert porous beads.[55,56] The smaller the size of the protein the higher its chance of entering the bead pore, thus allowing it to spend more time stationary before eluting through the column, while larger proteins elute through the column without entering the beads. Thus, proteins elute off the column based on their size; the larger the protein the faster its elution. In gel filtration, the length and diameter of the column is directly proportional to the resolution. An important aspect of gel filtration is column packing, as improper packing could decrease the resolution.

4.3.2.3 *Reverse Phase Liquid Chromatography (RPLC)*

RPLC involves the use of a hydrophobic stationary phase and a polar mobile phase. The hydrophobic molecules in the polar mobile phase tend to adsorb to the stationary phase while the hydrophilic molecules pass through the column.[57,58] The strength of adsorption depends on the hydrophobicity of the molecule. The adsorbed hydrophobic molecules are then eluted by decreasing the polarity of the mobile phase by using organic solvents. The most

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

popular matrix used as a stationary phase in RPLC is porous silica beads attached to alkyl chains of varying lengths (C2, C4, C5, C8, and C18). Protein or peptides in aqueous solution are loaded onto a reverse-phase column, and the hydrophobic patches of the peptides interact with the stationary phase. Polar aprotic organic solvents such as acetonitrile, methanol and isopropanol are used for the elution of bound peptides. The lower the polarity of the solvent the greater the elution potential. Isopropanol has the strongest eluting power but has high viscosity, which results in high back pressure and low column efficiency. Acetonitrile and methanol are the most commonly used organic solvents and are less viscous than isopropanol. Acetonitrile is used most commonly as it has satisfactory hydrophobic properties and does not consociate very strongly with water.[59] Thus it helps in the maintenance of binary equilibrium between the stationary and mobile phase. Increasing the percentage of acetonitrile results in higher hydrophobicity of the mobile phase and hydrophobic proteins bound to the stationary phase are eluted in a manner of increasing hydrophobicity with increasing acetonitrile concentration. In RPLC the matrix bead composition in terms of alkyl chain length[60,61] and pore size[62,63] are important parameters. For instance, the separation of peptides is usually carried out using C18[64,65] while for separation of intact proteins short alkyl chain columns such as C4 or C8 are generally used.[66,67]

4.3.2.4 Hydrophilic Interaction Liquid Chromatography (HILIC)

HILIC has the same principle as RPLC except that the proteins or peptides move over a hydrophilic stationary phase and are eluted by the increasing polarity of the mobile phase, which leads to elution based on the hydrophobicity of the molecules, where the most hydrophobic molecules are eluted first. The stationary phase is made up of underivatized silica with functional groups like siloxanes, silanols[68] and derivatized silica with weak cations[69] or anion exchangers,[70] zwitterionic HILIC[71] and saccharides.[72]

4.3.2.5 Ion Exchange Chromatography

Ion exchange chromatography separates the molecules based on their charge. It is frequently used in the separation of proteins or peptides. The charged molecules in the solution bind to the ionizable functional group in the stationary phase through ionic interactions, while passing through the column.[73] Ion exchange chromatography is classified depending on the types of ions to be separated. In cation exchange chromatography the stationary phase consists of negatively charged functional groups and hence binds to cations, while in anion exchange chromatography, the stationary phase consists of positively charged functional groups enabling the separation of anions. The bound cations or anions are then eluted from the column either by using a higher concentration of the ions or by varying the pH. Commonly used anionic resins are diethylaminoethyl, quaternary aminoethyl, and quaternary ammonium, while commonly used cation resins are functional

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

carboxymethyl, sulfopropyl, and methylsulfonate. Cation exchange chromatography is extensively used in proteomics studies for protein or peptide separation based on ion strength.

4.3.2.6 In-solution IEF: Offgel Fractionation

Offgel electrophoresis is a modified form of the IEF technique[74] where proteins/peptides are separated according to their isoelectric points and the separated proteins/peptides are recovered in liquid phase.[75,76] The offgel electrophoresis device chamber is constructed in such a manner that the tray with channels is placed between the two electrodes. Dried IPG strips are placed onto this tray and a "well frame" consisting of a multiple column is placed over it. The IPG strips are then rehydrated and in each of the wells a diluted protein sample is added and the well sealed with a cover seal to prevent evaporation during electrophoresis. High voltages are applied to the electrodes at both ends of the tray such that the electric field is perpendicular to the direction of the pH gradient and flow of proteins. The IPG gel buffers the thin layer of protein solution allowing convection-free diffusion across the wells due to which proteins with pI close to the pH of the gel remain in contact with the IPG strip in a particular well of the array. At this pH proteins are neutral and therefore do not move further under the effect of the electric field. This has various advantages over the conventional 2-DGE since the protein/peptides remain in liquid throughout and can be directly analyzed by an MS, eliminating the necessity of cutting spots from the gel.[75] This technique is highly compatible with other downstream proteomics techniques such as LC-MS, gel-based identification and separation, immuno-depletion,[77–79] *etc.*

4.4 Multidimensional Protein Identification Technology (MudPIT)

MudPIT is a powerful tool that eliminates the necessity of separating proteins using gel-based methods. This is particularly useful for complex biological samples with high dynamic range. Large numbers of peptides obtained from proteolysis of complex biological samples are directly subjected to strong cation exchange (SCX) chromatography followed by RPLC before direct injection into an MS for identification. This is also known as 2D LC (2D-LC). After SCX chromatography the samples can be transferred individually to RPLC coupled to ESI MS.[80] This is known as discontinuous multidimensional chromatography. However, technologies have emerged where fractions eluted from the first dimension (size exclusion chromatography, ion exchange chromatography) are injected online to an RPLC MS coupled system by column-switching valves that switch the column in line.[81–84] This is known as continuous multidimensional chromatography. In addition, multidimensional separation is also achieved using biphasic columns whose proximal and distal parts are filled with different types of matrix

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

corresponding to two different chromatography techniques such as reverse phase and cation exchange.[85] The distal part is mostly RPLC. The proximal column allows stepped elution and the distal part allows gradient elution using a solvent system compatible with both arrangements. The method of continuous multidimensional chromatography with a biphasic column was modified and developed by Yates and colleagues.[86,87] Desirable results can also be achieved without pre-fractionation by using long columns with small bead sizes and longer gradients at particular temperatures. The yeast proteome with high coverage was performed using 35 cm long, 1.8 µm bead size, C18 columns with 4 h gradient.[88,89] More recently, similar coverage has been reported using 1.7 µm beads with 1.3 h long gradients.[90]

4.5 MS for Proteomics

A MS measures the mass-to-charge ratio (m/z) of ions in vacuum. It consists of an ionizer, mass analyzer and a detector. A combination of different types of ionizers, analyzers and detectors results in varied types of MSs, each serving a different purpose. MS instrumentation is described elsewhere in this book.

4.5.1 Ionization

Samples infused into an MS are first subjected to ionization *via* an ionization source. These ions are then accelerated in an electric field and forced into mass analyzers where they are separated based on their m/z. Ionization techniques have evolved over the years from electron ionization to electrospray ionization. Since proteomics studies involve whole proteins or peptides, which are sensitive to temperature, it is important to have efficient soft ionization techniques that easily desorb into the gaseous phase to minimize sample degradation.[91,92] The two most widely used ionization techniques used in proteomics are matrix-assisted laser desorption/ionization (MALDI)[93,94] and electrospray ionization (ESI).[91]

4.5.1.1 MALDI

MALDI is a solid phase ionization technique where the sample is mixed with an excess matrix, which has a high capacity to absorb laser radiation, in a metal plate and allowed to dry. The most common type of matrices used in proteomics include 2,5-dihydroxybenzoic acid (DHB),[95] 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid),[96,97] 4-hydroxy-3-methoxycinnamic acid,[96,97] and CHCA.[98] These are prepared by mixing them in a mixture of organic solvent, usually acetonitrile, and pure water in 1 : 1 ratio. Trifluoroacetic acid is added to this mixture to help generate good $(M + H)^+$ ions. The incident energy from the laser is absorbed by the matrix and is transferred to the protein/peptide sample, which desorbs and becomes ionized.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

The gaseous matrix accelerates the ionization of analytes as mostly singly charged $(M + H)^+$ ions. The ionization is achieved by numerous laser shots, usually in the hundreds, to provide adequate signal-to-noise ratios.[99] The charged molecules then reach the mass analyzers. The mass analyzers commonly coupled with MALDI are time of flight (TOF), separating ions using the mass-to-charge ratio. MALDI has a range of up to 300 000 Da, which makes it suitable for the identification of proteins with high molecular weight. A major limitation of this method is that it offers low shot-to-shot reproducibility and resolution, which is entirely dependent on the sample preparation and homogeneity of the matrix.[100,101]

4.5.1.2 ESI

In ESI,[91] the liquid solution of peptides and biological molecules eluted from the liquid chromatographic system is directly converted into ionized molecules in gaseous form. The fine spray of liquid is generated *via* passing the solution through fine metal tips held at a high voltage of ~2–6 kV. In the presence of a high voltage and strong electric field, a spray of highly charged droplets is produced. These droplets are electrostatically attracted towards the MS orifice. Applying temperature and a sheath gas at the metallic tip further disperses these droplets and evaporates the solvent, facilitating the entry of ions into the MS. With de-solvation, the size of droplets decreases, resulting in increased electric field density at the surface. As the size of the charged droplet decreases, the charge density on its surface increases. The mutual Coulombic repulsion between like charges on this surface becomes so great that it exceeds the forces of surface tension, and ions are ejected from the droplet through a "Taylor cone".[102] The charged ions leaving the droplets enter the orifice of the MS towards the mass analyzer. ESI came as a boon for the study of biological samples as nano-flow LC systems can be attached before the MS[103,104] along with the convention micro-flow LC. The nano LC-ESI MS has provided an advantage for protein analysis in terms of a flow rate in nanoliters, which requires low protein concentration and continuous online analysis of peptides separated on the capillary reverse phase (RP) columns.[104,105]

The ionized samples accelerate to the mass analyzers and are separated on the basis of their *m*/*z* values.

4.5.2 Mass Analyzers

The mass analyzer plays an important role after ionization. The ions accelerated in the MS are separated inside the analyzers on the basis of their *m*/*z* values. The actual potential of different types of MS is realized on the basis of the type of analyzer used inside the MS. The mass analyzer mainly regulates parameters such as mass range, resolution and accuracy, scan rate, detection limit, and sensitivity of the instrument. The selection of instrument for a particular application comprises major adjustment of these parameters.

The different types of analyzers that are used in proteomics include ion trap, TOF, quadrupole, Orbitrap, ion mobility and Fourier transform ion cyclotron resonance (FTICR) based analyzers. The details of analyzers and detection technologies are described elsewhere in this book.

4.6 Tandem MS (MS/MS)

MS/MS is important as it allows the fragmentation of peptides or proteins, which provides information on the sequence of the peptides. Fragmentation achieved by collision is known as collision induced dissociation (CID). Using CID, an ion with the desired *m*/*z* is filtered by the mass analyzer and trapped in collision cell supplied with collision gas, mainly argon (Ar). Collision of Ar with ions results in fragmentation, producing daughter ions constituting MS2, MS3 up to MS*n*. Therefore, it is sometimes referred to as MS*n* where *n* represents the generation of ions being analyzed. The digested protein mixtures are subjected to MS/MS. It consists of an ionization source, mass analyzer, collision cell, second mass analyzer, and detector. The initial analyzer helps in resolving peptides with a particular mass range of *m*/*z* and isolates one specific *m*/*z* ion at a time. These ions are then sent to the collision cell for fragmentation by inert gas. The peptides are usually fragmented on or nearby the amide bond, generating a sequence ladder of fragments defining the structure of the peptide. The second analyzer then determines the mass of the fragmented peptide ions yielding the amino acid sequence information. The large amount of information yielded by the several peptide fragments sequence, called peptide sequence tags, can be searched in the protein databases for the identification of proteins in a peptide mixture.[106,107] The MS/MS approach, as compared to peptide fingerprinting, provides more information about peptide sequences for protein identification than merely measuring the mass of the peptides.

4.7 Approaches in Proteomics

There are various approaches used for the identification and quantification of proteins depending upon the complexity of the sample and the problem to be addressed. While protein identification and characterization is carried out using top-down[108,109] and bottom-up proteomics,[110,111] quantitative analysis is performed using the directed[112] and targeted proteomics approach.[113,114] A representative figure of the different proteomics approaches is shown in Figure 4.3.

4.7.1 Top-down Proteomics Approach

In the top-down proteomics approach the intact protein is introduced into the MS for fragmentation, which helps in the identification of the protein. The top-down proteomics approach has advantages as it helps in the whole

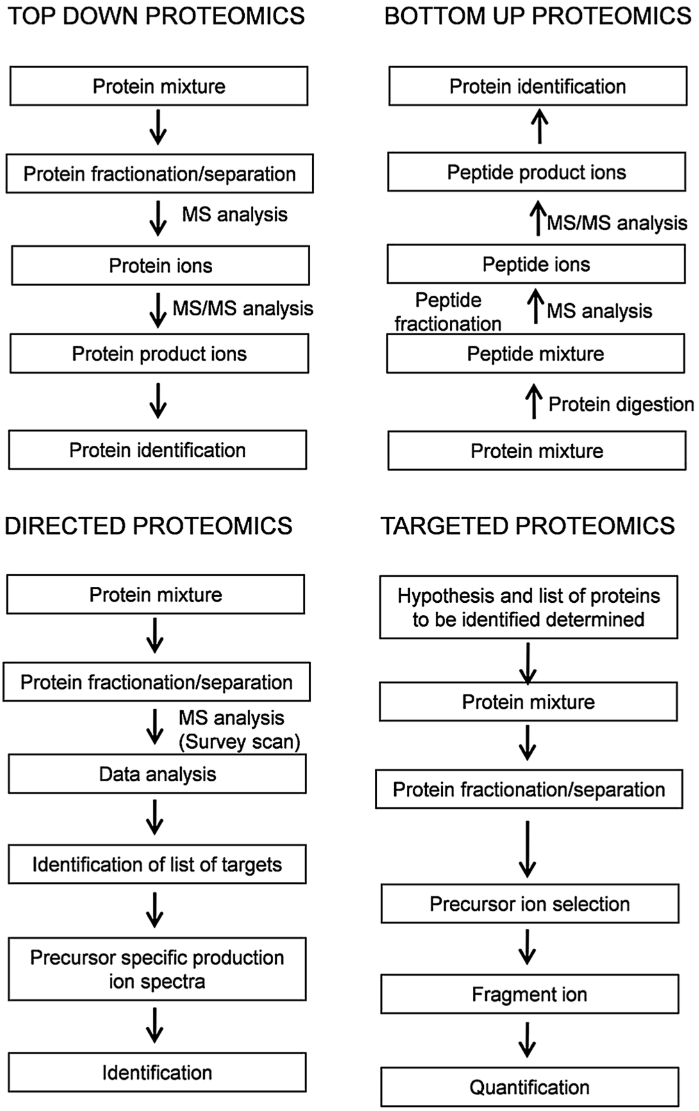

Figure 4.3 A schematic diagram highlighting the different types of proteomics approaches.

proteome-wide characterization of proteins as compared to the other three MS strategies discussed above.[115,116] The limitations of shotgun methods, such as the detection of a fewer number of peptides, which limits the information of post-translational modifications (PTMs), sequence variants and the possibility of the same peptide coming from different proteins or a similar protein, are not problems in top-down proteomics. Separation of proteins from a complex mixture is achieved either offline[117] or online using chromatographic (such as RPLC, HILIC and ion exchange) and electrophoretic techniques (such as tube GE, capillary electrophoresis and IEF).[118] However, offline separation of intact proteins is preferred since it may be necessary to use multiple separation techniques to separate complex protein mixtures. Offline top-down proteomics has also been used to separate certain classes of proteins. HILIC has been used to separate modified histones[119,120] and membrane proteins.[121,122] Ion exchange chromatography has been used in combination with RPLC for intact protein studies in *E. coli*[123] and yeast,[124] proteotyping in human leukocyte[125] and chromatofocusing in breast cancer studies.[126] Tube GE is used in combination with online RPLC to achieve the fractionation of yeast proteome.[127] Gel-eluted liquid fraction entrapment electrophoresis came as the new improvement in 2009 in top-down proteomics.[128,129] Capillary electrophoresis (CE)[130] is a useful method for the separation of intact proteins.[130] This technique utilizes a small capillary (inner diameter <100 μm) and separation is achieved by applying voltage (10–30 kV) thus making the process quicker. The most significant achievement of this technology is the separation of alpha and beta sub-units of hemoglobin isolated from a single erythrocyte,[131,132] ribosomes of *E. coli*[132] and glycoforms of various plasma proteins, including erythropoietin, fetuin, and α 1-acid glycoprotein.[133] In capillary IEF (cIEF)-MS,[134,135] another variant of CE, proteins are focused and introduced into the MS directly to an electrospray inlet, improving the speed of data acquisition.[136,137] cIEF coupled with RPLC provides second dimension separation and also removes ampholytes coming along with proteins.[138,139]

Top-down proteomics is mostly performed using an ESI source coupled with FT-ion cyclotron resonance (ICR)-MS[140] or an Orbitrap instrument.[141,142] The fragmentation is either achieved by electron capture dissociation (ECD)[143] and electron transfer dissociation (ETD).[144,145] In ECD and ETD an electron is captured by an intact protein molecule, which results in the breaking of the backbone of the large molecule and provides more information than CID.[146,147] Top-down proteomics helps in achieving almost one hundred percent sequence coverage. As the full sequence is covered post-translational modifications[148] and proteoforms,[149,150] which are generated due to alternative splicing and genetic variations, can be studied efficiently. As this technique involves the analysis of intact proteins,[151] it is a good technique for the analysis of *de novo* protein characterization,[152] expressed sequence tag proteins analysis[153,154] as well as identification of single proteins or a simple mixture of proteins. Although the top-down proteomics approach is very effective, it has still not gained popularity due to instrument limitations and the absence of the desired online separation techniques. However, with the advancement of new separation technologies and newer instruments, top-down proteomics will be the method of choice in the near future. Quantification in top-down proteomics using label-free[155] and labeled approaches such as stable isotope labeling with amino acids in cell culture (SILAC)[142] are also developing.

4.7.2 Bottom-up Proteomics Approach

Bottom-up proteomics, also known as shotgun proteomics, derived its name from shotgun genomics. In the shotgun approach proteins are enzymatically digested to produce peptide mixtures, which are then fractionated using a suitable chromatographic system and subjected to MS analysis. The resultant peptide masses and the sequences obtained after mass analysis are used for identification and quantification of proteins. The data is acquired as series of MS (survey scan) and dependent MS*n* scans. For this reason it is also known as data dependent acquisition (DDA) or tandem data acquisition. This utilizes collision-activated dissociation (CAD) or CID for fragmentation of peptide precursors selected from the survey scan. Typically, 10–20 precursor ions per survey scan are automatically selected by the MS, for fragmentation based on the signal intensity, using CID. The fragmentation pattern of each selected precursor ion is recorded as fragment ion spectra, also known as product ion scanning. A full cycle consists of a survey scan and a product ion scan of ~100 ms, which is faster than the chromatographic separation of peptides ~30 s.[156,157] The MSs show impressive data acquisition rates by acquiring thousands of fragment ion spectra in a fraction of a second from the peptides eluted from a typical LC-MS/MS system. The resolution and accuracy achieved in the MS scan affects MS/MS ion spectra obtained after fragmentation. The fragment ion spectra of peptides, along with precursor MS ion information, is subsequently used for amino acid sequence analysis of peptides, which in turn helps in identifying the protein from which the peptides originated. The most widely used shotgun tandem mass acquisition method involves the alignment of experimentally generated MS*n* fragment data of peptides to an *in silico* generated peptide fragmentation library of a given database.[158,159] A wide varieties of instruments can be used for bottom-up proteomics, which include Orbitrap, Q-TOF, MALDI-TOF,[160] *etc.* With the development of high-accuracy and high-resolution instruments the accuracy of peptide identification has increased dramatically. One of the limitations of shotgun proteomics is reproducibility, which is compromised as the number of peptides generated after each tryptic digestion varies greatly. This could be circumvented by extensive fractionation and repeated analysis of the same sample.[161] This increases the coverage and reproducibility but is not cost effective. Further, in most cases bottom-up proteomics results in limited protein coverage[162] and could lead to the loss of labile post translational modifications. The bottom-up method is mostly used for qualitative analysis and is a method of choice for the discovery phase experimentation where no prior knowledge is required. However, of late it is also increasingly being used for differential protein analysis[163,164] (relative quantification) using both labeled and label-free approaches, which are discussed in Section 4.8.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

4.7.3 Directed Proteomics Approach

Directed proteomics, as the name suggests, fragments a defined set of peptides obtained from a survey scan.[165–167] This mass and time tagging concept for directed analysis was introduced in 2000.[168]In this two-step approach, a desired list of peptides, called the inclusion list of precursor molecules, is created from an LC-MS survey scan in the first step, containing information on their *m*/*z* value, retention time and charge state. In the second step, the same sample is infused for a repeat of the survey scan followed by a product ion scan in a data-dependent manner based on the inclusion list. To trigger data-dependent CID of precursor ions, a peak intensity above a certain signal-to-noise threshold and its presence in the survey scan is mandatory. This method allows precise quantification as the accurate mass of the precursors is taken into account. The directed approach is used for discovery proteomics in a hypothesis driven manner of mostly low abundant protein species, which are missed as noise due to low signal intensities in conventional LC-MS experiments.[112]

The full coverage of the proteome with a large number of peptide identifications can be achieved by repeated analysis of the sample with a new inclusion list every time to avoid repetition of the same peptide in the subsequent LC-MS/MS runs until no new features are left.[165] A reverse strategy of the exclusion list is also employed in directed proteomics, which is also known as accurate mass-exclusion-based data-dependent acquisition (AMEx) or standard data-dependent acquisition.[169] In AMEx the previously studied precursor ions are excluded in the subsequent survey and product ion scans which improves new protein identification.[170] In directed proteomics the sequence of the precursor ions may or may not be known. Directed MS with known sequence assignment is used for peptide screening (of the whole proteome, a subset of the protein or their modifications), single reaction monitoring (SRM)[167] assays and absolute quantification using SILAC.[171] This is mostly based on the determination of proteotypic peptides (PTPs) which have a unique amino acid sequence and are flyable[172] in MS and determination of their LC elution time. Knowing the PTPs considerably reduces the repetition of data by reducing the number of scans required to cover the whole proteome set. PTP libraries can be created from previous extensively acquired LC-MS/MS experiments on platforms such as PeptideAtlas, PRIDE, PROWL[173–176] or by computationally predicting PTPs.[177,178] These PTPs are subjected to directed MS/MS analysis in subsequent runs for deep sequencing as well as for SRM assays where peptides for the biomarkers are defined. In absolute quantification the heavy labeled PTPs of the peptides[179] or the labeled intact protein standards before digestion are spiked in the biological samples.[180] This helps in the absolute quantification of endogenous peptides present in a sample based on the signal intensity and modifications can also be studied.[181,182] This advanced form of data-dependent acquisition holds

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

true potential in terms of a proteome-wide study and reproducibility, but is also very expensive.

4.7.4 Targeted Proteomics Approach

The targeted proteomics approach is similar to the directed approach. In this approach, pairs of precursor and product ions, along with their *m*/*z* values, retention time, charge state and collision energy is defined for each pair to be targeted. Unlike the directed approach, in the targeted approach the presence of precursor ions in the survey scan is not mandatory. Thus, this method is extensively used for SRM or multiple reaction monitoring (MRM). The SRM assays are more popular for small molecule studies and are well established.[183] A triple quadrupole instrument is mostly suitable for targeted proteomics due to the very high sensitivity of these instruments. This is absolutely a hypothesis-based approach as precursor targeting and method development needs prior information. High-throughput assays can be designed based on a set of synthetically synthesized peptide libraries for a large number of proteins. In targeted proteomics, the initial survey scan is not performed. Instead, scanning is directly done for a pair of precursor and product ions also known as transitions in the first quadrupole and third quadrupole of QqQ instruments. This method is highly selective for precursor molecules, independent of their presence in the sample and almost 1000 transitions per LC-MS/MS run can be processed without compromising sensitivity.[113] With highly sensitive instruments even a minimal amount can be detected and quantified with high reproducibility. Prior information for an experiment such as specific peptides, elution times and charge states are mostly obtained from well-established databases such as SRM/MRM-Atlas[184–186] or PeptideAtlas,[174] which can be directly fed for analysis to highly integrated software such as Skyline,[187,188] *etc.* Comparing the targeted fragment ion intensities with the isotopically labeled fragments for the same peptide provides quantification. Parallel reaction monitoring (PRM) is considered to be the most acceptable form of SRM as in the case of PRMs a higher number of transitions are considered per protein, leading to higher confidence in identification and quantification.[189] The performance of these approaches is entirely dependent on the dwell time, *i.e.* time per transition. The higher the dwell time the fewer transitions or peptides can be studied per cycle time, limiting the detection abilities.

4.8 Quantitative Proteomics

MS is inherently not a quantitative technique. However, using various methods, researchers have been successful in obtaining relative or absolute quantification of proteins using MS. The following section discusses various quantitative techniques used in proteomics.

4.8.1 2D Difference GE (2D-DIGE)

One of the first methods used for relative quantification of proteins was through 2-DGE. However, inter-gel variations and variable exposures to stains limited its use as a quantitative method. These limitations were to some extent circumvented by running multiple gels for each condition and comparing these gels. However, this was time consuming and hence did not gain popularity. One of the major improvements was the introduction of fluorescence dyes, which could be tagged to proteins.[190,191] This has not only increased the sensitivity but also eliminated the need for running multiple gels to compare the spot intensities between two conditions. This technique, known as difference in GE (DIGE), utilizes fluorescence tagging of two protein samples with two different dyes. The tagged proteins are run on the same 2D gel and post-run fluorescence imaging of the gel is used to create two images, which are superimposed to identify spots with different intensities. The dyes used for the purpose are cyanine dyes (Cy), which are synthetic dyes belonging to the polymethine group.[192] They are fluorescent dyes and depending on the structure they cover the spectrum from IR to UV, as shown in Table 4.1. The dyes are designed to ensure that proteins common to both samples have the same relative mobility (since the dyes are of similar charge and molecular weight) regardless of the dye used to tag them.[193,194]

These dyes are very sensitive (about 25 pg of a single protein can be detected) and most importantly have a linear response to protein concentrations up to five orders of magnitude. Two types of Cy dye are used for DIGE: minimal labeling dyes[192–196] and saturation labeling dyes.[197] In minimal labeling dyes, the cyanine derivatives contain a *N*-hydroxysuccinimide group and bind to the ε-amino group of protein Lys residues. These are normally used in DIGE experiments and labeling of 50 μg protein is possible. To prevent precipitation the minimal labeling technique[198] is chosen where only 1–2% of the proteins are labeled such that only a single Lys per protein is labeled. The saturation dyes contain a maleimide group, which binds to cysteine residues of the protein.[194] These are used when samples are precious or available in small amounts and labeling of 5 μg protein is possible. These dyes are hydrophobic, which can result in precipitation.

For quantification purposes, two samples are tagged with two different dyes (usually Cy3 and Cy5) and equal mixtures of the two samples are tagged with a third dye (usually Cy2), which acts as an internal control (containing equal mixtures of the two samples).[199] All these three tagged protein samples are then mixed together in equal proportions and subjected to 2-DGE.

Table 4.1 Fluorophores used for 2D-DIGE.

Flurophore	Excitation peak (nm)	Emission peak (nm)	Mol. wt.
Cyanine, Cy2	492	510	714
Indocarbocyanine, Cy3	550	570	766
Indodicarbocyanine, Cy5	650	670	792

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

Post electrophoresis, three scans corresponding to Cy2, Cy3 and Cy5 are taken for each gel. The scanned images of each sample and the internal standard are then analyzed using software, which overlaps the spots present in each scan, identifying the position of each spot in the gel. Use of internal standards eliminates gel-to-gel variation thereby enabling comparisons between multiple gels.[200,201] The internal standard image has the maximum number of spots and is assigned as the "master". Once the protein spots have been matched, the ratio of protein abundance between samples is determined by comparing the ratio of their intensities. However, since the concentrations of proteins in the gel are very low, excising the spots with varying intensities for identification of the protein by MS is not possible. Thus, a "preparatory" gel is run using 500 µg to 1 mg of a 1 : 1 mixture of the initial protein samples. This gel is then stained with Coomassie or silver stain and the scanned image of the gel is overlaid with that of the fluorescent gel for matching of spots between the two gels. Once the spots are matched, those with significantly different intensities observed in the analytical master gel are excised and subjected to MS for identification. One of the major advantages of this technique is its ability to identify protein isoforms and it is still a method of choice where the objective is to compare the intensities of various isoforms of a protein. Although this method eliminates the need for comparing multiple gels for relative quantification, it still is limited by the number and type of proteins that can be detected, especially when comparing complex proteomes, and is also time consuming when compared with gel-free methods. A schematic diagram of DIGE is shown in Figure 4.4.

4.8.2 Labeled Quantification or Stable Isotope Labeling

With the development of MSs with high accuracy capable of unit mass resolution across the relevant mass range, various methods of relative quantification where proteins or peptides tagged with isotopically labeled forms have emerged. Relative quantification of proteins using stable isotope labeling is done based on MS signal intensity and is either performed *in vivo* or *in vitro* depending upon the objective of the study.

4.8.2.1 In vivo *Methods of Metabolic Labeling*

This method is used primarily to compare relative expressions of proteins in cell cultures. Cells are grown either in media with heavy or light isotopes for the desired time period. The label becomes attached to the proteins during cell growth, division and metabolism. This method of metabolic labeling is precise and eliminates the chances of any human error. The technique was first used for studying differential phosphorylation.[202] The cells to be studied were grown separately in ^{14}N (light) and ^{15}N (heavy) labeled media incorporating the two isotopes in the proteins.[203] The proteins from two different conditions were pooled and subjected to mass spectrometric analysis. The light and heavy isotope labeled peptides show mass shift, which is clearly resolved in the MS at the MS level. The ratio of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

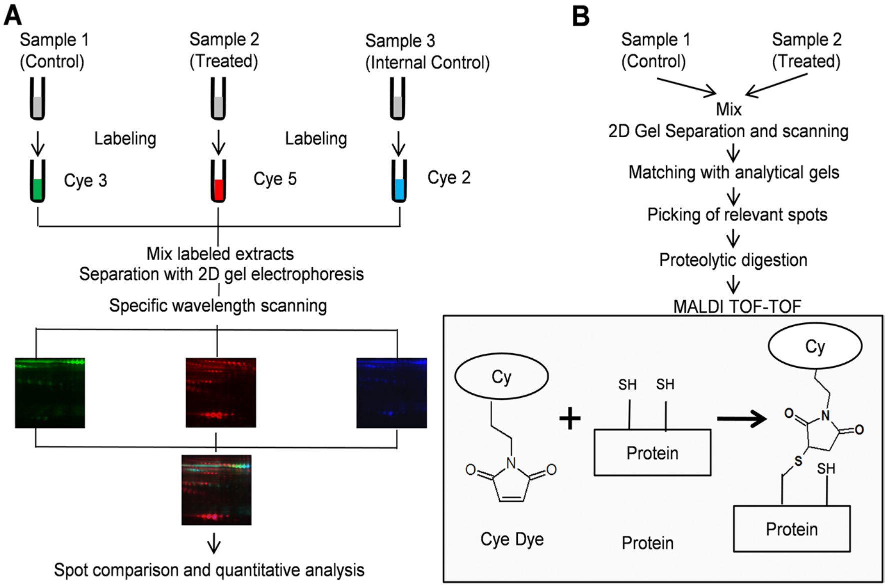
A
Sample 1 (Control)
Sample 2 (Treated)
Sample 3 (Internal Control)
Labeling
Labeling
Cye 3
Cye 5
Cye 2
Mix labeled extracts
Separation with 2D gel electrophoresis
Specific wavelength scanning
Spot comparison and quantitative analysis
B
Sample 1 (Control)
Sample 2 (Treated)
Mix
2D Gel Separation and scanning
Matching with analytical gels
Picking of relevant spots
Proteolytic digestion
MALDI TOF-TOF
Cy
O
N
O
Cye Dye
+
SH SH
Protein
Protein
Cy
O
N
O
S
SH
Protein

Figure 4.4 A schematic diagram of 2D-DIGE. (A) An experimental flowchart for Cy dye labeling. (B) A flowchart for preparative gel for spot picking and identification of the relevant spots obtained from the 2D-DIGE labeling experiment. The inset consists of the saturated labeling reaction where the maleimide dye labels the cysteine residues. Unlike saturated labeling, in minimal labeling the NHS ester reactive dye binds with the amine group of Lys.

the intensities of light and heavy labeled peptide pairs for a protein is an indicator of the relative expression of the protein under the two conditions. However, in this method, as all the nitrogen atoms in a peptide are labeled, quantification of an unknown protein is difficult since there is no *a priori* knowledge of the initial peptide mass or the mass shift expected due to the heavy label.

Using culture media containing a specific labeled amino acid circumvented this problem. This method is known as stable isotope labeling with amino acids in cell culture (SILAC). Since only a specific amino acid is labeled in this method, the intracellular proteins will contain either light or heavy isotopes of the specific amino acid. Labeled Lys and Arg are usually used for this purpose. The heavy isotopes of Lys and Arg may contain heavy isotopes of ^{2}H, ^{13}C, or ^{15}N as compared to light isotopes of H, ^{12}C, and ^{14}N. A combination of these isotopes can thus be used for multiplexing: $^{12}C_6{}^{14}N_4$–Arg, $^{13}C_6{}^{14}N_4$–Arg and $^{13}C_6{}^{15}N_4$–Arg introducing 0 DA, 6 Da and 10 Da change in mass, respectively.[204] A combination of Lys and Arg labeling can also be used such as Lys0–Arg0, Lys4–Arg6, and Lys8–Arg10.[205] The proteins are quantitated on the basis of the ratios of the intensities of heavy and light isotopes.[206] A major disadvantage of this technique is its limited utility for only cells that can be grown in labeled media and cannot be used for tissues and body fluids. Further, the

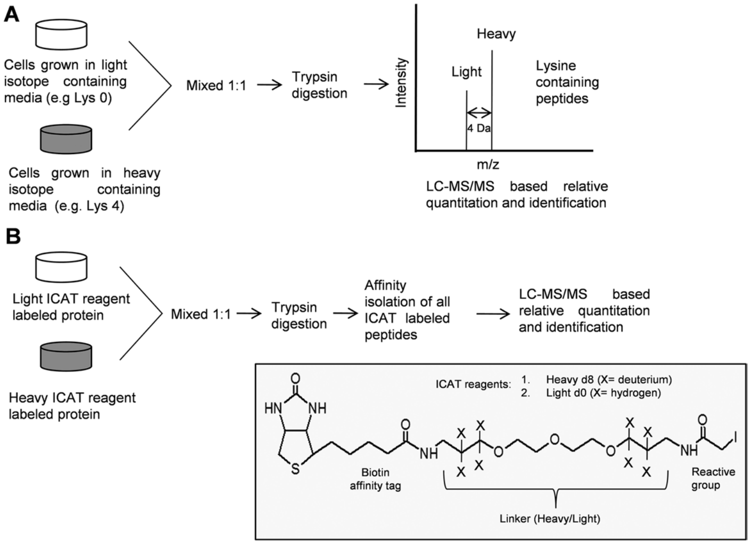
A
Cells grown in light isotope containing media (e.g Lys 0)
Cells grown in heavy isotope containing media (e.g. Lys 4)
Mixed 1:1
Trypsin digestion
Intensity
Heavy
Light
4 Da
Lysine containing peptides
m/z
LC-MS/MS based relative quantitation and identification
B
Light ICAT reagent labeled protein
Heavy ICAT reagent labeled protein
Mixed 1:1
Trypsin digestion
Affinity isolation of all ICAT labeled peptides
LC-MS/MS based relative quantitation and identification
ICAT reagents: 1. Heavy d8 (X= deuterium) 2. Light d0 (X= hydrogen)
Biotin affinity tag
Linker (Heavy/Light)
Reactive group

Figure 4.5 (A) An experimental flowchart for SILAC. (B) An experimental flowchart for ICAT.

conversion of Arg to proline can happen during cell division, which complicates the analysis. A schematic diagram of SILAC is shown in Figure 4.5.

4.8.2.2 In vitro *Methods: Enzymatic and Chemical Labeling*

4.8.2.2.1 Enzymatic Labeling. In this method, the C-terminal carboxyl groups of peptides are labeled with ^{18}O in the presence of trypsin or another protease such as Lys-C, Glu-C, *etc.* and ^{18}O labeled water ($H_2{}^{18}O$).[207] Protein digestion occurs in the presence of ^{18}O labeled water ($H_2{}^{18}O$) introducing a mass shift of 2–4 Da depending on partial or complete labeling of the carboxyl group, which have two oxygen moieties.[208] The heavy (^{18}O) and the light (^{16}O) isotopic forms of the peptide help in comparison- and ratio-based quantification at the MS1 level.[209–211] This is a very simple and cost effective method of labeling and quantification but is not preferred widely because of certain disadvantages. The data analysis is challenging due to incomplete labeling and back exchange of oxygen,[212,213] which depend on the pH, enzyme amount and activity, digestion time, purity of the ^{18}O labeled water and dilution of the ^{18}O label after sample pooling. However, these effects can be minimized by processing the samples separately, the use of immobilized trypsin, or pre-trypsinization followed by incubation in ^{18}O labeled water at low pH in the presence of immobilized trypsin[210,214,215] and the use of computational algorithms that can apply correction factors.[216]

4.8.2.2.2 Chemical Labeling. Chemical labeling is the most widely used method for relative quantification. In this peptides are tagged with reagents that bind to specific functional groups (like –SH or –NH_2) of amino acid

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

residues of a peptide. The relative quantification is achieved either at the MS1 or at MS2 levels depending on the methodology used. In isotope-coded affinity tagging (ICAT), mass-coded abundance tagging (MCAT), isotope-coded protein labeling (ICPL), isotope-differentiated binding energy shift tagging (IDBEST), and dimethyl labeling the ratio of the peak intensities at the MS1 level is used for relative quantification while identification of the protein is achieved at the MS2 level. In contrast, in isobaric labeling, which includes isobaric tags for relative and absolute quantitation (iTRAQ), tandem mass tagging (TMT) and isobaric peptide termini labeling (IPTL), both relative quantification and identification is done at the MS2 level.

In ICAT labeling, the sulfhydryl group of cysteine residues of the peptides are labeled with a reagent containing a chain of eight hydrogen molecules in non-deuterated (^{1}H) or deuterated (^{2}H) form.[163] This linker of hydrogen molecules also has a biotin tag for affinity purification of labeled peptides. The samples are digested, labeled, affinity purified and subjected to MS. The heavy and light peptide from two different samples shows a mass difference of 8 Da at the MS1 level. The cysteine-rich peptides are specifically enriched using this technique, which significantly reduces the complexity of the peptide mixture, but the proteome level coverage for relative quantification is poor since only peptides containing cysteine residues will bind to the reagent. Further, the biotin tags used as a linker complicates the mass spectra analysis as it provide hydrophobicity to peptides, hindering elution and creating a difference in retention of light and deuterated peptides due to early elution of heavy isotopes during RPLC.[217] This drawback of retention time has been taken care of by the introduction of an advanced cICAT method where a ^{13}C labeled version of ICAT reagents that generates a mass shift of 9 Da[218] is used. ICAT has been used widely for biomarker discovery.[219,220] Another modification of the ICAT method is the OxICAT (oxidation state determination using ICAT), which helps in the determination of the oxidized form of the cysteine residues such as disulfide and post translational modifications such as cysteine sulfenic acid or *S*-nitrosocysteine.[221] This is known as redox proteomics analysis.[222,223] In OxICAT, all the free cysteine residues in a peptide are first irreversibly labeled with a light isotope. It is then reduced in the presence of a strong thiol reductant such as tris(2-carboxyethyl)phosphine, glutaredoxin and ascorbate, and labeled with a heavy isotope (^{13}C). This enables identification of disulfide bonded cysteines.[224]

In ICPL, MCAT and dimethyl labeling techniques, the free amino groups are labeled. In ICPL, free amino groups of intact proteins are labeled and can be used for all different types of biological samples.[225,226] The labeling is done by the derivatization of the amino group with *N*-hydroxysuccinimide-nicotinic acid ester (Nic-NHS). Using this technique up to four different samples can be analyzed by derivatization with Nic-NHS containing deuterated hydrogen ^{2}H and ^{13}C such as ICPL_0, ICPL_4 (four ^{2}H), ICPL_6 (six ^{13}C) and ICPL_10 (four ^{2}H + six ^{13}C) reagents, which will have a mass shift of 4 Da, 6 Da and 10 Da, respectively.[227] Since the protein labeling is done prior to digestion, proteins retain their physical and chemical properties, which can be utilized for separation technologies. The labeled samples are combined, separated, digested enzymatically and subjected to MS. The ratio of the isotope peaks at the MS level is an indicator of the relative peptide and hence protein abundance in a sample. A variant of this technique uses *N*-acetoxysuccinimide and 1-nicotinoyloxysuccinimide for labeling to expose free N-termini only. However, nicotine derivatives enhance the fragmentation in MS but the labeling of Lys residues reduces the efficiency of trypsin digestion resulting in longer and bigger peptides hindering fragmentation. A combinatorial digestion strategy such as trypsin/Glu-C helps in better digestion of the proteins.[228]

The MCAT technology involves the modification of the ε-amino group at the C-terminal Lys residues of digested proteins using *O*-methylisourea. The control sample is not usually labeled while the test sample is labeled. The two peptide samples are pooled, fractionated and analyzed using an LC/MS system. The y-ions of a labeled peptide are generated with a mass shift of 42 Da (for singly charged). Unlike the methods described above, quantification is done by comparing the intensities of mass shifted y-ions of a peptide at the MS/MS level.[229] Another labeling method is dimethyl labeling,[230] which labels all primary amines (N-terminal amino group and ε-amino groups of Lys) of proteolytically digested peptides by formaldehyde, which results in a Schiff base intermediate reduced by cyanoborohydride. Three proteolytic peptide samples can be analyzed in parallel[231] in a single MS run by using combinations of formaldehyde and cyanoborohydride (C^1H_2O + NaB^1H_3CN) resulting in a mass shift of 28 Da per amino group of a peptide, two deuterated ^2H formaldehyde + cyanoborohydride (C^2H_2O + NaB^1H_3CN) resulting in mass shift of 32 Da per amino group of a peptide and two deuterated ^2H, one ^{13}C formaldehyde + ^2H cyanoborohydride ($^{13}C^2H_2O$ + NaB^2H_3CN) resulting in mass shift of 36 Da per amino group of a peptide.[231] However, similar to ICAT it results in a shift of retention time due to the deuterated label.[217] This is a cost effective, fast method and can be used for biological samples.

The IDBEST method is capable of relative as well as absolute quantification. This method is based on the introduction of a mass defect. A mass defect can be defined as the energy equivalent liberated after the packing of nucleons (protons and neutrons) in a nucleus of an atom. The energy released on packing is known as the nuclear binding energy and according to the theory of relativity this energy is equivalent to mass error. Therefore, by convention, all the elements differ from their theoretical masses with some mass defect. ^{12}C is considered to have a zero mass defect. The commonly occurring atoms such as C, H, N, and O have a negligible mass defect based on the number of nucleons. However, atoms of ^{35}Br (Bromine) and ^{63}Eu (Europium) result in a mass defect of −0.1 amu as compared to ^{12}C, which can be easily captured by high-end MSs. The IDBEST labeling uses the well-established top-down proteomics approach of inverted mass ladder sequencing patented by Hall and Schneider.[232] In this technique Cys-reactive IDBEST™ tags (IGBP) or the amino-reactive group of *N*-hydroxysuccinimidyl ester with ^{12}C or ^{13}C is used for light and heavy isotope labeling. For relative quantification the ratio of intensities of isotopic peaks obtained at the MS1 level are considered.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

Absolute quantification is carried out by determining the number or concentration of bromine coming from each tagged peptide captured. In addition, the introduction of mass tags results in the separation of signal from chemical noise. One of the advantages of this method is its ability to identify low abundant proteins and thus has utility in biomarker identifications.[233]

4.8.2.2.3 Chemical Labeling by Isobaric Tagging. This is the most widely used method of quantification as it enables efficient multiplexing based isobaric labeling. There are two isobaric labeling techniques: iTRAQ developed by Ross *et al.*, in 2004[234] and TMT developed by Thompson *et al.*, in 2003.[235] Both the techniques involve derivatization of the peptides at the ε-amino Lys and N-terminal amine with reagents such that the mass of a modified peptide from different samples remains the same. Isobaric labeling leads to elution of all isobaric peptides from different samples at the same time and is recorded as a single *m*/*z* peak in the MS1 scan. The MS/MS fragmentation of the peptides result in the release of the reporter ions from the peptides and the relative peptide abundance in different samples is obtained from the relative intensities of these reporter ions. In iTRAQ labeling either four or eight samples can be multiplexed with the reporter ions having a mass of 114–117 Da, for 4-plex and 113–121 Da (except 120), for 8-plex experiments. The iTRAQ reagent consists of three parts: reporter, balancer and an amine reactive group (NHS ester). The combined mass of the reporter and balancer is 145 Da (for 4-plex) and 305 Da (for 8-plex). If a reporter ion has a mass of 114 Da, the balancer will have a mass of 31 Da. Thus, the mass of the reporter varies from 114–117 (4-plex) and 113–121 Da except 120 (8-plex), while that of the balancer varies from 31 Da to 28 Da (4-plex) and 184 Da to 192 Da (8-plex). Proteolytically digested peptides from different samples are tagged with the isobaric reagents. The samples are then pooled, fractionated using cation exchange and RPLC, followed by mass spectrometric analysis. The peptide masses obtained in the MS1 scan are further fragmented (MS/MS) to release the reporter ions for quantification. Thus, both quantification and identification of a peptide takes place at the MS/MS level. TMT is another method of isobaric labeling, which is capable of multiplexing two, six or ten samples based on the isobaric tags used. In principle, the reagent is similar to iTRAQ, and also consists of three parts: TMT tags, a normalization group and a reactive group. The TMT tag has masses of 126–127 (for 2-plex) and 126–131 (for 6-plex and 10-plex) involving a rearrangement of ^{13}C and ^{15}N atoms. The most commonly used MSs for performing chemical labeling are Q-TOF, TOF/TOF, Orbitrap, ion trap and QQQ. For TMT experiments, a high-energy collision dissociation (HCD)[236] or electron transfer dissociation (ETD)[237] are generally used for fragmentation. However, CID can also be used.[238] A disadvantage of this technique is the mass range of the reporter ion, which overlaps with other compounds. However, the use of HCD where MS2 helps in identification and MS3 helps in quantification of reporter ions[239] circumvents this problem to a great extent.

Another form of chemical isobaric labeling is IPTL, which is less commonly used. The principle of IPTL is similar to iTRAQ and TMT. However,

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

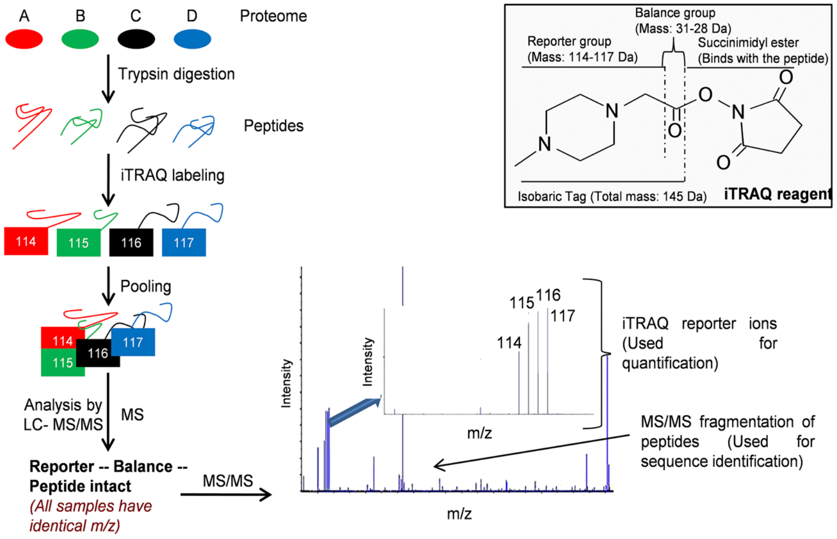
A
B
C
D
Proteome
Trypsin digestion
Peptides
iTRAQ labeling
114
115
116
117
Pooling
114
116
117
115
Analysis by LC- MS/MS
MS
Reporter -- Balance -- Peptide intact
(All samples have identical m/z)
MS/MS
Intensity
Intensity
115
116
117
114
m/z
m/z
iTRAQ reporter ions (Used for quantification)
MS/MS fragmentation of peptides (Used for sequence identification)
Balance group (Mass: 31-28 Da)
Reporter group (Mass: 114-117 Da)
Succinimidyl ester (Binds with the peptide)
O
N
N
O
O
N
O
Isobaric Tag (Total mass: 145 Da)
iTRAQ reagent

Figure 4.6 A schematic diagram of iTRAQ experimentation. The inset consists of the iTRAQ reagent.

it is different from the other two in the mode of digestion, which is usually performed by endopeptidase Lys-C instead of trypsin. The labeling is done at the C-terminal Lys by 2-methoxy-4, 5-dihydro-1H-imidazole (MDHI) or with deuterated form MDHI-d4 followed by tagging the N-termini of the peptide with deuterated succinic anhydride (SA-d4) or normal SA, respectively. Thus, balancing of the deuterated form at the N- and C-termini can create isobaric tags. The quantification is done at the MS2 level but in this case all the b and y ions generated contain the signature tag, which helps in robust quantification of the peptides. There are also some modifications of this method (Figure 4.6).[240–243]

4.8.3 Label-free Quantification

In contrast to quantification using various labeled techniques, label-free quantification is simple, cost effective, and does not involve complex labeling procedures. This has led to increased use of label-free techniques for relative or absolute quantification. There are various types of label-free approaches such as the protein-based spectral count method, extracted ion chromatogram (XIC), and a peptide-based method measuring ion intensity data independent acquisition (DIA).

4.8.3.1 Spectral Counting

The simplest form of label-free quantification method is peptide counting, which is based on the count of unique peptides obtained for a protein and is used as a measure of protein abundance.[86] Spectrum counting,[244] which is

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

based on the sum of intensities of all peptide spectra obtained for a protein, is a better method than peptide counting in terms of reproducibility. This is also known as peptide spectral matching (PSM).[245] PSM counting is based on MS/MS intensity where the average intensities of all the MS/MS ions generated are taken into account for the comparison of protein abundance between two samples.[246] Thus, spectral counting is a simple tool for quantification. Different normalization and statistical methods have been used for quantification,[247] which include the protein abundance index (PAI) and the exponentially modified PAI (emPAI), absolute protein expression (APEX), spectral counting (SpC), spectral abundance factor (SAF), normalized SAF (NSAF), spectral index (SI) or normalized SI (SI_N). PAI gives the abundance of protein and is calculated on the basis of the number of observed peptides divided by the number of theoretically possible tryptic peptides.[248] The exponentially modified form of PAI (emPAI)[249] is calculated as 10^{PAI}–1, which directly indicates protein concentration in terms of molar percentage but the disadvantage is saturation of emPAI calculations for high abundant proteins. A major disadvantage of SpC is its dependence on the physico-chemical properties of protein. This was taken care of in the APEX method, which takes into account the number of observed tryptic peptides and the probability of a peptide being detected, and is corrected computationally by using a machine learning algorithm. It introduces a correction factor (Oi) for the probability of obtaining a tryptic peptide.[177,250] This method provides good accuracy in the measurement of protein abundance and is an openly available tool for analysis.[250,251] Similar to this is the method of modified spectral count index, which is obtained by dividing the observed peptides by the probability of relative identification of protein.[252] Spectral counting is considered to be influenced by the length of the protein. The longer the protein, the higher the number of peptides, and the higher their chance of becoming fragmented and detected. Thus, in SpC, the number of spectral counts for a protein (all MS/MS spectra per peptide) is normalized with the length of the protein to calculate the SAF (SpC/L).[253] Whereas in the NSAF, the SAF of a protein is divided by the sum of the SAFs of all the proteins. Later, it is observed that the length of the protein does not influence spectral counting to a great extent.[247] Thus, another index was introduced, which comprises all the unique peptides identified, MS/MS spectra per peptide, *i.e.* SpC, and the fragment ion intensity from MS/MS spectra taken to calculate the SI called SI_N (N stands for normalized).[254] The SI for a protein is divided by the SIs of all the protein for normalization. The SI_N method is very promising. Another method of absolute and relative quantification is robust intensity-based averaged ratio (RIBAR). In the RIBAR method, calculation is done by taking the average of the $\log_2$ of peptide ratios for a protein. The peptide ratio is calculated using the sum of all the MS/MS intensities obtained from the fragment ions of all the unique peptides, *i.e.* the PSMs for a protein.[255] These methods are based on data-dependent shotgun proteomics and hence are heavily dependent on protein abundance, chromatographic separation and peak width, ion inclusion criteria, instrument type and its speed, reproducibility, and false discovery rate analysis for PSMs. These parameters should be kept in mind when considering the method of analysis.[256]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

4.8.3.2 Extracted Ion Current (XIC)

This system of quantification is based on the calculation of the peak area. For the purpose of comparing two samples, the area under curve (AUC) of all the peptides of a protein are integrated and compared. The AUC for a peptide correlates linearly with the concentration of peptide or protein.[257,258] XIC usually measures the ion abundance or ion counts for a peptide at a particular retention time (RT). The major drawback of this method is its dependence on sample preparation, ionization and reproducibility. Further, peak widening, which results in peptide peak overlap, or the occurrence of multiple peaks for the same peptide are the other drawbacks with this method. Other technical variations like a shift in RT, ion suppression, high background noise and chemical interference also need to be accounted for. Currently, advanced software takes care of some of these issues, such as RT correction within samples, "ideal" peak picking, noise correction, peak normalization and alignment of RT from different samples for intensity comparison.[259] Statistical analysis after proper normalization, peak alignment and multiple runs from the same sample ensure the robustness of this technique.[260] Quantification though XIC requires a defined cycle time for the acquisition of MS and MS/MS spectra, and hence the acquisition of MS/MS spectra should be commensurate with cycle time. The simultaneous MS and MS/MS acquisition makes it difficult to find the right balance, as one may interfere with the other. For instance, increasing the fragmentation using CID may result in a higher number of co-eluting peptides at the MS scan but it may not change the identification through MS/MS events and *vice versa*. To tackle this problem, accurate mass–retention time pair (AMRT) analysis is performed, where the sample is analyzed multiple times separately for MS and MS/MS. MS cycle acquisition in low collision energy helps to generate a precursor ion scan with accurate mass and RT. This can be used to generate a targeted list for the MS/MS cycle obtained in high collision energy mode for identification.[261] The AMRT was the first data independent acquisition technique. This strategy was additionally developed for parallel MS and MS/MS acquisition using low and high collision energy settings alternatively, thereby fragmenting all the peptides (including isotopes and different charge states). This is known as LCMSE or the data independent acquisition mode.[262,263] The cycle time is decided based on the chromatographic elution time such that the MS scan has sufficient data points for peak integration at a specific RT with accurate *m*/*z* and AUC. LC-MS^E utilizes the full cycle time of 4 s for extensive qualitative and quantitative analysis. Intensity-based absolute quantitation is another technique that takes into account the ratio of the sum of all the intensities from the peptides of a protein to the theoretical number of peptides for a protein.[264] This method was developed in 2011,[265] and was found to be more robust and provided better coverage of the proteome than other spectral counting methods.[266]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

4.8.3.3 Data-independent Acquisition (DIA)

The most widely used proteomics approaches are shotgun proteomics and targeted proteomics. Though shotgun proteomics is well suited for identification of large number of proteins it has poor reproducibility and requires extensive fractionation. On the other hand, targeted proteomics approaches can provide quantification along with reproducibility but a large number of proteins cannot be quantified as only few transitions can be targeted per run. To tackle this situation a new method known as DIA was introduced.[262,263] DIA utilizes the advantages of DDA and SRM. This method helps in absolute quantification without limitations to a set of targeted transitions.[267–274] In DIA, analysis is performed in small windows of 10–20 Da and all the MS peaks are allowed to fragment. This continues until the entire mass range is covered for a given proteome. Data analysis is usually done by searching for standard MS/MS spectra in databases using the average m/z value from the given range of the m/z window as the parent ion[268,270] or by searching the pseudo reconstructed precursor and product ion lineage from the co-eluting precursor ions and their potential fragment ions.[263,272,275–277] As large amount of data is generated due to multiple windows, each fragment ion generated should be linked to its corresponding precursor ion for absolute quantification. Further, larger m/z isolation windows result in frequent acquisition but also lead to the co-fragmentation of many precursors making it impossible to trace the lineage of the fragment ion. The use of CID with a narrow isolation window pre-defines the precursor mass, thus eliminating the need for finding the lineage of the fragment molecules.[268,270] Thus the use of narrow windows can improve the data analysis. A modified method of DIA known as sequential window acquisition of all theoretical mass spectra[278,279] has been developed, which generates a fragment ion spectra map resolved in time with 25 Da windows within a mass range of 400–1200 m/z. The data analysis couples the DIA with the spectral library matching approach.[280] This is a targeted data analysis approach where the peptides are identified and quantified on the basis of the information present in spectral libraries generated by DDA. The spectral libraries contain information on the fragment ions, their comparative intensities, and chromatographic correspondence, and are used to mine deeper into the fragment ion spectra map obtained from DIA for collecting the information about the targeted peptide intensity and hence quantification.

4.8.4 Absolute Quantification

4.8.4.1 Absolute Quantification Using the Targeted Proteomics Method (MRM, SRM, PRM)

Absolute quantification using MRM/SRM is achieved by the use of synthetic internal standards. These internal standards are isotopically labeled peptides, which help in the quantification of the corresponding unlabeled

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

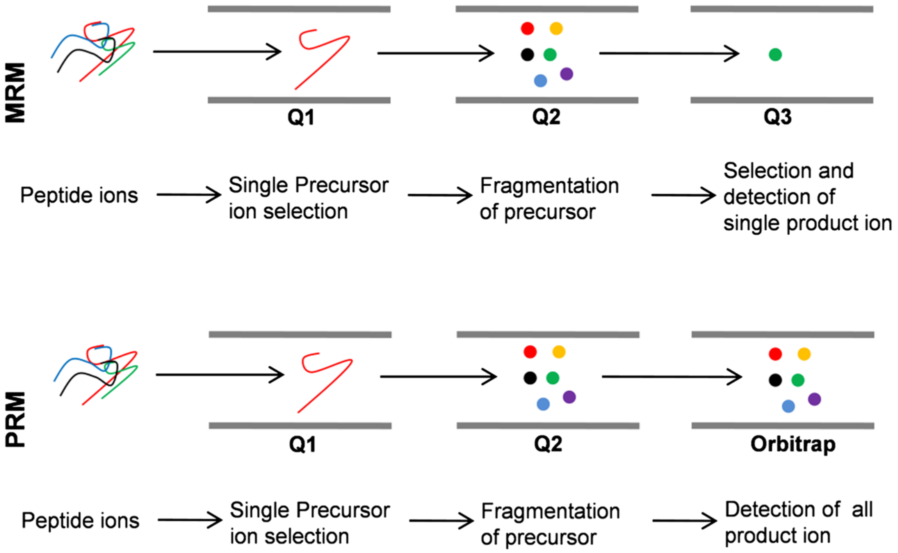
MRM
Q1
Q2
Q3
Peptide ions
Single Precursor ion selection
Fragmentation of precursor
Selection and detection of single product ion
PRM
Q1
Q2
Orbitrap
Peptide ions
Single Precursor ion selection
Fragmentation of precursor
Detection of all product ion

Figure 4.7 A schematic diagram highlighting the differences between MRM and PRM.

peptide in the sample. A standard curve is plotted with the MRM response of known concentrations of the isotopically labeled internal standards.[182,281] The concentration of the desired peptide in the sample is then calculated from the MRM response of the peptide using the standard curve. MRM quantification is mostly done using a triple quadrupole MS or hybrid ion trap MS with pre-defined precursor and product ions. Since these MSs have very high sensitivity, even low abundant peptides (attomoles) can be quantified using MRM[282] with a dynamic range up to 5.[283,284] To calculate the concentration of a protein it is important to quantify at least three or four peptides of the protein for better accuracy.[285] MRM holds great potential in biomarker discovery and validation, and has been used to quantify differentially expressed proteins in various diseases (Figure 4.7).[284,286,287]

4.8.4.2 AQUA Signal Based Quantification

AQUA is a labeled method for absolute quantification and was introduced in 2003 by Gerber *et al.*[182,288] It involves the spiking of a heavy labeled internal standard peptide. The spiked internal standard (of known concentration) and the native peptide ionize with the same efficiency and elute chromatographically at the same time. The standards are spiked before the digestion of the sample and hence are not affected by the loss during the preparation protocol. Quantification is then performed either by XIC or SRM to generate the peak ratios, which aid in getting the absolute quantity of the desired peptide since the concentration of the labeled peptide is known. This technique

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

is also known as stable isotope dilution-multiple reaction monitoring-mass spectrometry (SID-MRM-MS)[289] and has been used extensively for quantification of post-translationally modified peptides. The synthetic peptides are prepared using covalent modifications such as phosphorylation, methylation, acetylation, *etc.* for PTM studies or by using heavy amino acids such as L-alanine (3 ^{13}C,^{15}N), L-proline (5 ^{13}C,^{15}N), L-valine (5 ^{13}C,^{15}N), L-isoleucine (6 ^{13}C,^{15}N), L-leucine (6 ^{13}C, ^{15}N), L-lysine (6 ^{13}C, 2 ^{15}N), L-arginine (6^{13}C, 4 ^{15}N), and L-phenylanaline (9 ^{13}C,^{15}N). Since, the synthesis of the labeled standards is very costly the method can only be used for the study of a limited number of proteins.

4.9 Computational Methods of Proteomics Data Analysis

One of the critical steps in proteomics is the identification of the protein using various databases and search algorithms. The peaks obtained either in MS1 or MS2 are subjected to a database search for identification of peptides and then proteins. Initially there was no dedicated protein database and the information was stored as nucleic acid translated form in databases such as GenBank, the EMBL and the DNA Database of Japan. In 1986 SWISS-PROT was developed as a dedicated database for proteins, and currently contains over 50 000 manually curated proteins. Another database, TrEMBL, contains automated translations of nucleic acid sequence from the EMBL database. The most widely used databases are the Entrez Protein database and the Reference Sequence (RefSeq) database from the US National Center for Biotechnology Information (NCBI), the UniProt–Swiss-Prot database and its extended version TrEMBL, and the International Protein Index (IPI) database from the European Bioinformatics Institute (EBI). Depending on the type of MS used, the raw data is commonly acquired in formats such as Xcalibur/RAW, Analyst/WIFF, MassLynx/RAW or BAF. For analysis of the raw data the spectra is converted to plain formats such as mgf, dta and pkl files and the output files are in the formats of XML-files, mzXML, mzDATA, pepXML and protXML. The most important part of data analysis is the database search tool. These tools can be divided into categories such as database search tools, *de novo* sequencing tools, statistical tools for validation of peptide and proteins, protein quantification tools, and storage and mining tools. Some of the databases, which are available freely or in licensed form, have been listed in Tables 4.2 and 4.3. The database searching tools provides the best match scores with some statistically significant values (E-values) to peptides/proteins by matching experimental spectra with the theoretical spectra in the database. These scores are highly variable depending on the type of instrument and algorithm used. Similarly, the spectral matching method of identification is based on the matching of the observed spectra with the spectral libraries generated experimentally. Identification of the protein by *de novo* sequencing requires high quality spectra and allows the identification of peptides that are not

Table 4.2 Types of tools for proteomics data analysis.

Database search tools	HTML link
SEQUEST[158]	http://www.thermo.com
MASCOT[159]	http://matrixscience.coma
OMSSA[292]	http://pubchem.ncbi.nlm.nih.gov/omssaa,b
MassMatrix	http://www.massmatrix.net/
TANDEM[293]	http://www.thegpm.orga,b
SpectrumMill	http://www.chem.agilent.com
Phenyx[294]	http://www.phenyx-ms.com
VEMS[295]	http://personal.cicbiogune.es/rmatthiesenb
MyriMatch[296]	https://svn.code.sf.net/p/proteowizard/code/trunk/pwiz/pwiz_tools/Bumbershoot/myrimatch/doc/index.html
MassWiz[297]	https://sourceforge.net/projects/masswiz/
PEAKS DB[298]	http://www1.bioinfor.com/peaks/features/peaksdb.html
ProbID[299]	http://tools.proteomecenter.org/wiki/index.,php?title=Software:ProbIDb
ProteinProspector[300]	http://prospector.ucsf.edua
De novo sequencing tools	HTML link
PepNovo[301]	http://peptide.ucsd.edu/pepnovo.pya,b
PEAKS De Novo[302]	http://www1.bioinfor.com/peaks/features/denovo.html
Sequit	http://www.proteomefactory.com
InSpecT[303]	http://proteomics.ucsd.edu/ProteoSAFe/
lutefisk[304]	http://www.hairyfatguy.com/lutefisk/
MSNovo[305]	http://msms.usc.edu/supplementary/msnovo
Statistical validation of peptide and proteins	HTML link
PeptideProphet[306]	http://www.proteomecenter.org/software.phpb
ProteinProphet[307]	http://www.proteomecenter.org/software.phpb
Scaffold[308]	http://www.proteomesoftware.com
XTandem Parser[309]	http://compomics.github.io/projects/xtandem-parser.html
X!TandemPipeline	X!TandemPipeline
Databases for storage and mining	HTML link
PeptideAtlas[310]	http://www.peptideatlas.org
Proteios	http://www.proteios.org
ProteoWizard[311]	http://proteowizard.sourceforge.net/
PRIDE[312]	http://www.ebi.ac.uk/pride
XCMS	https://metlin.scripps.edu/xcms/
TPP	http://tools.proteomecenter.org/software.php
SBEAMS	http://sbeams.org
Proteios	http://www.proteios.org/
OpenMS	http://www.openms.de/about/
CPFP[313]	http://cpfp.sourceforge.net/
CPAS[314]	https://www.labkey.org

Table 4.3 Types of quantification tools for proteomics.

Approaches	Method	Tools	HTML link
2D-DIGE		MSQuant[315]	http://msquant.sourceforge.net
		PDQuest	http://www.bio-rad.com/
		Progenesis	http://www.nonlinear.com/
		SameSpots Melanie[316]	http://www.genebio.com/
		Progenesis LCMS msInspect[317]	http://www.nonlinear.com/
		DeCyder MS[318]	http://www.gelifesciences.com
Labeled relative quantification	Methods	Tools	HTML link
Metabolic labeling	15 N	MSQuant[315]	http://msquant.sourceforge.net
		XPRESS[219]	http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS
		Scaffold[308]	http://www.proteomesoftware.com/
		MSQuant[319]	http://msquant.sourceforge.net/
	SILAC	AYUMS[320]	www.csml.org/ayums/
		MaxQUANT[321,322]	www.maxquant.org
		ASAP Ratio[323]	http://tools.proteomecenter.org/wiki/index.php?title=Software:ASAPRatio
		XPRESS[219]	http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS
Enzymatic and chemical labeling	ICAT,18O labeling	ASAP Ratio[323]	http://tools.proteomecenter.org/wiki/index.php?title=Software:ASAPRatio
		Census[324]	http://fields.scripps.edu/census/index.php
		MSQuant[315]	http://msquant.sourceforge.net
		XPRESS[219]	http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS
		ZoomQuant[325]	http://proteomics.mcw.edu/zoomquant.html
		RAAMS	http://informatics.mayo.edu/svn/trunk/mprc/raams/index.html
		QUIL[326]	

(*continued*)

Table 4.3 (*continued*)

Approaches	Method	Tools	HTML link
Chemical labeling by isobaric tagging	iTRAQ	i-Tracker	www.cranfield.ac.uk/health/researchareas/bioinformatics/page6801.jsp
		jTraqX	http://sourceforge.net/projects/protms/
		QUANT	https://sourceforge.net/projects/protms/
		ProteinPilot[327]	http://www.absciex.com/
		IsobariQ	http://norwegian-proteomics-society.uio.no/isobariq/
	TMT	Multi-Q[328]	http://ms.iis.sinica.edu.tw/Multi-Q-Web/
		IsobariQ[329]	http://norwegian-proteomics-society.uio.no/isobariq/
	IPTL	IsobariQ[329]	http://norwegian-proteomics-society.uio.no/isobariq/
	ICPL	ASAP Ratio[323]	http://tools.proteomecenter.org/wiki/index.php?title=Software:ASAPRatio
		Census[324]	http://fields.scripps.edu/census/index.php
		MSQuant[315]	http://msquant.sourceforge.net
		XPRESS[219]	http://tools.proteomecenter.org/wiki/index.php?title=Software:XPRESS
		ZoomQuant[325]	http://proteomics.mcw.edu/zoomquant.html
		PepC[330]	http://sashimi.svn.sourceforge.net/viewvc/sashimi/trunk/trans_proteomic_pipeline/src/Quantitation/Pepc/
Label free relative quantification	Methods	Tools	HTML link
Spectrum counting		ProteoIQ	http://www.bioinquire.com
		Scaffold[308]	http://www.proteomesoftware.com/
		Census[324]	http://fields.scripps.edu
	PAI and exponentially modified PAI	APEX[251]	http://pfgrc.jcvi.org/index.php/bioinformatics/apex.html
		Mascot	http://www.matrixscience.com/
		emPAI Calc	http://empai.iab.keio.ac.jp
	APEX	APEX[251]	http://pfgrc.jcvi.org/index.php/bioinformatics/apex.html
		Mascot[159]	http://www.matrixscience.com/

Extracted ion chromatogram (XIC)		MSight[331]	http://www.expasy.org/Msight
		TOPP	
		PEPPeR[332]	http://www.broadinstitute.org
		SuperHirn[333]	http://prottools.ethz.ch/muellelu/web/SuperHirn.php
		DeCyder MS[318]	http://www.gelifesciences.com
		SIEVE	http://www.thermo.com
		ProteinLynx[262]	http://www.waters.com
		Elucidator	http://www.rosettabio.com
		Expressionist	http://www.genedata.com
		Progenesis-LC	http://www.nonlinear.com
		Census[324]	http://fields.scripps.edu
		Corra[334]	http://corra.sourceforge.net
		MSInspect[317]	http://proteomics.fhcrc.org
		MSQuant[319,335]	http://msquant.sourceforge.net
		Serac[317,334]	
		SpecArray[336]	http://sourceforge.net/projects/sashimi/files/SpecArray
Data independent acquisition (DIA)	SWATH	Skyline[187]	https://brendanx-uw1.gs.washington.edu
Absolute quantification	MRM,SRM, PRM	MRMer[337]	http://proteomics.fhcrc.org/CPL/MRMer.html
		Skyline[187]	https://brendanx-uw1.gs.washington.edu
		MaxQuant[321]	http://maxquant.org/
		ATAQS[338]	http://tools.proteomecenter.org/ATAQS/ATAQS.html

available in the database, leading to the discovery of new proteins. The new algorithms for statistical validation of proteins and peptides uses the false discovery rate (FDR) where experimental peptides are matched and scored with a cut-off value to a targeted database and a decoy database (randomly shuffled and reverse sequenced for the same target database) separately. The FDR is then calculated as the ratio of peptides matched in the decoy database divided by the number of peptides matched in the target database above the cut-off score. The FDR helps in reducing the number of false positives in the data. Now, freely available data storage and mining tools are also available, which has improved the quality of data interpretation and analysis.

4.10 Applications of Proteomics

Proteomics have emerged as a fascinating tool for addressing biological phenomena in a more comprehensive and robust manner. It holds immense potential for discovering newer biomarkers for disease biology to understand the pathophysiology of a particular biological state at a given time-scale. The evolution of the next-generation MS-based approaches has enabled much higher resolution for investigating the proteome more comprehensively. It is evident that analyzing proteins in general is fraught with several challenges. The field of biochemistry has provided a fundamental scaffold for researchers to study the structure–functional aspect of several proteins with respect to specific biological interest. However, the application of proteomics holds enormous importance as it enables the study of the entire complement of proteins present within a specific cell, tissue or organ. Furthermore, the state of proteins is highly dynamic in terms of different translational modifications which include another layer of complexity. To dissect the complex network of proteome, the MS-based approaches discussed above, mainly developed in the last decade, have opened up a new avenue of scientific advancement in biological research. The power of unbiased screening across proteomes in several biological states through proteomics provides the handle to such discoveries. However, the scope of proteomics has been extended to various other dimensions. One of the important aspects is detection and quantification of different PTMs in a given time and space. The combination of a biochemical method to enrich specific PTMs, such as phosphorylation, with high resolution mass spectrometric analysis for the detection and quantification of these modified peptides actually integrates the study of comprehensive cellular signaling pathways and cross-talk, which are really difficult and challenging to probe. Quantitative phosphoproteomics-based proteomics have revealed newer cellular signaling cross-talk in many diseases such as cancer, cardiovascular diseases, diabetes, *etc.*[290,291] In hyperglycemic conditions, the advanced glycation end products and several other site-specific N-linked and O-linked glycosylations have been identified, which play critical roles in development and disease pathogenesis. In this context, the parallel development of newer bioinformatics approaches to analyze MS-based data is the key to success. Many researchers have made significant progress in analyzing this complex set of mass spectrometric data acquired with newer fragmentation methods such as ETD and HCD, as mentioned earlier. The dramatically rapid progress in resolution, accuracy, sensitivity and speed of high-end MSs coupled with the development of newer and better bioinformatics tools have been a major breakthrough in the proteomics arena. However, even in the current scenario we believe that proteomics is still in its infancy, with huge potential as a more accurate, sensitive and specific analytical platform. The development of single-cell-based proteomics approaches exploring quantification, delineating several PTMs, and identifying signaling cross-talk is also on the horizon.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

Acknowledgement

We acknowledge the support from Council of Scientific and Industrial Research (CSIR), Ministry of Science and Technology, Govt of India, CSIR-Institute of Genomics and Integrative Biology, Academy of Scientific & Innovative Research (AcSIR).

References

1. H. Hartley, Origin of the Word ‘Protein’, *Nature*, 1951, **168**, 244.
2. B. Kuska, Beer, Bethesda, and biology: how "genomics" came into being, *J. Natl. Cancer Inst.*, 1998, **90**(2), 93.
3. M. J. Dunn, 2D Electrophoresis: From Protein Maps to Genomes. Proceedings of the International Meeting. Siena, Italy, 1994, *Electrophoresis*, 1995, **16**(7), 1077–1322.
4. M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith and D. F. Hochstrasser, *et al.*, Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it, *Biotechnol. Genet. Eng. Rev.*, 1996, **13**, 19–50.
5. M. R. Wilkins, I. Lindskog, E. Gasteiger, A. Bairoch, J. C. Sanchez and D. F. Hochstrasser, *et al.*, Detailed peptide characterization using PEPTIDEMASS–a World-Wide-Web-accessible tool, *Electrophoresis*, 1997, **18**(3–4), 403–408.
6. N. L. Anderson and N. G. Anderson, Proteome and proteomics: new technologies, new concepts, and new words, *Electrophoresis*, 1998, **19**(11), 1853–1861.
7. P. H. O'Farrell, High resolution two-dimensional electrophoresis of proteins, *J. Biol. Chem.*, 1975, **250**(10), 4007–4021.
8. J. Klose, Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals, *Humangenetik*, 1975, **26**(3), 231–243.
9. G. A. Scheele, Two-dimensional gel analysis of soluble proteins. Charaterization of guinea pig exocrine pancreatic proteins, *J. Biol. Chem.*, 1975, **250**(14), 5375–5385.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

10. P. Edman and G. Begg, A protein sequenator, *Eur. J. Biochem.*, 1967, **1**(1), 80–91.
11. S. C. Moyer, B. A. Budnik, J. L. Pittman, C. E. Costello and P. B. O'Connor, Attomole peptide analysis by high-pressure matrix-assisted laser desorption/ionization Fourier transform mass spectrometry, *Anal. Chem.*, 2003, **75**(23), 6449–6454.
12. M. Balogh, Debating resolution and mass accuracy in mass spectrometry, *Spectroscopy*, 2004, **19**, 34–40.
13. L. Pasa-Tolic, C. Masselon, R. C. Barry, Y. Shen and R. D. Smith, Proteomic analyses using an accurate mass and time tag strategy, *BioTechniques*, 2004, **37**(4), 621–624.
14. A. Makarov, E. Denisov, O. Lange and S. Horning, Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer, *J. Am. Soc. Mass Spectrom.*, 2006, **17**(7), 977–982.
15. R. J. Weber, A. D. Southam, U. Sommer and M. R. Viant, Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification, *Anal. Chem.*, 2011, **83**(10), 3737–3743.
16. D. F. Hochstrasser and J.-C. Sanchez, The dynamic range of protein expression: a challenge for proteomic research, *Proteins*, 2000, **1109**, 4.
17. L. Wu and D. K. Han, Overcoming the dynamic range problem in mass spectrometry-based shotgun proteomics, *Expert Rev. Proteomics*, 2006, **3**(6), 611–619.
18. N. L. Anderson and N. G. Anderson, The human plasma proteome: history, character, and diagnostic prospects, *Mol. Cell. Proteomics*, 2002, **1**(11), 845–867.
19. N. L. Anderson, M. Polanski, R. Pieper, T. Gatlin, R. S. Tirumalai and T. P. Conrads, *et al.*, The human plasma proteome: a nonredundant list developed by combination of four separate sources, *Mol. Cell. Proteomics*, 2004, **3**(4), 311–326.
20. D. J. Pappin, P. Hojrup and A. J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting, *Curr. Biol.*, 1993, **3**(6), 327–332.
21. S. Ahmad, E. Sundaramoorthy, R. Arora, S. Sen, G. Karthikeyan and S. Sengupta, Progressive degradation of serum samples limits proteomic biomarker discovery, *Anal. Biochem.*, 2009, **394**(2), 237–242.
22. M. Fukuda, N. Islam, S. H. Woo, A. Yamagishi, M. Takaoka and H. Hirano, Assessing matrix assisted laser desorption/ionization-time of flight-mass spectrometry as a means of rapid embryo protein identification in rice, *Electrophoresis*, 2003, **24**(7–8), 1319–1329.
23. P. Giansanti, L. Tsiatsiani, T. Y. Low and A. J. Heck, Six alternative proteases for mass spectrometry-based proteomics beyond trypsin, *Nat. Protoc.*, 2016, **11**(5), 993–1006.
24. H. Hahne, S. Wolff, M. Hecker and D. Becher, From complementarity to comprehensiveness–targeting the membrane proteome of growing Bacillus subtilis by divergent approaches, *Proteomics*, 2008, **8**(19), 4123–4136.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

25. Y. Fang, D. P. Robinson and L. J. Foster, Quantitative analysis of proteome coverage and recovery rates for upstream fractionation methods in proteomics, *J. Proteome Res.*, 2010, **9**(4), 1902–1912.
26. S. R. Piersma, U. Fiedler, S. Span, A. Lingnau, T. V. Pham and S. Hoffmann, *et al.*, Workflow comparison for label-free, quantitative secretome proteomics for cancer biomarker discovery: method evaluation, differential analysis, and verification in serum, *J. Proteome Res.*, 2010, **9**(4), 1913–1922.
27. S. Magdeldin, S. Enany, Y. Yoshida, B. Xu, Y. Zhang and Z. Zureena, *et al.*, Basics and recent advances of two dimensional- polyacrylamide gel electrophoresis, *Clin. Proteomics*, 2014, **11**(1), 16.
28. A. Gorg, G. Boguth, C. Obermaier, A. Posch and W. Weiss, Two-dimensional polyacrylamide gel electrophoresis with immobilized pH gradients in the first dimension (IPG-Dalt): the state of the art and the controversy of vertical *versus* horizontal systems, *Electrophoresis*, 1995, **16**(7), 1079–1086.
29. A. Chrambach, P. Doerr, G. R. Finlayson, L. E. Miles, R. Sherins and D. Rodbard, Instability of pH gradients formed by isoelectric focusing in polyacrylamide gel, *Ann. N. Y. Acad. Sci.*, 1973, **209**, 44–64.
30. P. G. Righetti, E. Gianazza, C. Gelfi, M. Chiari and P. K. Sinha, Isoelectric focusing in immobilized pH gradients, *Anal. Chem.*, 1989, **61**(15), 1602–1612.
31. A. Gorg, W. Postel and S. Gunther, The current state of two-dimensional electrophoresis with immobilized pH gradients, *Electrophoresis*, 1988, **9**(9), 531–546.
32. A. Gorg, G. Boguth, C. Obermaier and W. Weiss, Two-dimensional electrophoresis of proteins in an immobilized pH 4-12 gradient, *Electrophoresis*, 1998, **19**(8–9), 1516–1519.
33. A. Gorg, C. Obermaier, G. Boguth and W. Weiss, Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures, *Electrophoresis*, 1999, **20**(4–5), 712–717.
34. M. R. de Moreno, J. F. Smith and R. V. Smith, Mechanism studies of coomassie blue and silver staining of proteins, *J. Pharm. Sci.*, 1986, **75**(9), 907–911.
35. R. H. Butt and J. R. Coorssen, Coomassie blue as a near-infrared fluorescent stain: a systematic comparison with Sypro Ruby for in-gel protein detection, *Mol. Cell. Proteomics*, 2013, **12**(12), 3834–3850.
36. K. Berggren, T. H. Steinberg, W. M. Lauber, J. A. Carroll, M. F. Lopez and E. Chernokalskaya, *et al.*, A luminescent ruthenium complex for ultrasensitive detection of proteins immobilized on membrane supports, *Anal. Biochem.*, 1999, **276**(2), 129–143.
37. T. H. Steinberg, L. J. Jones, R. P. Haugland and V. L. Singer, SYPRO orange and SYPRO red protein gel stains: one-step fluorescent staining of denaturing gels for detection of nanogram levels of protein, *Anal. Biochem.*, 1996, **239**(2), 223–237.

38. C. Winkler, K. Denker, S. Wortelkamp and A. Sickmann, Silver- and Coomassie-staining protocols: detection limits and compatibility with ESI MS, *Electrophoresis*, 2007, **28**(12), 2095–2099.
39. R. Westermeier and R. Marouga, Protein detection methods in proteomics research, *Biosci. Rep.*, 2005, **25**(1–2), 19–32.
40. S. P. Gygi, G. L. Corthals, Y. Zhang, Y. Rochon and R. Aebersold, Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**(17), 9390–9395.
41. R. Slibinskas, R. Razanskas, R. Zinkeviciute and E. Ciplys, Comparison of first dimension IPG and NEPHGE techniques in two-dimensional gel electrophoresis experiment with cytosolic unfolded protein response in Saccharomyces cerevisiae, *Proteome Sci.*, 2013, **11**(1), 36.
42. H. Nordvarg, J. Flensburg, O. Ronn, J. Ohman, R. Marouga and B. Lundgren, *et al.*, A proteomics approach to the study of absorption, distribution, metabolism, excretion, and toxicity, *J. Biomol. Tech.*, 2004, **15**(4), 265–275.
43. E. Zacharia, E. Kostopoulou, D. Maroulis and S. Kossida, A Spot Segmentation Approach for 2D Gel Electrophoresis Images Based on 2D Histograms, *20th International Conference on Pattern Recognition (ICPR)*, IEEE, 2010, pp. 2540–2543.
44. O. Carrette, P. R. Burkhard, J. C. Sanchez and D. F. Hochstrasser, State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research, *Nat. Protoc.*, 2006, **1**(2), 812–823.
45. M. R. Richardson, S. Liu, H. N. Ringham, V. Chan and F. A. Witzmann, Sample complexity reduction for two-dimensional electrophoresis using solution isoelectric focusing prefractionation, *Electrophoresis*, 2008, **29**(12), 2637–2644.
46. T. Stasyk and L. A. Huber, Zooming in: fractionation strategies in proteomics, *Proteomics*, 2004, **4**(12), 3704–3716.
47. H. Schagger and G. von Jagow, Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form, *Anal. Biochem.*, 1991, **199**(2), 223–231.
48. H. Eubel, H. P. Braun and A. H. Millar, Blue-native PAGE in plants: a tool in analysis of protein-protein interactions, *Plant Methods*, 2005, **1**(1), 11.
49. Q. Meng, L. Rao, X. Xiang, C. Zhou, X. Zhang and Y. Pan, A systematic strategy for proteomic analysis of chloroplast protein complexes in wheat, *Biosci., Biotechnol., Biochem.*, 2011, **75**(11), 2194–2199.
50. J. P. Lasserre and A. Menard, Two-dimensional blue native/SDS gel electrophoresis of multiprotein complexes, *Methods Mol. Biol.*, 2012, **869**, 317–337.
51. A. Vertommen, B. Panis, R. Swennen and S. C. Carpentier, Challenges and solutions for the identification of membrane proteins in non-model plants, *J. Proteomics*, 2011, **74**(8), 1165–1181.
52. X. Zhao, G. Li and S. Liang, Several affinity tags commonly used in chromatographic purification, *J. Anal. Methods Chem.*, 2013, **2013**, 581093.

53. Y. Lv, X. Bao, H. Liu, J. Ren and S. Guo, Purification and characterization of calcium-binding soybean protein hydrolysates by Ca2+/Fe3+ immobilized metal affinity chromatography (IMAC), *Food Chem.*, 2013, **141**(3), 1645–1650.
54. K. Karkra, K. K. R. Tetala and M. A. Vijayalakshmi, A structure based plasma protein pre-fractionation using conjoint immobilized metal/chelate affinity (IMA) system, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2017, **1052**, 1–9.
55. A. J. De Vries, M. LePage, R. Beau and C. L. Guillemin, Evaluation of porous silica beads as a new packing material for chromatographic columns. Application in gel permeation chromatography, *Anal. Chem.*, 1967, **39**(8), 935–939.
56. A. Rembaum, S.-P. S. Yen and W. J. Dreyer, Crosslinked, porous, polyacrylate beads, Google Patents, 1976.
57. J. G. Dorsey and W. T. Cooper, Retention mechanisms of bonded-phase liquid chromatography, *Anal. Chem.*, 1994, **66**(17), 857A–867A.
58. R. I. Boysen and M. T. Hearn, HPLC of peptides and proteins: standard operating conditions, *Curr. Protoc. Mol. Biol.*, 2001, **10**, 3.
59. J. L. Rafferty, J. I. Siepmann and M. R. Schure, Mobile phase effects in reversed-phase liquid chromatography: a comparison of acetonitrile/water and methanol/water solvents as studied by molecular simulation, *J. Chromatogr. A*, 2011, **1218**(16), 2203–2213.
60. M. J. Hetem, J. W. De Haan, H. A. Claessens, L. J. Van de Ven, C. A. Cramers and J. N. Kinkel, Influence of alkyl chain length on the stability of n-alkyl-modified reversed phases. 1. Chromatographic and physical analysis, *Anal. Chem.*, 1990, **62**(21), 2288–2296.
61. J. Pearson and F. Regnier, The Influence of Reversed-Phase n-Alkyl Chain Length on Protein Retention, Resolution and Recovery: Implications for Preparative HPLC, *J. Liq. Chromatogr.*, 1983, **6**(3), 497–510.
62. H. Minakuchi, K. Nakanishi, N. Soga, N. Ishizuka and N. Tanaka, Effect of domain size on the performance of octadecylsilylated continuous porous silica columns in reversed-phase liquid chromatography, *J. Chromatogr. A*, 1998, **797**(1), 121–131.
63. J. Larmann, J. DeStefano, A. Goldberg, R. Stout, L. Snyder and M. Stadalius, Separation of macromolecules by reversed-phase high-performance liquid chromatography: Pore-size and surface-area effects for polystyrene samples of varying molecular weight, *J. Chromatogr. A*, 1983, **255**, 163–189.
64. Y. Van Wanseele, J. Viaene, L. Van den Borre, K. Dewachter, Y. Vander Heyden and I. Smolders, *et al.*, LC-method development for the quantification of neuromedin-like peptides. Emphasis on column choice and mobile phase composition, *J. Pharm. Biomed. Anal.*, 2017, **137**, 104–112.
65. A. L. Capriotti, C. Cavaliere, A. Cavazzini, F. Gasparrini, G. Pierri and S. Piovesana, *et al.*, A multidimensional liquid chromatography-tandem mass spectrometry platform to improve protein identification in high-throughput shotgun proteomics, *J. Chromatogr. A*, 2017, **1498**, 176–182.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

66. G. Hong, M. Gao, G. Yan, X. Guan, Q. Tao and X. Zhang, Optimization of two-dimensional high performance liquid chromatographic columns for highly efficient separation of intact proteins, *Se Pu*, 2010, **28**(2), 158–162.
67. Q. Zhao, L. Sun, Y. Liang, Q. Wu, H. Yuan and Z. Liang, *et al.*, Prefractionation and separation by C8 stationary phase: effective strategies for integral membrane proteins analysis, *Talanta*, 2012, **88**, 567–572.
68. W. Naidong, Bioanalytical liquid chromatography tandem mass spectrometry methods on underivatized silica columns with aqueous/organic mobile phases, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2003, **796**(2), 209–224.
69. H. Lindner, B. Sarg, C. Meraner and W. Helliger, Separation of acetylated core histones by hydrophilic-interaction liquid chromatography, *J. Chromatogr. A*, 1996, **743**(1), 137–144.
70. A. J. Alpert, Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides, *Anal. Chem.*, 2008, **80**(1), 62–76, Epub 2007/11/22.
71. P. Hagglund, J. Bunkenborg, F. Elortza, O. N. Jensen and P. Roepstorff, A new strategy for identification of *N*-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation, *J. Proteome Res.*, 2004, **3**(3), 556–566.
72. Z. Guo, A. Lei, Y. Zhang, Q. Xu, X. Xue and F. Zhang, *et al.*, "Click saccharides": novel separation materials for hydrophilic interaction liquid chromatography, *Chem. Commun.*, 2007, (24), 2491–2493.
73. I. Ali, H. Y. Aboul-Enein, P. Singh, R. Singh and B. Sharma, Separation of biological proteins by liquid chromatography, *Saudi Pharm. J.*, 2010, **18**(2), 59–73.
74. A. Ros, M. Faupel, H. Mees, J. Oostrum, R. Ferrigno and F. Reymond, *et al.*, Protein purification by Off-Gel electrophoresis, *Proteomics*, 2002, **2**(2), 151–156.
75. P. E. Michel, F. Reymond, I. L. Arnaud, J. Josserand, H. H. Girault and J. S. Rossier, Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing, *Electrophoresis*, 2003, **24**(1–2), 3–11.
76. P. Horth, C. A. Miller, T. Preckel and C. Wenz, Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis, *Mol. Cell. Proteomics*, 2006, **5**(10), 1968–1974.
77. M. R. Pergande and S. M. Cologna, Isoelectric Point Separations of Peptides and Proteins, *Proteomes*, 2017, **5**(1), 4.
78. M. Heller, P. E. Michel, P. Morier, D. Crettaz, C. Wenz and J. D. Tissot, *et al.*, Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma, *Electrophoresis*, 2005, **26**(6), 1174–1188.
79. S. Krishnan, M. Gaspari, A. Della Corte, P. Bianchi, M. Crescente and C. Cerletti, *et al.*, OFFgel-based multidimensional LC-MS/MS approach to the cataloguing of the human platelet proteome for an interactomic profile, *Electrophoresis*, 2011, **32**(6–7), 686–695.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

80. A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize and D. R. Morris, *et al.*, Direct analysis of protein complexes using mass spectrometry, *Nat. Biotechnol.*, 1999, **17**(7), 676–682.
81. G. J. Opiteck, S. M. Ramirez, J. W. Jorgenson and M. A. Moseley 3rd, Comprehensive two-dimensional high-performance liquid chromatography for the isolation of overexpressed proteins and proteome mapping, *Anal. Biochem.*, 1998, **258**(2), 349–361.
82. L. A. Holland and J. W. Jorgenson, Separation of nanoliter samples of biological amines by a comprehensive two-dimensional microcolumn liquid chromatography system, *Anal. Chem.*, 1995, **67**(18), 3275–3283.
83. Z. Tian, R. Zhao, N. Tolic, R. J. Moore, D. L. Stenoien and E. W. Robinson, *et al.*, Two-dimensional liquid chromatography system for online top-down mass spectrometry, *Proteomics*, 2010, **10**(20), 3610–3620.
84. G. J. Opiteck and J. W. Jorgenson, Two-dimensional SEC/RPLC coupled to mass spectrometry for the analysis of peptides, *Anal. Chem.*, 1997, **69**(13), 2283–2291.
85. A. C. Paoletti, B. Zybailov and M. P. Washburn, Principles and applications of multidimensional protein identification technology, *Expert Rev. Proteomics*, 2004, **1**(3), 275–282.
86. M. P. Washburn, D. Wolters and J. R. Yates 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nat. Biotechnol.*, 2001, **19**(3), 242–247.
87. H. Wang and S. Hanash, Multi-dimensional liquid phase based separations in proteomics, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2003, **787**(1), 11–18.
88. N. Nagaraj, N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr and O. Hoerning, *et al.*, System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap, *Mol. Cell. Proteomics*, 2012, **11**(3), M111 013722.
89. S. S. Thakur, T. Geiger, B. Chatterjee, P. Bandilla, F. Frohlich and J. Cox, *et al.*, Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation, *Mol. Cell. Proteomics*, 2011, **10**(8), M110 003699.
90. A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin and M. S. Westphall, *et al.*, The one hour yeast proteome, *Mol. Cell. Proteomics*, 2014, **13**(1), 339–347.
91. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse, Electrospray ionization for mass spectrometry of large biomolecules, *Science*, 1989, **246**(4926), 64–71.
92. S. Banerjee and S. Mazumdar, Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte, *Int. J. Anal. Chem.*, 2012, **2012**, 282574.
93. M. Karas, D. Bachmann, U. E. Bahr and F. Hillenkamp, Matrix-assisted ultraviolet laser desorption of non-volatile compounds, *Int. J. Mass Spectrom. Ion Processes*, 1987, 7853–7868.

94. F. Hillenkamp and M. Karas, Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization, *Methods Enzymol.*, 1990, **193**, 280–295.
95. K. Strupat, M. Karas and F. Hillenkamp, 2,5-Dihydroxybenzoic acid a new matrix for laser desorption—ionization mass spectrometry, *Int. J. Mass Spectrom. Ion Processes*, 1991, **111**, 89–102.
96. R. C. Beavis and B. T. Chait, Matrix-assisted laser-desorption mass spectrometry using 355 nm radiation, *Rapid Commun. Mass Spectrom.*, 1989, **3**(12), 436–439.
97. R. C. Beavis and B. T. Chait, Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins, *Rapid Commun. Mass Spectrom.*, 1989, **3**(12), 432–435.
98. R. C. Beavis, T. Chaudhary and B. T. Chait, α-Cyano-4-hydroxycinnamic acid as a matrix for matrixassisted laser desorption mass spectromtry, *Org. Mass Spectrom.*, 1992, **27**(2), 156–158.
99. P. C. Liao and J. Allison, Dissecting matrix-assisted laser desorptionionization mass spectra, *J. Mass Spectrom.*, 1995, **30**(5), 763–766.
100. L. H. Silvertand, J. S. Torano, G. J. de Jong and W. P. van Bennekom, Improved repeatability and matrix-assisted desorption/ionization–time of flight mass spectrometry compatibility in capillary isoelectric focusing, *Electrophoresis*, 2008, **29**(10), 1985–1996.
101. J. Zheng, N. Li, M. Ridyard, H. Dai, S. M. Robbins and L. Li, Simple and robust two-layer matrix/sample preparation method for MALDI MS/MS analysis of peptides, *J. Proteome Res.*, 2005, **4**(5), 1709–1716.
102. G. B. Foote and P. R. Brazier-Smith, Disruption of water drops by electrical forces, *J. Geophys. Res.*, 1972, **77**(9), 1695–1699.
103. Y. Shen, R. J. Moore, R. Zhao, J. Blonder, D. L. Auberry and C. Masselon, *et al.*, High-efficiency on-line solid-phase extraction coupling to 15-150-microm-i.d. column liquid chromatography for proteomic analysis, *Anal. Chem.*, 2003, **75**(14), 3596–3605.
104. Y. Shen, R. Zhao, S. J. Berger, G. A. Anderson, N. Rodriguez and R. D. Smith, High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics, *Anal. Chem.*, 2002, **74**(16), 4235–4249.
105. M. R. Emmett and R. M. Caprioli, Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins, *J. Am. Soc. Mass Spectrom.*, 1994, **5**(7), 605–613.
106. M. Mann, A shortcut to interesting human genes: peptide sequence tags, expressed-sequence tags and computers, *Trends Biochem. Sci.*, 1996, **21**(12), 494–495.
107. M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer and T. Fotsis, *et al.*, Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry, *Nature*, 1996, **379**(6564), 466–469.
108. K. Breuker, M. Jin, X. Han, H. Jiang and F. W. McLafferty, Top-down identification and characterization of biomolecules by mass spectrometry, *J. Am. Soc. Mass Spectrom.*, 2008, **19**(8), 1045–1053.

109. G. E. Reid and S. A. McLuckey, 'Top down' protein characterization *via* tandem mass spectrometry, *J. Mass Spectrom.*, 2002, **37**(7), 663–675.

110. W. H. McDonald and J. R. Yates 3rd, Shotgun proteomics and biomarker discovery, *Dis. Markers*, 2002, **18**(2), 99–105.

111. B. T. Chait, Chemistry. Mass spectrometry: bottom-up or top-down? *Science*, 2006, **314**(5796), 65–66.

112. A. Schmidt, M. Claassen and R. Aebersold, Directed mass spectrometry: towards hypothesis-driven proteomics, *Curr. Opin. Chem. Biol.*, 2009, **13**(5–6), 510–517.

113. J. Stahl-Zeng, V. Lange, R. Ossola, K. Eckhardt, W. Krek and R. Aebersold, *et al.*, High sensitivity detection of plasma proteins by multiple reaction monitoring of *N*-glycosites, *Mol. Cell. Proteomics*, 2007, **6**(10), 1809–1817.

114. P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon and R. Aebersold, Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics, *Cell.*, 2009, **138**(4), 795–806.

115. D. R. Ahlf, P. D. Compton, J. C. Tran, B. P. Early, P. M. Thomas and N. L. Kelleher, Evaluation of the compact high-field orbitrap for top-down proteomics of human cells, *J. Proteome Res.*, 2012, **11**(8), 4308–4314.

116. A. D. Catherman, K. R. Durbin, D. R. Ahlf, B. P. Early, R. T. Fellers and J. C. Tran, *et al.*, Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence, *Mol. Cell. Proteomics*, 2013, **12**(12), 3465–3473.

117. A. L. Capriotti, C. Cavaliere, P. Foglia, R. Samperi and A. Lagana, Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics, *J. Chromatogr. A*, 2011, **1218**(49), 8760–8776.

118. A. D. Catherman, O. S. Skinner and N. L. Kelleher, Top Down proteomics: facts and perspectives, *Biochem. Biophys. Res. Commun.*, 2014, **445**(4), 683–693.

119. J. J. Pesavento, B. A. Garcia, J. A. Streeky, N. L. Kelleher and C. A. Mizzen, Mild performic acid oxidation enhances chromatographic and top down mass spectrometric analyses of histones, *Mol. Cell. Proteomics*, 2007, **6**(9), 1510–1526.

120. J. J. Pesavento, C. R. Bullock, R. D. LeDuc, C. A. Mizzen and N. L. Kelleher, Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry, *J. Biol. Chem.*, 2008, **283**(22), 14927–14937.

121. J. Carroll, M. C. Altman, I. M. Fearnley and J. E. Walker, Identification of membrane proteins by tandem mass spectrometry of protein ions, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(36), 14330–14335.

122. J. Carroll, I. M. Fearnley and J. E. Walker, Definition of the mitochondrial proteome by measurement of molecular masses of membrane proteins, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**(44), 16170–16175.

123. K. M. Millea, I. S. Krull, S. A. Cohen, J. C. Gebler and S. J. Berger, Integration of multidimensional chromatographic protein separations with a combined "top-down" and "bottom-up" proteomic strategy, *J. Proteome Res.*, 2006, **5**(1), 135–146.
124. B. A. Parks, L. Jiang, P. M. Thomas, C. D. Wenger, M. J. Roth and M. T. Boyne 2nd, *et al.*, Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers, *Anal. Chem.*, 2007, **79**(21), 7984–7991.
125. M. J. Roth, B. A. Parks, J. T. Ferguson, M. T. Boyne 2nd and N. L. Kelleher, "Proteotyping": population proteomics of human leukocytes using top down mass spectrometry, *Anal. Chem.*, 2008, **80**(8), 2857–2866.
126. B. E. Chong, F. Yan, D. M. Lubman and F. R. Miller, Chromatofocusing nonporous reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry of proteins from human breast cancer whole cell lysates: a novel two-dimensional liquid chromatography/mass spectrometry method, *Rapid Commun. Mass Spectrom.*, 2001, **15**(4), 291–296.
127. Y. Du, F. Meng, S. M. Patrie, L. M. Miller and N. L. Kelleher, Improved molecular weight-based processing of intact proteins for interrogation by quadrupole-enhanced FT MS/MS, *J. Proteome Res.*, 2004, **3**(4), 801–806.
128. J. C. Tran and A. A. Doucette, Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation, *Anal. Chem.*, 2008, **80**(5), 1568–1573.
129. J. E. Lee, J. F. Kellie, J. C. Tran, J. D. Tipton, A. D. Catherman and H. M. Thomas, *et al.*, A robust two-dimensional separation for top-down tandem mass spectrometry of the low-mass proteome, *J. Am. Soc. Mass Spectrom.*, 2009, **20**(12), 2183–2191.
130. R. Haselberg, G. J. de Jong and G. W. Somsen, Capillary electrophoresis-mass spectrometry for the analysis of intact proteins, *J. Chromatogr. A*, 2007, **1159**(1–2), 81–109.
131. P. Cao and M. Moini, Separation and detection of the alpha- and beta-chains of hemoglobin of a single intact red blood cells using capillary electrophoresis/electrospray ionization time-of-flight mass spectrometry, *J. Am. Soc. Mass Spectrom.*, 1999, **10**(2), 184–186.
132. M. Moini and H. Huang, Application of capillary electrophoresis/electrospray ionization-mass spectrometry to subcellular proteomics of Escherichia coli ribosomal proteins, *Electrophoresis*, 2004, **25**(13), 1981–1987.
133. E. Balaguer and C. Neususs, Glycoprotein characterization combining intact protein and glycan analysis by capillary electrophoresis-electrospray ionization-mass spectrometry, *Anal. Chem.*, 2006, **78**(15), 5384–5393.
134. F. Kilar and S. Hjerten, Fast and high resolution analysis of human serum transferrin by high performance isoelectric focusing in capillaries, *Electrophoresis*, 1989, **10**(1), 23–29.

135. P. K. Jensen, L. Pasa-Tolic, K. K. Peden, S. Martinovic, M. S. Lipton and G. A. Anderson, *et al.*, Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures, *Electrophoresis*, 2000, **21**(7), 1372–1380.
136. L. Yang, C. S. Lee, S. A. Hofstadler, L. Pasa-Tolic and R. D. Smith, Capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for protein characterization, *Anal. Chem.*, 1998, **70**(15), 3235–3241.
137. P. K. Jensen, L. Pasa-Tolic, G. A. Anderson, J. A. Horner, M. S. Lipton and J. E. Bruce, *et al.*, Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, *Anal. Chem.*, 1999, **71**(11), 2076–2084.
138. F. Zhou and M. V. Johnston, Protein characterization by on-line capillary isoelectric focusing, reversed-phase liquid chromatography, and mass spectrometry, *Anal. Chem.*, 2004, **76**(10), 2734–2740.
139. F. Zhou, T. E. Hanson and M. V. Johnston, Intact protein profiling of Chlorobium tepidum by capillary isoelectric focusing, reversed-phase liquid chromatography, and mass spectrometry, *Anal. Chem.*, 2007, **79**(18), 7145–7153.
140. F. W. McLafferty, D. M. Horn, K. Breuker, Y. Ge, M. A. Lewis and B. Cerda, *et al.*, Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance, *J. Am. Soc. Mass Spectrom.*, 2001, **12**(3), 245–249.
141. B. Macek, L. F. Waanders, J. V. Olsen and M. Mann, Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer, *Mol. Cell. Proteomics*, 2006, **5**(5), 949–958.
142. L. F. Waanders, S. Hanke and M. Mann, Top-down quantitation and characterization of SILAC-labeled proteins, *J. Am. Soc. Mass Spectrom.*, 2007, **18**(11), 2058–2064.
143. R. A. Zubarev, N. L. Kelleher and F. W. McLafferty, Electron capture dissociation of multiply charged protein cations. A nonergodic process, *J. Am. Chem. Soc.*, 1998, **120**(13), 3265–3266.
144. J. J. Coon, B. Ueberheide, J. E. Syka, D. D. Dryhurst, J. Ausio and J. Shabanowitz, *et al.*, Protein identification using sequential ion/ion reactions and tandem mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(27), 9463–9468.
145. J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt, Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**(26), 9528–9533.
146. L. Elviri, *ETD and ECD Mass Spectrometry Fragmentation for the Characterization of Protein Post Translational Modifications*, INTECH Open Access Publisher, 2012.
147. H. Molina, R. Matthiesen, K. Kandasamy and A. Pandey, Comprehensive comparison of collision induced dissociation and electron transfer dissociation, *Anal. Chem.*, 2008, **80**(13), 4825–4835.

148. J. Zhang, M. J. Guy, H. S. Norman, Y. C. Chen, Q. Xu and X. Dong, *et al.*, Top-down quantitative proteomics identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure, *J. Proteome Res.*, 2011, **10**(9), 4054–4065.
149. L. M. Smith and N. L. Kelleher, Proteoform: a single term describing protein complexity, *Nat. Methods*, 2013, **10**(3), 186–187.
150. J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman and K. R. Durbin, *et al.*, Mapping intact protein isoforms in discovery mode using top-down proteomics, *Nature*, 2011, **480**(7376), 254–258.
151. A. Resemann, D. Wunderlich, U. Rothbauer, B. Warscheid, H. Leonhardt and J. Fuchser, *et al.*, Top-down *de Novo* protein sequencing of a 13.6 kDa camelid single heavy chain antibody by matrix-assisted laser desorption ionization-time-of-flight/time-of-flight mass spectrometry, *Anal. Chem.*, 2010, **82**(8), 3283–3292.
152. D. M. Horn, R. A. Zubarev and F. W. McLafferty, Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.*, 2000, **97**(19), 10313–10317.
153. B. J. Cargile, S. A. McLuckey and J. L. Stephenson Jr, Identification of bacteriophage MS2 coat protein from E. coli lysates *via* ion trap collisional activation of intact protein ions, *Anal. Chem.*, 2001, **73**(6), 1277–1285.
154. E. Mortz, P. B. O'Connor, P. Roepstorff, N. L. Kelleher, T. D. Wood and F. W. McLafferty, *et al.*, Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**(16), 8264–8267.
155. I. Ntai, K. Kim, R. T. Fellers, O. S. Skinner, A. D. Smith and B. P. Early, *et al.*, Applying label-free quantitation to top down proteomics, *Anal. Chem.*, 2014, **86**(10), 4961–4968.
156. S. Kreimer, M. E. Belov, W. F. Danielson, L. I. Levitsky, M. V. Gorshkov and B. L. Karger, *et al.*, Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling, *J. Proteome Res.*, 2016, **15**(10), 3563–3573.
157. F. Ciregia, L. Kollipara, L. Giusti, R. P. Zahedi, C. Giacomelli and M. R. Mazzoni, *et al.*, Bottom-up proteomics suggests an association between differential expression of mitochondrial proteins and chronic fatigue syndrome, *Transl. Psychiatry*, 2016, **6**(9), e904.
158. J. K. Eng, A. L. McCormack and J. R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom.*, 1994, **5**(11), 976–989.
159. D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, 1999, **20**(18), 3551–3567.
160. J. R. Yates, C. I. Ruse and A. Nakorchevsky, Proteomics by mass spectrometry: approaches, advances, and applications, *Annu. Rev. Biomed. Eng.*, 2009, **11**, 49–79.

161. P. Mallick and B. Kuster, Proteomics: a pragmatic perspective, *Nat. Biotechnol.*, 2010, **28**(7), 695–709.
162. A. Michalski, J. Cox and M. Mann, More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS, *J. Proteome Res.*, 2011, **10**(4), 1785–1793.
163. S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat. Biotechnol.*, 1999, **17**(10), 994–999.
164. M. Schnolzer, P. Jedrzejewski and W. D. Lehmann, Protease-catalyzed incorporation of 18O into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry, *Electrophoresis*, 1996, **17**(5), 945–953.
165. A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell and M. Mueller, *et al.*, An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures, *Mol. Cell. Proteomics*, 2008, **7**(11), 2138–2150.
166. B. Domon and S. Broder, Implications of new proteomics strategies for biology and medicine, *J. Proteome Res.*, 2004, **3**(2), 253–260.
167. J. D. Jaffe, H. Keshishian, B. Chang, T. A. Addona, M. A. Gillette and S. A. Carr, Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification, *Mol. Cell. Proteomics*, 2008, **7**(10), 1952–1962.
168. T. P. Conrads, G. A. Anderson, T. D. Veenstra, L. Pasa-Tolic and R. D. Smith, Utility of accurate mass tags for proteome-wide protein identification, *Anal. Chem.*, 2000, **72**(14), 3349–3354.
169. E. L. Rudomin, S. A. Carr and J. D. Jaffe, Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx), *J. Proteome Res.*, 2009, **8**(6), 3154–3160.
170. N. Wang and L. Li, Exploring the precursor ion exclusion feature of liquid chromatography-electrospray ionization quadrupole time-of-flight mass spectrometry for improving protein identification in shotgun proteome analysis, *Anal. Chem.*, 2008, **80**(12), 4696–4710.
171. S. Hanke, H. Besir, D. Oesterhelt and M. Mann, Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level, *J. Proteome Res.*, 2008, **7**(3), 1118–1130.
172. W. S. Sanders, S. M. Bridges, F. M. McCarthy, B. Nanduri and S. C. Burgess, Prediction of peptides observable by mass spectrometry applied at the experimental set level, *BMC Bioinf.*, 2007, **8**(suppl. 7), S23.
173. E. W. Deutsch, H. Lam and R. Aebersold, PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows, *EMBO Rep.*, 2008, **9**(5), 429–434.
174. F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick and J. Eng, *et al.*, The PeptideAtlas project, *Nucleic Acids Res.*, 2006, **34**(Database issue), D655–D658.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

175. P. Jones, R. G. Cote, S. Y. Cho, S. Klie, L. Martens and A. F. Quinn, *et al.*, PRIDE: new developments and new datasets, *Nucleic Acids Res.*, 2008, **36**(Database issue), D878–D883.
176. R. Beavis and D. Fenyö, Finding protein sequences using PROWL, *Curr. Protoc. Bioinformatics*, 2004, **13**, 2.
177. P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee and D. Martin, *et al.*, Computational prediction of proteotypic peptides for quantitative proteomics, *Nat. Biotechnol.*, 2007, **25**(1), 125–131.
178. V. A. Fusaro, D. R. Mani, J. P. Mesirov and S. A. Carr, Prediction of high-responding peptides for targeted protein assays by mass spectrometry, *Nat. Biotechnol.*, 2009, **27**(2), 190–198.
179. A. Schmidt, M. Beck, J. Malmstrom, H. Lam, M. Claassen and D. Campbell, *et al.*, Absolute quantification of microbial proteomes at different states by directed mass spectrometry, *Mol. Syst. Biol.*, 2011, **7**, 510.
180. J. C. Silva, M. V. Gorenstein, G. Z. Li, J. P. Vissers and S. J. Geromanos, Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition, *Mol. Cell. Proteomics*, 2006, **5**(1), 144–156.
181. J. Malmstrom, M. Beck, A. Schmidt, V. Lange, E. W. Deutsch and R. Aebersold, Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans, *Nature*, 2009, **460**(7256), 762–765.
182. S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner and S. P. Gygi, Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**(12), 6940–6945.
183. J. D. Baty and P. R. Robinson, Single and multiple ion recording techniques for the analysis of diphenylhydantoin and its major metabolite in plasma, *Biomed. Mass Spectrom.*, 1977, **4**(1), 36–41.
184. U. Kusebauch, D. S. Campbell, E. W. Deutsch, C. S. Chu, D. A. Spicer and M. Y. Brusniak, *et al.*, Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome, *Cell.*, 2016, **166**(3), 766–778.
185. R. Huttenhain, S. Surinova, R. Ossola, Z. Sun, D. Campbell and F. Cerciello, *et al.*, *N*-glycoprotein SRMAtlas: a resource of mass spectrometric assays for *N*-glycosites enabling consistent and multiplexed protein quantification for clinical applications, *Mol. Cell. Proteomics*, 2013, **12**(4), 1005–1016.
186. O. T. Schubert, J. Mouritsen, C. Ludwig, H. L. Rost, G. Rosenberger and P. K. Arthur, *et al.*, The Mtb proteome library: a resource of assays to quantify the complete proteome of Mycobacterium tuberculosis, *Cell Host Microbe*, 2013, **13**(5), 602–612.
187. B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney and B. Frewen, *et al.*, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, *Bioinformatics*, 2010, **26**(7), 966–968.
188. V. Sharma, J. Eckels, G. K. Taylor, N. J. Shulman, A. B. Stergachis and S. A. Joyner, *et al.*, Panorama: a targeted proteomics knowledge base, *J. Proteome Res.*, 2014, **13**(9), 4205–4210.

189. A. C. Peterson, J. D. Russell, D. J. Bailey, M. S. Westphall and J. J. Coon, Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics, *Mol. Cell. Proteomics*, 2012, **11**(11), 1475–1488.
190. M. Unlu, M. E. Morgan and J. S. Minden, Difference gel electrophoresis: a single gel method for detecting changes in protein extracts, *Electrophoresis*, 1997, **18**(11), 2071–2077.
191. R. Tonge, J. Shaw, B. Middleton, R. Rowlinson, S. Rayner and J. Young, *et al.*, Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology, *Proteomics*, 2001, **1**(3), 377–396.
192. M. E. Jung, W. J. Kim, N. K. Avliyakulov, M. Oztug and M. J. Haykinson, Synthesis and validation of cyanine-based dyes for DIGE, *Methods Mol. Biol.*, 2012, **854**, 67–85.
193. T. L. Wu, Two-dimensional difference gel electrophoresis, *Methods Mol. Biol.*, 2006, **328**, 71–95.
194. T. Hrebicek, K. Durrschmid, N. Auer, K. Bayer and A. Rizzi, Effect of CyDye minimum labeling in differential gel electrophoresis on the reliability of protein identification, *Electrophoresis*, 2007, **28**(7), 1161–1169.
195. G. Van den Bergh and L. Arckens, Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics, *Curr. Opin. Biotechnol.*, 2004, **15**, 38–43.
196. B. Chakravarti, S. R. Gallagher and D. N. Chakravarti, Difference gel electrophoresis (DIGE) using CyDye DIGE fluor minimal dyes, *Curr. Protoc. Mol. Biol.*, 2005, **10**, 23.
197. B. Sitek, J. Luttges, K. Marcus, G. Kloppel, W. Schmiegel and H. E. Meyer, *et al.*, Application of fluorescence difference gel electrophoresis saturation labelling for the analysis of microdissected precursor lesions of pancreatic ductal adenocarcinoma, *Proteomics*, 2005, **5**(10), 2665–2679.
198. C. May, F. Brosseron, P. Chartowski, H. E. Meyer and K. Marcus, Differential proteome analysis using 2D-DIGE, *Methods Mol. Biol.*, 2012, **893**, 75–82.
199. N. S. Tannu and S. E. Hemby, Two-dimensional fluorescence difference gel electrophoresis for comparative proteomics profiling, *Nat. Protoc.*, 2006, **1**(4), 1732–1742.
200. A. Alban, S. O. David, L. Bjorkesten, C. Andersson, E. Sloge and S. Lewis, *et al.*, A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard, *Proteomics*, 2003, **3**(1), 36–44.
201. K. Engelen, A. Sifrim, B. Van de Plas, K. Laukens, L. Arckens and K. Marchal, Alternative experimental design with an applied normalization scheme can improve statistical power in 2D-DIGE experiments, *J. Proteome Res.*, 2010, **9**(10), 4919–4926.
202. Y. Oda, K. Huang, F. R. Cross, D. Cowburn and B. T. Chait, Accurate quantitation of protein expression and site-specific phosphorylation, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**(12), 6591–6596.

203. M. P. Washburn, R. Ulaszek, C. Deciu, D. M. Schieltz and J. R. Yates 3rd, Analysis of quantitative proteomic data generated *via* multidimensional protein identification technology, *Anal. Chem.*, 2002, **74**(7), 1650–1657.

204. B. Blagoev, S. E. Ong, I. Kratchmarova and M. Mann, Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics, *Nat. Biotechnol.*, 2004, **22**(9), 1139–1145.

205. J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar and P. Mortensen, *et al.*, Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks, *Cell.*, 2006, **127**(3), 635–648.

206. T. Geiger, J. Cox, P. Ostasiewicz, J. R. Wisniewski and M. Mann, Super-SILAC mix for quantitative proteomics of human tumor tissue, *Nat. Methods*, 2010, **7**(5), 383–385.

207. I. I. Stewart, T. Thomson and D. Figeys, 18O labeling: a tool for proteomics, *Rapid Commun. Mass Spectrom.*, 2001, **15**(24), 2456–2465.

208. M. Miyagi and K. C. Rao, Proteolytic 18O-labeling strategies for quantitative proteomics, *Mass Spectrom. Rev.*, 2007, **26**(1), 121–136.

209. O. A. Mirgorodskaya, Y. P. Kozmin, M. I. Titov, R. Korner, C. P. Sonksen and P. Roepstorff, Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards, *Rapid Commun. Mass Spectrom.*, 2000, **14**(14), 1226–1232.

210. X. Yao, A. Freas, J. Ramirez, P. A. Demirev and C. Fenselau, Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus, *Anal. Chem.*, 2001, **73**(13), 2836–2842.

211. M. Heller, H. Mattou, C. Menzel and X. Yao, Trypsin catalyzed 16O-to-18O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-QTOF, and ESI-ion trap mass spectrometers, *J. Am. Soc. Mass Spectrom.*, 2003, **14**(7), 704–718.

212. X. Ye, B. Luke, T. Andresson and J. Blonder, 18O stable isotope labeling in MS-based proteomics, *Briefings Funct. Genomics Proteomics*, 2009, **8**(2), 136–144.

213. M. J. Castillo, K. J. Reynolds, A. Gomes, C. Fenselau and X. Yao, Quantitative protein analysis using enzymatic [(1)(8)O]water labeling, *Curr. Protoc. Protein Sci.*, 2014, **76**, 23.4.1–23.4.9.

214. K. Bezstarosti, A. Ghamari, F. G. Grosveld and J. A. Demmers, Differential proteomics based on 18O labeling to determine the cyclin dependent kinase 9 interactome, *J. Proteome Res.*, 2010, **9**(9), 4464–4475.

215. X. Chen, S. W. Cushman, L. K. Pannell and S. Hess, Quantitative proteomic analysis of the secretory proteins from rat adipose cells using a 2D liquid chromatography-MS/MS approach, *J. Proteome Res.*, 2005, **4**(2), 570–577.

216. A. Ramos-Fernandez, D. Lopez-Ferrer and J. Vazquez, Improved method for differential expression proteomics using trypsin-catalyzed 18O labeling with a correction for labeling efficiency, *Mol. Cell. Proteomics*, 2007, **6**(7), 1274–1286.

217. R. Zhang, C. S. Sioma, S. Wang and F. E. Regnier, Fractionation of isotopically labeled peptides in quantitative proteomics, *Anal. Chem.*, 2001, **73**(21), 5142–5149.
218. K. C. Hansen, G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin and A. L. Burlingame, Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography, *Mol. Cell. Proteomics*, 2003, **2**(5), 299–314.
219. D. K. Han, J. Eng, H. Zhou and R. Aebersold, Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry, *Nat. Biotechnol.*, 2001, **19**(10), 946–951.
220. X. Zhang, B. Huang, X. Zhou and C. Chen, Quantitative proteomic analysis of S-nitrosated proteins in diabetic mouse liver with ICAT switch method, *Protein Cell*, 2010, **1**(7), 675–687.
221. G. Chiappetta, S. Ndiaye, A. Igbaria, C. Kumar, J. Vinh and M. B. Toledano, Proteome screens for Cys residues oxidation: the redoxome, *Methods Enzymol.*, 2010, **473**, 199–216.
222. N. Brandes, D. Reichmann, H. Tienson, L. I. Leichert and U. Jakob, Using quantitative redox proteomics to dissect the yeast redoxome, *J. Biol. Chem.*, 2011, **286**(48), 41893–41903.
223. L. I. Leichert, F. Gehrke, H. V. Gudiseva, T. Blackwell, M. Ilbert and A. K. Walker, *et al.*, Quantifying changes in the thiol redox proteome upon oxidative stress *in vivo*, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**(24), 8197–8202.
224. S. Garcia-Santamarina, S. Boronat, A. Domenech, J. Ayte, H. Molina and E. Hidalgo, Monitoring *in vivo* reversible cysteine oxidation in proteins using ICAT and mass spectrometry, *Nat. Protoc.*, 2014, **9**(5), 1131–1145.
225. A. Schmidt, J. Kellermann and F. Lottspeich, A novel strategy for quantitative proteomics using isotope-coded protein labels, *Proteomics*, 2005, **5**(1), 4–15.
226. G. Maccarrone, M. Lebar and D. Martins-de-Souza, Brain quantitative proteomics combining GeLC-MS and isotope-coded protein labeling (ICPL), *Methods Mol. Biol.*, 2014, **1156**, 175–185.
227. F. Lottspeich and J. Kellermann, ICPL labeling strategies for proteome research, *Methods Mol. Biol.*, 2011, **753**, 55–64.
228. A. Turtoi, G. D. Mazzucchelli and E. De Pauw, Isotope coded protein label quantification of serum proteins–comparison with the label-free LC-MS and validation using the MRM approach, *Talanta*, 2010, **80**(4), 1487–1495.
229. G. Cagney and A. Emili, De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging, *Nat. Biotechnol.*, 2002, **20**(2), 163–170.
230. J. L. Hsu, S. Y. Huang, N. H. Chow and S. H. Chen, Stable-isotope dimethyl labeling for quantitative proteomics, *Anal. Chem.*, 2003, **75**(24), 6843–6852.

231. P. J. Boersema, T. T. Aye, T. A. van Veen, A. J. Heck and S. Mohammed, Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates, *Proteomics*, 2008, **8**(22), 4624–4632.

232. M. P. Hall, S. Ashrafi, I. Obegi, R. Petesch, J. N. Peterson and L. V. Schneider, Mass defect" tags for biomolecular mass spectrometry, *J. Mass Spectrom.*, 2003, **38**(8), 809–816.

233. M. P. Hall and L. V. Schneider, Isotope-differentiated binding energy shift tags (IDBEST) for improved targeted biomarker discovery and validation, *Expert Rev. Proteomics*, 2004, **1**(4), 421–431.

234. P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker and S. Hattan, *et al.*, Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents, *Mol. Cell. Proteomics*, 2004, **3**(12), 1154–1169.

235. A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz and G. Schmidt, *et al.*, Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS, *Anal. Chem.*, 2003, **75**(8), 1895–1904.

236. M. Schirle, E. C. Petrella, S. M. Brittain, D. Schwalb, E. Harrington and I. Cornella-Taracido, *et al.*, Kinase inhibitor profiling using chemoproteomics, *Methods Mol. Biol.*, 2012, **795**, 161–177.

237. R. I. Viner, T. Zhang, T. Second and V. Zabrouskov, Quantification of post-translationally modified peptides of bovine alpha-crystallin using tandem mass tags and electron transfer dissociation, *J. Proteomics*, 2009, **72**(5), 874–885.

238. L. Dayon, A. Hainard, V. Licker, N. Turck, K. Kuhn and D. F. Hochstrasser, *et al.*, Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags, *Anal. Chem.*, 2008, **80**(8), 2921–2931.

239. L. Ting, R. Rad, S. P. Gygi and W. Haas, MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics, *Nat. Methods*, 2011, **8**(11), 937–940.

240. C. J. Koehler, M. Strozynski, F. Kozielski, A. Treumann and B. Thiede, Isobaric peptide termini labeling for MS/MS-based quantitative proteomics, *J. Proteome Res.*, 2009, **8**(9), 4333–4341.

241. M. O. Arntzen, C. J. Koehler, A. Treumann and B. Thiede, Quantitative proteome analysis using isobaric peptide termini labeling (IPTL), *Methods Mol. Biol.*, 2011, **753**, 65–76.

242. C. J. Koehler, M. O. Arntzen, M. Strozynski, A. Treumann and B. Thiede, Isobaric peptide termini labeling utilizing site-specific N-terminal succinylation, *Anal. Chem.*, 2011, **83**(12), 4775–4781.

243. C. J. Koehler, M. O. Arntzen, G. A. de Souza and B. Thiede, An approach for triplex-isobaric peptide termini labeling (triplex-IPTL), *Anal. Chem.*, 2013, **85**(4), 2478–2485.

244. C. Vogel and E. M. Marcotte, Label-free protein quantitation using weighted spectral counting, *Methods Mol. Biol.*, 2012, **893**, 321–341.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

245. H. Liu, R. G. Sadygov and J. R. Yates 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics, *Anal. Chem.*, 2004, **76**(14), 4193–4201.
246. J. M. Asara, H. R. Christofk, L. M. Freimark and L. C. Cantley, A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen, *Proteomics*, 2008, **8**(5), 994–999.
247. D. H. Lundgren, S. I. Hwang, L. Wu and D. K. Han, Role of spectral counting in quantitative proteomics, *Expert Rev. Proteomics*, 2010, **7**(1), 39–53.
248. J. Rappsilber, U. Ryder, A. I. Lamond and M. Mann, Large-scale proteomic analysis of the human spliceosome, *Genome Res.*, 2002, **12**(8), 1231–1245.
249. Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu and J. Rappsilber, *et al.*, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein, *Mol. Cell. Proteomics*, 2005, **4**(9), 1265–1272.
250. P. Lu, C. Vogel, R. Wang, X. Yao and E. M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation, *Nat. Biotechnol.*, 2007, **25**(1), 117–124.
251. J. C. Braisted, S. Kuntumalla, C. Vogel, E. M. Marcotte, A. R. Rodrigues and R. Wang, *et al.*, The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results, *BMC Bioinf.*, 2008, **9**, 529.
252. A. Sun, J. Zhang, C. Wang, D. Yang, H. Wei and Y. Zhu, *et al.*, Modified spectral count index (mSCI) for estimation of protein abundance by protein relative identification possibility (RIPpro): a new proteomic technological parameter, *J. Proteome Res.*, 2009, **8**(11), 4934–4942.
253. D. W. Powell, C. M. Weaver, J. L. Jennings, K. J. McAfee, Y. He and P. A. Weil, *et al.*, Cluster analysis of mass spectrometry data reveals a novel component of SAGA, *Mol. Cell. Biol.*, 2004, **24**(16), 7249–7259.
254. N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore and Y. Li, *et al.*, Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis, *Nat. Biotechnol.*, 2010, **28**(1), 83–89.
255. N. Colaert, K. Gevaert and L. Martens, RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification, *J. Proteome Res.*, 2011, **10**(7), 3183–3189.
256. M. Bantscheff, S. Lemeer, M. M. Savitski and B. Kuster, Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present, *Anal. Bioanal. Chem.*, 2012, **404**(4), 939–965.
257. P. V. Bondarenko, D. Chelius and T. A. Shaler, Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry, *Anal. Chem.*, 2002, **74**(18), 4741–4749.

258. D. Chelius and P. V. Bondarenko, Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry, *J. Proteome Res.*, 2002, **1**(4), 317–323.
259. L. N. Mueller, M. Y. Brusniak, D. R. Mani and R. Aebersold, An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data, *J. Proteome Res.*, 2008, **7**(1), 51–61.
260. J. Listgarten and A. Emili, Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry, *Mol. Cell. Proteomics*, 2005, **4**(4), 419–434.
261. J. C. Silva, R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass and G. Z. Li, *et al.*, Quantitative proteomic analysis by accurate mass retention time pairs, *Anal. Chem.*, 2005, **77**(7), 2187–2200.
262. J. C. Silva, R. Denny, C. Dorschel, M. V. Gorenstein, G. Z. Li and K. Richardson, *et al.*, Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale, *Mol. Cell. Proteomics*, 2006, **5**(4), 589–607.
263. S. J. Geromanos, J. P. Vissers, J. C. Silva, C. A. Dorschel, G. Z. Li and M. V. Gorenstein, *et al.*, The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS, *Proteomics*, 2009, **9**(6), 1683–1695.
264. B. Fabre, T. Lambour, D. Bouyssié, T. Menneteau, B. Monsarrat and O. Burlet-Schiltz, *et al.*, Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry, *EuPa Open Proteomics*, 2014, **4**, 82–86.
265. B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt and J. Wolf, *et al.*, Global quantification of mammalian gene expression control, *Nature*, 2011, **473**(7347), 337–342.
266. L. Arike, K. Valgepea, L. Peil, R. Nahku, K. Adamberg and R. Vilu, Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli, *J. Proteomics*, 2012, **75**(17), 5437–5448.
267. S. Purvine, J. T. Eppel, E. C. Yi and D. R. Goodlett, Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer, *Proteomics*, 2003, **3**(6), 847–850.
268. J. D. Venable, M. Q. Dong, J. Wohlschlegel, A. Dillin and J. R. Yates, Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra, *Nat. Methods*, 2004, **1**(1), 39–45.
269. R. S. Plumb, K. A. Johnson, P. Rainville, B. W. Smith, I. D. Wilson and J. M. Castro-Perez, *et al.*, UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation, *Rapid Commun. Mass Spectrom.*, 2006, **20**(13), 1989–1994.
270. A. Panchaud, A. Scherl, S. A. Shaffer, P. D. von Haller, H. D. Kulasekara and S. I. Miller, *et al.*, Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean, *Anal. Chem.*, 2009, **81**(15), 6481–6488.
271. T. Geiger, J. Cox and M. Mann, Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation, *Mol. Cell. Proteomics*, 2010, **9**(10), 2252–2261.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

272. M. Bern, G. Finney, M. R. Hoopmann, G. Merrihew, M. J. Toth and M. J. MacCoss, Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry, *Anal. Chem.*, 2010, **82**(3), 833–841.
273. P. C. Carvalho, X. Han, T. Xu, D. Cociorva, G. Carvalho Mda and V. C. Barbosa, *et al.*, XDIA: improving on the label-free data-independent analysis, *Bioinformatics*, 2010, **26**(6), 847–848.
274. A. Panchaud, S. Jung, S. A. Shaffer, J. D. Aitchison and D. R. Goodlett, Faster, quantitative, and accurate precursor acquisition independent from ion count, *Anal. Chem.*, 2011, **83**(6), 2250–2257.
275. J. W. Wong, A. B. Schwahn and K. M. Downard, ETISEQ–an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics, *BMC Bioinf.*, 2009, **10**, 244.
276. G. Z. Li, J. P. Vissers, J. C. Silva, D. Golick, M. V. Gorenstein and S. J. Geromanos, Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures, *Proteomics*, 2009, **9**(6), 1696–1719.
277. K. Blackburn, F. Mbeunkui, S. K. Mitra, T. Mentzel and M. B. Goshe, Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation, *J. Proteome Res.*, 2010, **9**(7), 3621–3637.
278. L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek and L. Reiter, *et al.*, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, *Mol. Cell. Proteomics*, 2012, **11**(6), O111 016717.
279. S. Sidoli, S. Lin, L. Xiong, N. V. Bhanu, K. R. Karch and E. Johansen, *et al.*, Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) Analysis for Characterization and Quantification of Histone Post-translational Modifications, *Mol. Cell. Proteomics*, 2015, **14**(9), 2420–2428.
280. Q. Huang, L. Yang, J. Luo, L. Guo, Z. Wang and X. Yang, *et al.*, SWATH enables precise label-free quantification on proteome scale, *Proteomics*, 2015, **15**(7), 1215–1223.
281. V. Lange, P. Picotti, B. Domon and R. Aebersold, Selected reaction monitoring for quantitative proteomics: a tutorial, *Mol. Syst. Biol.*, 2008, **4**, 222.
282. W. J. Qian, J. M. Jacobs, T. Liu, D. G. Camp 2nd and R. D. Smith, Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications, *Mol. Cell. Proteomics*, 2006, **5**(10), 1727–1744.
283. A. Wolf-Yadlin, S. Hautaniemi, D. A. Lauffenburger and F. M. White, Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(14), 5860–5865.

284. L. Anderson and C. L. Hunter, Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins, *Mol. Cell. Proteomics*, 2006, **5**(4), 573–588.

285. M. H. Elliott, D. S. Smith, C. E. Parker and C. Borchers, Current trends in quantitative proteomics, *J. Mass Spectrom.*, 2009, **44**(12), 1637–1660.

286. E. Kuhn, J. Wu, J. Karl, H. Liao, W. Zolg and B. Guild, Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards, *Proteomics*, 2004, **4**(4), 1175–1186.

287. C. Huillet, A. Adrait, D. Lebert, G. Picard, M. Trauchessec and M. Louwagie, *et al.*, Accurate quantification of cardiovascular biomarkers in serum using Protein Standard Absolute Quantification (PSAQ) and selected reaction monitoring, *Mol. Cell. Proteomics*, 2012, **11**(2), M111 008235.

288. D. S. Kirkpatrick, S. A. Gerber and S. P. Gygi, The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications, *Methods*, 2005, **35**(3), 265–273.

289. A. Prakash, D. M. Tomazela, B. Frewen, B. Maclean, G. Merrihew and S. Peterman, *et al.*, Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development, *J. Proteome Res.*, 2009, **8**(6), 2733–2739.

290. T. Basak, S. Varshney, S. Akhtar and S. Sengupta, Understanding different facets of cardiovascular diseases based on model systems to human studies: a proteomic and metabolomic perspective, *J. Proteomics*, 2015, **127**(Pt A), 50–60.

291. T. Basak, V. S. Tanwar, G. Bhardwaj, N. Bhardwaj, S. Ahmad and G. Garg, *et al.*, Plasma proteomic analysis of stable coronary artery disease indicates impairment of reverse cholesterol pathway, *Sci. Rep.*, 2016, **6**, 28042.

292. L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu and D. M. Maynard, *et al.*, Open mass spectrometry search algorithm, *J. Proteome Res.*, 2004, **3**(5), 958–964.

293. R. Craig and R. C. Beavis, TANDEM: matching proteins with tandem mass spectra, *Bioinformatics*, 2004, **20**(9), 1466–1467.

294. J. Colinge, A. Masselot, M. Giron, T. Dessingy and J. Magnin, OLAV: towards high-throughput tandem mass spectrometry data identification, *Proteomics*, 2003, **3**(8), 1454–1463.

295. R. Matthiesen, M. B. Trelle, P. Hojrup, J. Bunkenborg and O. N. Jensen, VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins, *J. Proteome Res.*, 2005, **4**(6), 2338–2347.

296. D. L. Tabb, C. G. Fernando and M. C. Chambers, MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis, *J. Proteome Res.*, 2007, **6**(2), 654–661.

297. A. K. Yadav, D. Kumar and D. Dash, MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry, *J. Proteome Res.*, 2011, **10**(5), 2154–2160.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00082

298. J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie and D. Yuen, *et al.*, PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification, *Mol. Cell. Proteomics*, 2012, **11**(4), M111 010587.
299. N. Zhang, R. Aebersold and B. Schwikowski, ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data, *Proteomics*, 2002, **2**(10), 1406–1412.
300. K. R. Clauser, P. Baker and A. L. Burlingame, Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching, *Anal. Chem.*, 1999, **71**(14), 2871–2882.
301. A. Frank and P. Pevzner, PepNovo: *de novo* peptide sequencing *via* probabilistic network modeling, *Anal. Chem.*, 2005, **77**(4), 964–973.
302. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li and A. Doherty-Kirby, *et al.*, PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry, *Rapid Commun. Mass Spectrom.*, 2003, **17**(20), 2337–2342.
303. A. Frank, S. Tanner, V. Bafna and P. Pevzner, Peptide sequence tags for fast database search in mass-spectrometry, *J. Proteome Res.*, 2005, **4**(4), 1287–1295.
304. R. S. Johnson and J. A. Taylor, Searching sequence databases *via* de novo peptide sequencing by tandem mass spectrometry, *Mol. Biotechnol.*, 2002, **22**(3), 301–315.
305. L. Mo, D. Dutta, Y. Wan and T. Chen, MSNovo: a dynamic programming algorithm for *de novo* peptide sequencing *via* tandem mass spectrometry, *Anal. Chem.*, 2007, **79**(13), 4870–4878.
306. A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Anal. Chem.*, 2002, **74**(20), 5383–5392.
307. A. I. Nesvizhskii, A. Keller, E. Kolker and R. Aebersold, A statistical model for identifying proteins by tandem mass spectrometry, *Anal. Chem.*, 2003, **75**(17), 4646–4658.
308. B. C. Searle, Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies, *Proteomics*, 2010, **10**(6), 1265–1269.
309. T. Muth, M. Vaudel, H. Barsnes, L. Martens and A. Sickmann, XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results, *Proteomics*, 2010, **10**(7), 1522–1524.
310. F. Desiere, E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King and J. K. Eng, *et al.*, Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry, *Genome Biol.*, 2005, **6**(1), R9.
311. D. Kessner, M. Chambers, R. Burke, D. Agus and P. Mallick, ProteoWizard: open source software for rapid proteomics tools development, *Bioinformatics*, 2008, **24**(21), 2534–2536.
312. P. Jones and R. Cote, The PRIDE proteomics identifications database: data submission, query, and dataset comparison, *Methods Mol. Biol.*, 2008, **484**, 287–303.

313. D. C. Trudgian, B. Thomas, S. J. McGowan, B. M. Kessler, M. Salek and O. Acuto, CPFP: a central proteomics facilities pipeline, *Bioinformatics*, 2010, **26**(8), 1131–1132.
314. A. Rauch, M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman and P. Hussey, *et al.*, Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments, *J. Proteome Res.*, 2006, **5**(1), 112–121.
315. P. Mortensen, J. W. Gouw, J. V. Olsen, S. E. Ong, K. T. Rigbolt and J. Bunkenborg, *et al.*, MSQuant, an open source platform for mass spectrometry-based quantitative proteomics, *J. Proteome Res.*, 2010, **9**(1), 393–403.
316. R. D. Appel, J. R. Vargas, P. M. Palagi, D. Walther and D. F. Hochstrasser, Melanie II–a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms, *Electrophoresis*, 1997, **18**(15), 2735–2748.
317. M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph and P. Wang, *et al.*, A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS, *Bioinformatics*, 2006, **22**(15), 1902–1909.
318. C. Johansson, J. Samskog, L. Sundstrom, H. Wadensten, L. Bjorkesten and J. Flensburg, Differential expression analysis of Escherichia coli proteins using a novel software for relative quantitation of LC-MS/MS data, *Proteomics*, 2006, **6**(16), 4475–4485.
319. W. X. Schulze and M. Mann, A novel proteomic screen for peptide-protein interactions, *J. Biol. Chem.*, 2004, **279**(11), 10756–10764.
320. A. Saito, M. Nagasaki, M. Oyama, H. Kozuka-Hata, K. Semba and S. Sugano, *et al.*, AYUMS: an algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction, *BMC Bioinf.*, 2007, **8**, 15.
321. J. Cox and M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.*, 2008, **26**(12), 1367–1372.
322. J. Cox, I. Matic, M. Hilger, N. Nagaraj, M. Selbach and J. V. Olsen, *et al.*, A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics, *Nat. Protoc.*, 2009, **4**(5), 698–705.
323. X. J. Li, H. Zhang, J. A. Ranish and R. Aebersold, Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry, *Anal. Chem.*, 2003, **75**(23), 6648–6657.
324. S. K. Park, J. D. Venable, T. Xu and J. R. Yates 3rd, A quantitative analysis software tool for mass spectrometry-based proteomics, *Nat. Methods*, 2008, **5**(4), 319–322.
325. B. D. Halligan, R. Y. Slyper, S. N. Twigger, W. Hicks, M. Olivier and A. S. Greene, ZoomQuant: an application for the quantitation of stable isotope labeled peptides, *J. Am. Soc. Mass Spectrom.*, 2005, **16**(3), 302–306.

326. G. Wang, W. W. Wu, T. Pisitkun, J. D. Hoffert, M. A. Knepper and R. F. Shen, Automated quantification tool for high-throughput proteomics using stable isotope labeling and LC-MSn, *Anal. Chem.*, 2006, **78**(16), 5752–5761.

327. I. P. Shadforth, T. P. Dunkley, K. S. Lilley and C. Bessant, i-Tracker: for quantitative proteomics using iTRAQ, *BMC Genomics*, 2005, **6**, 145.

328. T. Muth, D. Keller, S. M. Puetz, L. Martens, A. Sickmann and A. M. Boehm, jTraqX: a free, platform independent tool for isobaric tag quantitation at the protein level, *Proteomics*, 2010, **10**(6), 1223–1225.

329. M. O. Arntzen, C. J. Koehler, H. Barsnes, F. S. Berven, A. Treumann and B. Thiede, IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT, *J. Proteome Res.*, 2011, **10**(2), 913–920.

330. N. L. Heinecke, B. S. Pratt, T. Vaisar and L. Becker, PepC: proteomics software for identifying differentially expressed proteins based on spectral counting, *Bioinformatics*, 2010, **26**(12), 1574–1575.

331. P. M. Palagi, D. Walther, M. Quadroni, S. Catherinet, J. Burgess and C. G. Zimmermann-Ivol, *et al.*, MSight: an image analysis software for liquid chromatography-mass spectrometry, *Proteomics*, 2005, **5**(9), 2381–2384.

332. J. D. Jaffe, D. R. Mani, K. C. Leptos, G. M. Church, M. A. Gillette and S. A. Carr, PEPPeR, a platform for experimental proteomic pattern recognition, *Mol. Cell. Proteomics*, 2006, **5**(10), 1927–1941.

333. L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller and M. Y. Brusniak, *et al.*, SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling, *Proteomics*, 2007, **7**(19), 3470–3480.

334. M. Y. Brusniak, B. Bodenmiller, D. Campbell, K. Cooke, J. Eddes and A. Garbutt, *et al.*, Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics, *BMC Bioinf.*, 2008, **9**, 542.

335. J. S. Andersen, C. J. Wilkinson, T. Mayor, P. Mortensen, E. A. Nigg and M. Mann, Proteomic characterization of the human centrosome by protein correlation profiling, *Nature*, 2003, **426**(6966), 570–574.

336. X. J. Li, E. C. Yi, C. J. Kemp, H. Zhang and R. Aebersold, A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry, *Mol. Cell. Proteomics*, 2005, **4**(9), 1328–1340.

337. D. B. Martin, T. Holzman, D. May, A. Peterson, A. Eastham and J. Eng, *et al.*, MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments, *Mol. Cell. Proteomics*, 2008, **7**(11), 2270–2278.

338. M. Y. Brusniak, S. T. Kwok, M. Christiansen, D. Campbell, L. Reiter and P. Picotti, *et al.*, ATAQS: A computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry, *BMC Bioinf.*, 2011, **12**, 78.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

CHAPTER 5

Mass Spectrometry for Discovering Natural Products

PAULO C. VIEIRA*, ANA CAROLINA A. SANTOS AND TAYNARA L. SILVA

Department of Chemistry, Federal University of São Carlos, Sao Carlos, SP, 13565-905, Brazil
*E-mail: dpcv@ufscar.br

5.1 Introduction

Due to the wide range of applications in several fields, natural products continue to attract scientists' efforts, mainly in medicinal chemistry for the discovery of new therapeutic agents. Plants, microorganisms and small animals are being used as principal sources in the discovery of novel drug candidates[1–5] and mass spectrometry (MS) has played a central role in making the identification and discovery easier and faster since the 1960s.[6] Initially, under high-voltage and high-vacuum conditions, Djerassi, Biemann, Williams and Budzikiewicz performed fragmentation studies of several natural products classes enabling the rapid identification of natural products from complex mixtures through gas chromatography (GC) MS (GC-MS) analysis.[6,7] For example, in 1967 Agurell and coworkers[8] performed analysis of extracts from *Banisteriopsis rusbyana* leaves, used in the preparation of the hallucinatory drink *ayahusca* from Amazonia, aiming to characterize its alkaloid content. The analysis revealed *N*,*N*-dimethyltryptamine as the major substance and *N*-methyl-tryptamine, 5-methoxy-*N*-*N*-dimethyltryptamine and

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

5-hydroxy-*N*,*N*-dimethyltryptamine as minor substances.[8] Sterol fragmentation pattern and retention time were also deeply studied in an attempt to make their identification in complex extracts easier.[9] 19-norcholesterol and its homologs were identified in the extracts of marine animals in an attempt at some progress in the biosynthesis of gorgosterol.[10] With the advance of many GC-MS libraries such as that for mass spectroscopy of recoiled ions (MSRI), the Automated Mass Spectral Deconvolution and Identification System (AMDIS), National Institute of Standards and Technology (NIST), and Wiley, GC-MS technique became widely used to characterize metabolites from complex matrixes of plants and microorganisms. GC-MS based metabolome analysis has many applications in biotechnological fields such as the discovery of bioactive compounds[11] and drugs candidates,[12] characterization of volatile compounds present in essential oils,[13] understanding of ecological processes[14] and the effects of alteration on the genomics of metabolomics,[15] *etc*.

However, high voltage, volatility, polarity and thermal stability restricted the applicability of the GC-MS technique to natural product identification as the majority of natural products consist of medium- to big-sized and very functionalized substances. Therefore, the development of chemical ionization (CI), fast atom bombardment (FAB), matrix assisted-laser desorption/ionization (MALDI) and electrospray ionization (ESI) methods was essential to making possible the analysis and identification of medium to highly polar compounds. Alkaloids, for example, were not effectively analyzed by electron ionization (EI) MS, probably due to the fact that these compounds are usually cyclic amines, often containing aromatic rings, suffering a beta-cleavage resulting in complete loss of the molecular ion. Due to the basicity of most alkaloids and the consequent stability of their protonated forms, the formation of quasimolecular ions through chemical ionization (CI) is facilitated.[16] Milne and co-workers performed a series of studies involving different classes of substances using CI, showing the applicability of this CI ionization technique.[16–18]

Pioneered by Barber and his colleagues, FAB allowed analyzing compounds with high polarity and molecular weights, operating in both positive and negative ion modes.[19] The applicability of this technique was principally for underivatized peptide and protein analysis in the range of 6 kDa to 20 kDa molecular weight. A liquid sample containing peptides and proteins is introduced into the MS in the condensed phase and the analytes are desorbed and ionized by bombardment with ions (Cs^+) or atoms (Ar, Xe) with energies between 5 and 30 keV. The compatibility with magnetic deflect ion mass analyzers and quadrupole instruments made this technique widely used.[20]

The desorption ionization methods included some variants that also used a matrix to make the ionization process easier and prevent the direct interaction of the analytes with the energy source. Among these variations, MALDI is the most important and is used in MS.[7] MALDI allows the ionization of large molecules like proteins and peptides. The development of spray ionization and desorption methods occurred in parallel. The spray ionization methods arose as a solution to the interfacing of liquid chromatography (LC) with MS

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

and culminated with the highly successful ESI method. The development of the spray ionization methods occurred concomitantly with the development of desorption methods. ESI allows production of ions from solutions and ionizing polar, non-volatile and non-thermally stable molecules. With the advances in both the MALDI and ESI methods and the perfect coupling to high-resolution mass analyzers, MALDI and ESI have been very widely applied as powerful tools in natural product analysis and are the ubiquitous ionization techniques of today.[7]

In parallel, the application of LC–MS in natural product analysis has grown dramatically over the last decade and has become essential for promoting the correct identification of components in complex matrices. Due to their vast applicability to a wide range of molecules, various methodologies using ESI and MALDI as an ionization source have been developed to promote fast analysis and identification. Dereplication and imaging MS methodologies were developed as a solution to this need.

5.2 GC-MS

It is well known that volatile organic compounds (VOCs) play a major role in microorganisms interactions[21] and are considered to be excellent infochemicals.[22] Essential oils (EOs) are aromatic oily liquids obtained from plants through different extraction processes, such as steam distillation, hydrodistillation, simultaneous distillation–extraction, supercritical carbon dioxide extraction, solvent extraction, *etc.*[23] They are complex mixtures of volatile organic compounds and the major classes are monoterpenes, sesquiterpenes and phenolic compounds. These oils are known to possess antimicrobial activity,[24] being applied as food additives,[25] antioxidants,[26] and antihelmintics,[27] and can also be used as repellents against mosquitoes.[28] Historically, GC-MS is a key technique for the separation and identification of low molecular weight natural products in complex mixtures, being suitable for the characterization of the chemical composition of EOs as its analyses provides valuable information.[29] In addition, the good chromatographic separation and resolution, and the highly reproducible MS data provides spectral fingerprinting that can be compared to different databases, as the ionization energy is usually 70 eV. Based on this, the technique is commonly used for VOC analysis.[30]

Various examples can be found in the literature (see Figure 5.1), such as the GC-MS analysis of clove leaves, *Syzygium aromaticum*, which led to the identification of 16 volatile compounds, eugenol being the major compound, followed by β-caryophyllene,[31] and α-pinene and *p*-cymene being identified as the major compounds of the essential oil of two *Eucalyptus* species, *E. robusta* and *E. saligna*, respectively, which showed antimicrobial activity against three bacterial strains.[32] Carvacrol is a monoterpene hydrocarbon present in oregano essential oil (*Origanum vulgare*) along with thymol.[33] Widely used in the perfumery and cosmetic industries, the essential oil

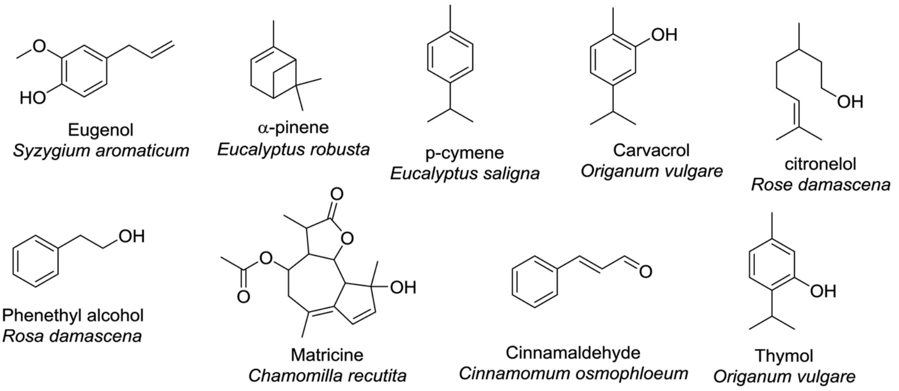

Figure 5.1 Volatile organic compounds identified from essential oil analyses using GC-MS.

produced by steam distillation of *Rosa damascena* is composed mostly of citronellol, while rose absolute, produced by solvent extraction, is composed mostly of phenethyl alcohol.[34] From leaves, the essential oil of *Cinnamomum osmophloeum*, cinnamaldehyde, was determined to be the major compound and after antimicrobial assay, the isolated compound showed potential to be used as an antibacterial additive.[35]

The use of GC-MS was fundamental for the analysis of complex mixtures of natural products, however, one of its limitations is its restriction to volatile compounds. It is interesting to highlight that to circumvent problems of volatility the sample to be separated in the chromatograph can be injected into the mode "on column". This approach was utilized in the analysis of rotenoids. Rotenoids are interesting compounds that display insecticidal activity, and are also known for their potent ichthyotoxicity. Usually this class of compounds is characterized by isolation either by column chromatography or high-performance LC (HPLC). The identification is also carried out by NMR and other spectroscopic data. Pereira and coworkers[36] demonstrated the use of high-temperature high-resolution GC (HT-HRGC) and HT-HRGC MS in on-column injection mode to separate and identify 18 rotenoids from *Tephrosia candida* without derivatization. The separation involved the use of a capillary column coated with PS-090 (20% phenyl, 80% methylpolysiloxane), and the identification was possible through direct analysis of the mass spectra. This analysis allowed the identification of very minor compounds in the mixtures of rotenoids. The fragmentation pattern observed for known major compounds was used to obtain the structures of the minor ones. The compounds identified had very close structures to rotenone and deguelin. Among the compounds identified were: α,α-rotenone; β,β-rotenone; rotenolone; 11-hydroxy-rotenone; 11,12a-dihydroxy-rotenone; 11,8′-dihydroxy-rotenone; 6a,12a-dehydro-rotenone; rotenone; deguelin; tephrosin; α-toxicarol; 12a-hydroxy-α-toxicarol; β-toxicarol; 6a,12a-dehydro-deguelin;

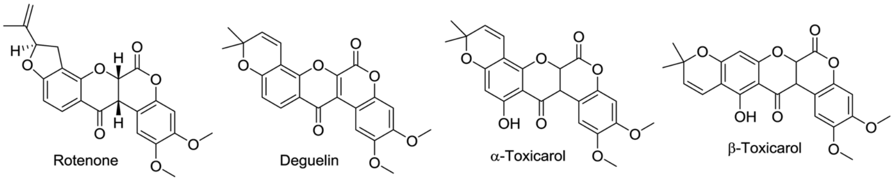

Figure 5.2 Some rotenoids identified from *T. candida.*

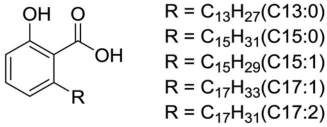

Figure 5.3 GAs from *G. biloba.*

6a,12a-dehydro-α-toxicarol, 6-hydroxy-6a,12a-dehydrodeguelin; 6a,12a-dehydro-β-toxicarol. Some of the structures of the rotenoids are displayed in Figure 5.2.

Another approach to overcoming problems of non-volatile compounds to be analyzed by GC-MS is to derivatize the samples before injection into the chromatograph. This methodology has been applied to the analysis of sugars, which are transformed into silicyl derivatives so that they have the appropriate volatility for injection. A similar approach was used for the investigation of gingko samples. Here the methyl esters of carboxylic acids were obtained for a better analysis of gingkolic acids. Wang and coworkers[37] developed high-resolution GC-MS with a selected ion monitor method to analyze ginkgolic acid (GA) mixtures (Figure 5.3) in *Ginkgo biloba* plant materials, extracts, and commercial products. The method also included quantification of each component in the mixtures. As a general procedure, samples were extracted, submitted to liquid–liquid partitioning and derivatized with trimethylsulfonium hydroxide. The separation was achieved using a polar HP-88 capillary GC column. The GC-MS method was also validated according to ICH guidelines. Samples of *G. biloba* from different sources were analyzed according to the established procedure. Other components, such as the ginkcolides (terpene trilactone) and flavonol glycosides, were determined by further ultra-HPLC-MS (UHPLC-MS). The method was used for the analysis of 19 *G. biloba* authenticated and commercial plant samples and 21 dietary supplements supposed to contain ginkgo leaf extracts.

Even though GC-MS is a technique that is very useful for separating and identifying natural products there are a few examples where separation is not necessary for the identification of compounds in mixture. This can be illustrated in the identification of a series of γ-lactones isolated from *Iryanthera* species.[38] It is proposed that these compounds can be biosynthetically derived from benzoyl or cinnamoyl CoA and pyruvyl CoA with the respective methylene chains of myristic, palmitic, stearic and oleic acids. The fatty

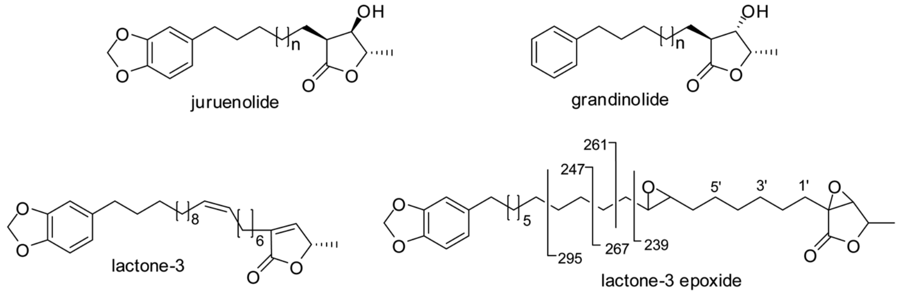

Figure 5.4 Lactones from *Iryanthera* species.

acids differ from each other in the length of the methylene chain, which can have 14, 16 or 18 carbon atoms. The identification of the lactones in mixture was possible through the analysis of EI spectra that afforded parent ions differing from each other by 28 Da, which corresponds to exactly two methylene carbons. One of the lactones had a double bond in the carbon chain and its correct positioning could not be directly determined using EI-MS. To solve this problem, compound 5 was transformed into its epoxide derivative, which, after fragmentation in the mass spectrometer, afforded fragment ions that allowed assigning of the correct position of the double bond. The analysis of the mass spectrum indicated three series of peaks spaced 28 Da apart. In the fragments highlighted in Figure 5.4, compound 5 is indicative of the position of the epoxide and consequently the double bonds in the parent compound.

5.3 Dereplication

Dereplication is directly connected to the process of discovering the identity of an unknown compound based on previous studies in order to avoid repeated and exhausting characterization procedures and reference standards.[39,40] Dereplication is most often applied by using HPLC or UHPLC coupled to diode array detection (DAD) and high-resolution MS (HRMS), in combination with database comparison.[40]

The rapid identification of known compounds present in a mixture is essential to the quick discovery of novel natural products. Dereplication methodologies are making this process feasible through the determination of the molecular formulas, MS/MS, retention time and UV/vis spectra comparison of different compounds in a mixture, and by comparing this information with databases.[41] The quality and availability of natural product databases directly affects the dereplication process.[42] In fact, several databases are available to assist the dereplication process. Some natural-product comprehensive databases that can be used are the Super Natural, the Dictionary of Natural Products, AntiBase, the Dictionary of Marine Natural Products, MarinLit, AntiMarin,[43,44] Metlin,[45] Napralet,[46] and GNPS,[47] among others.

The determination of molecular formula is crucial for natural product identification once reduced numbers of possible molecular formula are generated after accurate mass analysis. HRMS analyzers, such as ion trap, Orbitrap, time-of-flight MS (TOF-MS), quadrupole TOF, and Fourier transform ion cyclotron resonance MS instruments, provide accurate measurements of the *m*/*z* of the ions,[48] with low values of means errors (>1 ppm) depending on the instrument and its calibration.[49] The reliability of accurate masses for unknown compounds can be obtained by comparison with accurate masses of known compounds with close retention times.[50] Nevertheless, information on isotopic distribution and fragmentation patterns is necessary to provide correct elemental composition. Some caution about the elemental composition must be considered when ESI or Atmospheric pressure chemical ionization sources are being used due to the ease of adduct formation. Adducts such as $[M + H]^+$, $[M + Na]^+$, $[M + NH4]^+$, $[M + H + CH_3OH]^+$, $[M + H + ACN]^+$, $[M + K]^+$, *etc.*, are frequently formed in the positive mode of acquisition and sometimes their assignments are not so trivial.[40] Additionally, adduct formation may vary from one LC-MS system to another and can be favored by changing the ionization parameters. It also can change during a sequence because of sodium extraction from solvent glass bottles. Another point to be considered when the molecular formula is being assigned is the ease with which some substances fragment and lose some neutral fragments like water (H_2O), formic acid (HCO_2H), acetic acid (CH_3CO_2H) and carbon dioxide (CO_2).[51] If any of these points are ignored, the molecular mass can be erroneously assigned.

In general, LC-MS is frequently coupled to a DAD detector to obtain information on the UV spectra and, consequently, on chromophores, and this is often used as additional information in database searches to discern compounds with the same elemental composition. Furthermore, many secondary metabolites contain conjugated chromophore systems and some information on their biosynthetic pathways can be obtained, such as non-reduced PKs and NRPs.[40]

Alali and Tawaha demonstrated the use of dereplication techniques for analysis of the secondary metabolite constituents of the genus *Hypericum* using LC coupled to an ESI-MS and LC photodiode with DAD. They identified seven compounds; hypericin, pseudo-hypericin, proto-hypericin, protopseudo-hypericin, hyperforin, and adhyperforin (Figure 5.5), and the flavonoid rutin, from the crude methanolic extracts of the aerial parts (leaves, stems and flowers) based on data from LC-UV/PDA (distinctive UV-spectra for different classes), LC-ESI-MS (total ion chromatogram (TIC), and selected ion monitoring (SIM) chromatogram) and chromatographic elution patterns.[52]

Another aspect to be considered in the dereplication process is sample preparation. This step plays an important role in the final results of the dereplication process. Considering the chemical diversification of each class of secondary metabolite, and the different interaction with different solvents, the choice of the solvent must be carefully considered during the extraction and sample preparation processes. When methanol and ethanol are used as

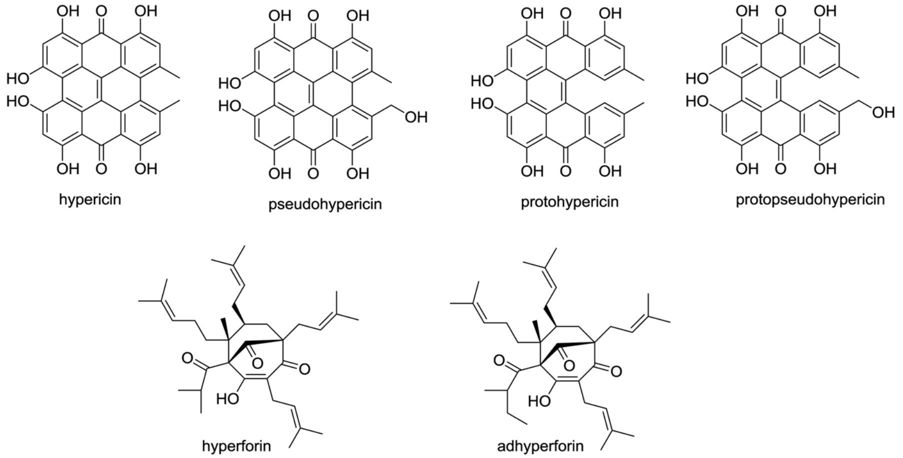

Figure 5.5 Metabolites of *H. triquterifolium* identified by dereplication process.

solvents in the extraction, large amounts of salts and other non-interesting very polar substances can be extracted, leading to interference in the analysis such as ion suppression.[40]

Aiming to solve this problem and make this step faster, Smedsgaard and Frisvad developed a methodology for the fast and efficient extraction of fungal metabolites.[53] This methodology consists in extracting several classes of microbial substances from plugs of cultured fungal on Agar by using a solvent mix and ultrasound bath.[54,55] During the studies, the micro-extraction methodology developed by Smedsgaard and Frisvad allowed the identification of various metabolites by LC-DAD from 395 fungal cultures by using a mix of solvents in the following proportions (v/v): methanol : dichloromethane : ethyl acetate 3 : 2 : 1, showing the efficiency of the extraction methodology.[53] The same methodology was widely used by Nielsen and coworkers showing good results.[40,41,51,56] In a study performed by Nielsen and colleagues, 719 microbial natural products were assigned from *Aspergillus nidulans* and *Aspergillus oryzae* extracts using an LC-DAD-HRMS in both positive (ESI-(+)) and negative (ESI-(−)) modes of ionization by comparison with reference standards and Antibase. About 93% of the substances were identified using ESI-(+), showing that the positive ionization mode was more versatile than the negative mode. Furthermore, unambiguous assignments were done by using adduct patterns; 56% of substances by ESI+ alone and 37% by ESI− alone.[41] Klitgaard and colleagues[51] also carried out the screening of several fungal extracts to identify known secondary metabolites and discover potential novel compounds based on dereplication methodologies. For this, a 15 min UHPLC–DAD–HRMS method combined ESI (ESI+ or ESI−), followed by 10–60 s of automated data analysis, was performed. This process allowed the identification of up to 3000 elemental compositions. Extracted ion chromatograms were automatically generated and overlaid with detected

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

compounds on the base peak chromatogram. By comparison with reference standards, already identified compounds and contaminants from solvents, media and filters, it was possible to assign the known compounds only identified initially by elemental composition and visualize all major potentially novel peaks. Peaks not assigned by elemental formula only were identified by comparison with adduct patterns, UV spectra, retention time compared with log D, co-identified biosynthetic related compounds, and elution order. In total, up to 3000 secondary metabolites could be identified by the dereplication methodology.

5.4 Imaging MS

The process of producing mass spectra, aiming at profiling the chemical composition in a two-dimensional space from intact cells, is called imaging MS (IMS).[57] This process consists in acquiring several mass spectra by scanning small spots of the sample surface and plotting a graph in two dimension (mass charge ratio, *m*/*z*, *vs.* space) after the combination of the acquired spectra with the spatial coordinates.[54,57] In general, the MS used to acquire this information is equipped with a MALDI or a desorption electrospray ionization source (DESI)[55] and a TOF mass analyzer. Obtaining the data can be done with the equipment set in full scan mode or MS/MS of analysis and many spectra are acquired in a single experiment to generate the final image.

IMS has been widely used in natural products to visualize the flow and distribution of bioactive compounds and understand interactions between different organisms[54] and physiological process (defense process against pathogens, for example) in plants.[55]

In the ongoing infection process, pathogens, and host and other microorganisms can interact through physical contact or chemical response (secretion of microbial substances). Direct analysis of microbial co-cultures grown on agar media is an important tool in understanding the interactions between the cultivated microorganisms.

Moniliophthora roreri is considered to be a phytopathogen responsible for infecting cocoa trees and devastating harvests. In fact, this pest is controlled by a biocontrol management pest with the application of endophytic fungi, such as *Trichoderma harzianum*. *T. harzianum* acts by antagonizing *M. roreri*, producing antifungal agents. Studies involving the investigation of the metabolic exchange during the antagonistic interaction between *M. roreri* and *T. harzianum* were performed using DESI-IMS. After three weeks of cultivation of *T. harzianum* and *M. roreri* separately and co-cultured, it was possible to detect the production of metabolites not present when the fungi were cultivated alone. Thorough studies were carried out to identify and localize some phytopathogen-dependent secondary metabolites: T39 butenolide, harzianolide, and sorbicillinol. In this case, the results obtained by MALDI-IMS provided a better understanding of the bioactive substances involved in the chemical ecology of *M. roreri* and *T. harzianum*.[58]

Patients affected by cystic fibrosis are very susceptible to secondary pulmonary infection by opportunistic pathogens, such as *Pseudomonas aeruginosa* and *Aspergillus fumigatus*. Medical observations have revealed that when a patient is infected with either *P. aeruginosa* or *A. fumigatus* they had better pulmonary function than patients who were infected with both *P. aeruginosa* and *A. fumigatus*.[59] This pattern of infection was reproduced in rats and results showed a different response of that observed in humans: rats co-infected with *P. aeruginosa* and *A. fumigatus* exhibited a higher survival rate than those infected with only *A. fumigatus*.[60] In an attempt to get more insight into the interaction between *A. fumigatus* and *P. aeruginosa* the microorganisms were cultivated on agar culture media analyzed by MALDI-IMS and compared to the controls (fungus and bacteria growth separately). The results showed that phenazine, a nitrogen-containing heterocyclic metabolite produced by *P. aeruginosa*, contributes to the *P. aeruginosa* colonization of the lungs of cystic fibrosis patients. In parallel, *A. fumigatus* converts phenazine in other substances with distinct biological activity. This fungal conversion consisted in the transformation of pyocyanin and phenazine-1- carboxylic acid into phenazine dimers and phenazine-1-carboxylic acid into 1-hydroxyphenazine. Additionally, *A. fumigatus* also converted 1-hydroxyphenazine into 1-methoxyphenazine and phenazine-1 sulfate.[61]

The infection *Xylella fastidiosa*, a Gram-negative bacterium, in the xylem of citrus plants causes diseases in several crops with high economical value. When *X. fastidiosa* infects the xylem of sweet orange, it causes a disease called citrus variegated chlorosis (CVC) resulting in crop loss.[62] In order to understand and evaluate the infection/defense process of a plant against the pathogen, the influence of rootstock on the content of bioactive compound studies of biotic and abiotic stress needed to be carried out, being the inoculation of the microorganism, a type of biotic stress.[63]

Aiming to investigate the infection of *X. fastidiosa* on *C. sinensis* and *C. limonia*, Santos and coworkers[63] developed a quantification method using HPLC-UV to quantify hesperidin and rutin levels in leaves and stems of *C. limonia* and *C. sinensis* grafted onto *C. limonia* with and without CVC symptoms after *X. fastidiosa* infection. It was observed that rutin levels remained constant but hesperidin was increased in symptomatic leaves. Scanning electron microscopy experiments on leaves with CVC symptoms revealed vessel occlusion by biofilm and the presence of crystallized material. Due to the difficulty of isolating and identifying these crystals through conventional methods, IMS was carried out. For this purpose, tissue sections were analyzed by MALDI-IMS to confirm the presence of hesperidin at the site of infection. Initially, MS/MS experiments were performed to obtain a specific ion as a diagnostic for hesperidin ($m/z = 483$) based on higher ion intensity for infected and healthy plants, mainly in the vessel regions. Then the images were constructed from MS/MS data with this specific diagnostic fragment ion ($m/z = 483$), confirming the presence of hesperidin in the infected tissues. These data suggest that hesperidin is directly involved in the plant–pathogen defense process, probably acting as a phytoanticipin (secondary metabolites produced by plants with defensive roles against microorganisms). This methodology was applied to *C. sinensis* and *C. limonia* seedlings, and the results showed that the amount of hesperidin was about 3.6 greater in the rootstock in the graft stem than in the stem of *C. sinensis* seedlings when compared with the graft. An increase in hesperidin content in rootstock can be related to induced internal defense mechanisms.[63]

5.5 MS and Quality Control of Herbal Medicines

In many countries, the control of herbal medicinal products (HMPs) is almost non-existent. Products are launched onto the market without any control of quality, clearly showing that no evaluation of toxicity and safety have been carried out. The consumption of such products is certainly dangerous and may cause severe health problems. Having in mind that it is imperative to ensure the quality of herbal products Nguyen and Kimaru[64] conducted a work to check for the main constituents of HMP obtained from a Somalian patient. The properties of this product were attributed to the *Commiphora molmol* species. Among the different techniques used for the analysis of the constituents of this plant, GC-MS was chosen for the identification of the volatile compounds that were obtained by both distillation and Soxhlet extraction. The extracts were injected into a chromatograph mounted with a 5% phenylmethylpolysiloxane capillary column and electron ionization. Compounds were identified by comparison of their mass spectra with a built-in library of the National Institute of Standards and Technology (NIST). The chemistry of this species is well documented and among the compounds identified were monoterpenoids, including α-pinene, camphene, β-pinene, myrcene, and limonene, and sesquiterpenoids, grouped in the main categories of: germacrane, eudesmane, guaiane, cadinane, elemane, bisabolane, and oplopane.

A number of sesquiterpenes related to *C. molmol* were identified together with some other compounds that are not characteristic of this species such as: acetonyldimethylcarbinol, isobutyl formate, secbutyl nitrite, acetic acid butyl ester, dimethylfulvene, isobutyl methacrylate, ethyl-3-propylacrolein, strophanthidol, decamethylcyclopentasiloxane, phenanthrenone, octahydronaphthalenone, and 1,2,3,4,5,6,7,8-octahydroanthracene. These compounds were probably found because they are contaminants formed either during the processing of plant resin, or degradation, and/or adulteration of the extracts due to storage conditions.

In a recent paper, Zhou and coworkers[65] discuss the use of LC-TOF-MS for the analysis of herbal medicines. In this case, all the aspects of LCMS-IT-TOF, a new type of mass spectrometer that combines ion traps (ITs) and TOF technologies, were presented, including stationary and mobile phases for LC, accurate mass, fragmentation and selectivity for MS. The application of the methodology for herbal samples describing a parallel between advantages and pitfalls of the approach for qualitative and quantitative analysis was also discussed. It is important to observe that the chemical characterization of herbal medicines is pivotal in establishing their safety and efficacy. For this purpose, the application of LC-TOF seems to be an appropriate method to achieve the correct information in the qualitative analysis of the chemical constituents of herbal samples.

Acknowledgements

The authors are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

1. L. H. Conover, in *Drug Discovery*, ed. B. Bloom, Washington DC, 1971, pp. 33–80.
2. R. M. P. Gutierrez, A. M. N. Gonzalez and A. M. Ramirez, *Curr. Med. Chem.*, 2012, **19**, 2992–3030.
3. A. Amedei and M. M. D'Elios, *Curr. Med. Chem.*, 2012, **19**, 3822–3840.
4. F. Javed, M. I. Qadir, K. H. Janbaz and M. Ali, *Crit. Rev. Microbiol.*, 2011, **37**, 245–249.
5. E. M. Costa-Neto, *An. Acad. Bras. Cienc.*, 2005, **77**, 33–43.
6. H. Budzikiewicz, C. Djerassi and D. H. Williams, *Structure Elucidation of Natural Products by Mass Spectrometry: Alkaloids*, Holden-Day, San Francisco, 1964, vol. 1.
7. C. Djerassi, B. Gilbert, J. N. Shoolery, L. F. Johnson and K. Biemann, *Experientia*, 1961, **17**, 162–163.
8. S. Agurell, B. Holmstedt and J. Lindgren, *Am. J. Pharm. Sci. Support. Public Health*, 1968, **140**, 148–151.
9. C. J. W. Brooks, *Philos. Trans. R. Soc. London A*, 1979, **293**, 53–67.
10. S. Popov, R. K. Carlson, A. Wegmann and C. Djeraasi, *Tetrahedron Lett.*, 1976, **6**, 3491–3494.
11. J. Maree, G. Kamatou, S. Gibbons, A. Viljoen and S. Van Vuuren, *Chemom. Intell. Lab. Syst.*, 2014, **130**, 172–181.
12. C. G. Hammar, B. Holmstedt, J. E. Lindgren and R. Tham, *Adv. Pharmacol.*, 1969, **7**, 53–89.
13. J. S. Dickschat, *Nat. Prod. Rep.*, 2014, **31**, 838–861.
14. C. Kuhlisch and G. Pohnert, *Nat. Prod. Rep.*, 2015, **32**, 937–955.
15. N. Schauer, D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M. G. Forbes, L. Willmitzer, A. R. Fernie and J. Kopka, *FEBS Lett.*, 2005, **579**, 1332–1337.
16. H. M. Fales, H. A. Lloyd and G. W. A. Milne, *J. Am. Chem. Soc.*, 1970, **92**, 1590–1597.
17. G. W. Milne, T. Axenrod and H. M. Fales, *J. Am. Chem. Soc.*, 1970, **92**, 5170–5175.
18. H. Fales, G. Milne and R. Nicholson, *Anal. Chem.*, 1971, **43,** 1785–1789.
19. M. Barber, R. S. Bordoli, R. D. Sedgwick and A. N. Tyler, *J. Chem. Soc., Chem. Commun.*, 1981, 325–327.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

20. S. K. Chowdhury and B. T. Chait, in *Annual Reports in Medicinal Chemistry*, ed. R. C. Allen, Academic Press, 1989, pp. 253–263.
21. R. Schmidt, D. W. Etalo, V. de Jager, S. Gerards, H. Zweers, W. de Boer and P. Garbeva, *Front. Microbiol.*, 2016, **6**.
22. R. Hung, S. Lee and J. W. Bennett, *Appl. Microbiol. Biotechnol.*, 2015, **99**, 3395–3405.
23. E. Cassel, R. M. F. Vargas, N. Martinez, D. Lorenzo and E. Dellacassa, *Ind. Crops Prod.*, 2009, **29**, 171–176.
24. J. Vergis, P. Gokulakrishnan, R. K. Agarwal and A. Kumar, *Crit. Rev. Food Sci. Nutr.*, 2015, **55**, 1320–1323.
25. S. Burt, *Int. J. Food Microbiol.*, 2004, **94**, 223–253.
26. G. Ruberto and M. T. Baratta, *Food Chem.*, 2000, **69**, 167–174.
27. L. Pessoa, S. Morais, C. M. Bevilaqua and J. H. Luciano, *Vet. Parasitol.*, 2002, **109**, 59–63.
28. J. U. Rehman, A. Ali and I. A. Khan, *Fitoterapia*, 2014, **95**, 65–74.
29. Z. Zhang and G. Li, *Microchem. J.*, 2010, **95**, 127–139.
30. P. Rubiolo, B. Sgorbini, E. Liberto, C. Cordero and C. Bicchi, *Flavour Fragrance J.*, 2010, **25**, 282–290.
31. V. K. Raina, S. K. Srivastava, K. K. Aggarwal, K. V. Syamasundar and S. Kumar, *Flavour Fragrance J.*, 2001, **16**, 334–336.
32. P. Sartorelli, A. D. Marquioreto, A. Amaral-Baroli, M. E. L. Lima and P. R. H. Moreno, *Phytother. Res.*, 2007, **21**, 231–233.
33. D. Vokou, S. Kokkini and J.-M. Bessiere, *Biochem. Syst. Ecol.*, 1993, **21**, 287–295.
34. S. Ulusoy, G. Boşgelmez-Tınaz and H. Seçilmiş-Canbay, *Curr. Microbiol.*, 2009, **59**, 554–558.
35. S.-T. Chang, P.-F. Chen and S.-C. Chang, *J. Ethnopharmacol.*, 2001, **77**, 123–127.
36. A. S. Pereira, A. C. Pinto, J. N. Cardoso, F. R. A. Neto, P. C. Vieira, J. B. Fernandes, M. F. das G. F. da Silva and C. C. Andrei, *J. High Resolut. Chromatogr.*, 1998, **21**, 513–518.
37. M. Wang, J. Zhao, B. Avula, Y.-H. Wang, C. Avonto, A. G. Chittiboyina, P. L. Wylie, J. F. Parcher and I. A. Khan, *J. Agric. Food Chem.*, 2014, **62**, 12103–12111.
38. P. C. Vieira, M. Yoshida, O. R. Gottlieb, H. F. Paulino Filho, T. J. Nagems and R. Braz-Filho, *Phytochemistry*, 1983, **22**, 711–713.
39. D. Krug and R. Müller, *Nat. Prod. Rep.*, 2014, **31**, 768–783.
40. K. F. Nielsen and T. O. Larsen, *Front. Microbiol.*, 2015, **6**, 1–15.
41. K. F. Nielsen, M. Månsson, C. Rank, J. C. Frisvad and T. O. Larsen, *J. Nat. Prod.*, 2011, **74**, 2338–2348.
42. S. P. Gaudêncio and F. Pereira, *Nat. Prod. Rep.*, 2015, **32**, 779–810.
43. J. Hubert, J. M. Nuzillard and J. H. Renault, *Phytochem. Rev.*, 2015, 1–41.
44. A. L. Harvey, R. Edrada-Ebel and R. J. Quinn, *Nat. Rev. Drug Discovery*, 2015, **14**, 111–129.
45. R. Tautenhahn, K. Cho, W. Uritboonthai, Z. Zhu, G. J. Patti and G. Siuzdak, *Nat. Biotechnol.*, 2012, **30**, 826–828.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00144

46. J. G. Graham and N. R. Farnsworth, *Comprehensive Natural Products II*, Elsevier, 2010, pp. 81–94.
47. M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crusemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderon, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodriguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Muller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. O. Palsson, K. Pogliano, R. G. Linington, M. Gutierrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.
48. T. Cai, Z.-Q. Guo, X.-Y. Xu and Z.-J. Wu, *Mass Spectrom. Rev.*, 2016, **47**, 987–992.
49. A. G. Marshall and C. L. Hendrickson, *Annu. Rev. Anal. Chem.*, 2008, **1**, 579–599.
50. Z.-J. Wu, G.-Y. Li, D.-M. Fang, H.-Y. Qi, W.-J. Ren and G.-L. Zhang, *Anal. Chem.*, 2008, **80**, 217–226.
51. A. Klitgaard, A. Iversen, M. R. Andersen, T. O. Larsen, J. C. Frisvad and K. F. Nielsen, *Anal. Bioanal. Chem.*, 2014, **406**, 1933–1943.
52. F. Q. Alali and K. Tawaha, *Saudi Pharm. J.*, 2009, **17**, 269–274.
53. J. Smedsgaard and J. C. Frisvad, *J. Microbiol. Methods*, 1996, **25**, 5–17.
54. C.-J. Shih, P.-Y. Chen, C.-C. Liaw, Y.-M. Lai and Y.-L. Yang, *Nat. Prod. Rep.*, 2014, **31**, 739–755.
55. N. Bjarnholt, B. Li, J. D'Alvise and C. Janfelt, *Nat. Prod. Rep.*, 2014, **31**, 818–837.
56. K. F. Nielsen and J. Smedsgaard, *J. Chromatogr. A.*, 2003, **1002**, 111–136.
57. E. Esquenazi, Y.-L. Yang, J. Watrous, W. H. Gerwick and P. C. Dorrestein, *Nat. Prod. Rep.*, 2009, **26**, 1521–1534.
58. A. Tata, C. Perez, M. L. Campos, M. A. Bayfield, M. N. Eberlin and D. R. Ifa, *Anal. Chem.*, 2015, **87**, 12298–12305.
59. R. Amin, A. Dupuis, S. D. Aaron and F. Ratjen, *Chest*, 2010, **137**, 171–176.

60. M. Yonezawa, H. Sugiyama, K. Kizawa, R. Hori, J. Mitsuyama, H. Araki, M. Shimakura, S. Minami, Y. Watanabe and K. Yamaguchi, *J. Infect. Chemother.*, 2000, **6**, 155–161.
61. W. J. Moree, V. V. Phelan, C.-H. Wu, N. Bandeira, D. S. Cornett, B. M. Duggan and P. C. Dorrestein, *Proc. Natl. Acad. Sci.*, 2012, **109**, 13811–13816.
62. A. A. de Souza, M. A. Takita, E. O. Pereira, H. D. Coletta-Filho and M. A. Machado, *Curr. Microbiol.*, 2005, **50**, 223–228.
63. M. S. Soares, D. F. da Silva, M. R. Forim, M. F. das G. F. da Silva, J. B. Fernandes, P. C. Vieira, D. B. Silva, N. P. Lopes, S. A. de Carvalho, A. A. de Souza and M. A. Machado, *Phytochemistry*, 2015, **115**, 161–170.
64. H. P. Nguyen and I. W. Kimaru, *LC-GC*, 2014, **12**, 14–18.
65. J.-L. Zhou, L.-W. Qi and P. Li, *J. Chromatogr. A*, 2009, **1216**, 7582–7594.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

CHAPTER 6

Applications of Mass Spectrometry in Synthetic Biology

ZAIRA BRUNA HOFFMAM[a], VIVIANE CRISTINA HEINZEN DA SILVA[a], MARINA C. M. MARTINS[a] AND CAMILA CALDANA*[a,b]

[a]Brazilian Bioethanol Science and Technology Laboratory (CTBE), Rua Giuseppe Máximo Scolfaro 10.000 CEP 13083-100 Campinas, São Paulo, Brazil; [b]Max Planck Partner Group at Brazilian Bioethanol Science and Technology Laboratory, Brazil
*E-mail: camila.caldana@bioetanol.org.br

6.1 Introduction

The emerging field of synthetic biology aims to make use of the principles of engineering to understand and re-design biological systems, rendering cells/organisms with predictable and novel functions.[1–3] Synthetic biology combines advances in several areas of knowledge, including molecular and cell biology, engineering, computational biology, biochemistry, and systems biology, with the ultimate intention of turning living systems into biological factories. Although this opens up an avalanche of possibilities for producing high-value substances, such as agrochemicals (*e.g.* bioherbicides), drugs (*e.g.* antibiotics, immunosuppressants), polymers (*e.g.* hydrogels), *etc*, reverse genetic tools must be available for the organism of interest. In addition, the optimization of multiple steps such as promoters, ribosome binding sites,

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

gene order arrangements, and control of gene expression must be also taken into consideration, as well as protein–protein interactions, availability of cofactors and precursors, and decreased competition from alternative reactions in order to achieve the synthesis of the final bio-product in a desirable range.[4]

Despite the advances in this field, we are still far from being able to engineer biological systems with the same precision and speed that electronic circuits are engineered.[5] In fact, the design and discovery of synthetic pathways and genes with new functions is a bottleneck that must be overcome to achieve this rationalization goal (see Figure 6.1). Thus, although it is important to draw together the combination of these parts to construct cell factories, it is also important to use robust techniques to characterize and explore the individual parts. The emergence of novel molecular profiling and analytical techniques that allow the identification and quantification of the basic functional components that form a biological systems are making it possible to uncover their interactions and obtain a more complete picture of how organisms respond to different perturbations and stimuli. In this context, mass spectrometry (MS) has made a huge contribution to understanding the dynamics of proteins and metabolites, facilitating the manipulation of living systems to produce a variety of bioproducts. In this chapter, we provide a concise overview of the progress made by the use of MS in the detection of target molecules, identification of metabolic engineering bottlenecks and

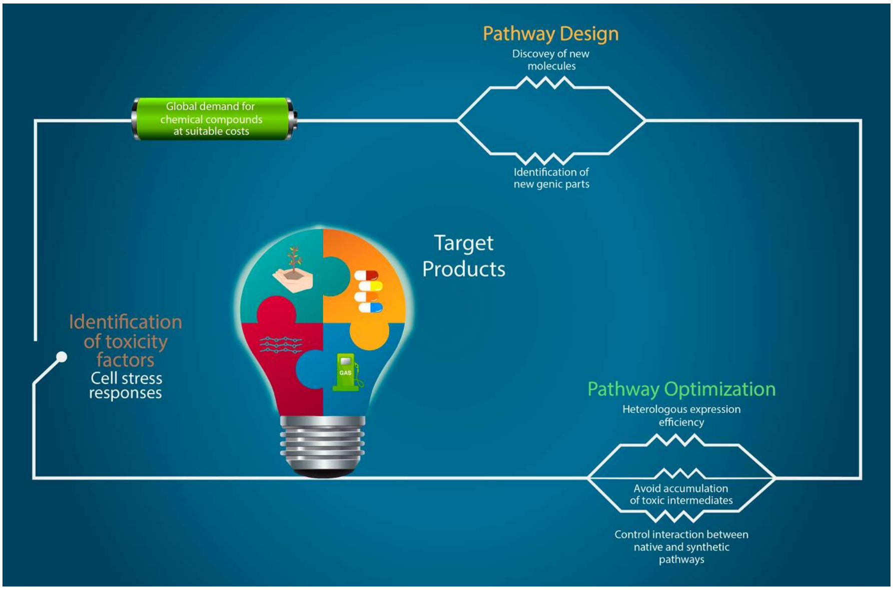
Pathway Design
Discovey of new molecules
Identification of new genic parts
Global demand for chemical compounds at suitable costs
Target Products
Identification of toxicity factors
Cell stress responses
GAS
Pathway Optimization
Heterologous expression efficiency
Avoid accumulation of toxic intermediates
Control interaction between native and synthetic pathways

Figure 6.1 Challenges to be overcome in building biosynthetic pathways. The range of compounds that can be obtained from recombinant microorganisms include agrochemicals, drugs, polymers and fuels.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

Table 6.1 Examples discussed in this chapter.

Target molecules/phenotype	Authors	MS technique applied
Antibiotics	Gillespie *et al.*[9]	HRFABMS (chemical structure elucidation)
Amorpha-4,11-diene	Martin *et al.*[18]	GC-MS (metabolomics)
Isopentenol	George *et al.*[19]	QTRAP LC-MS/MS (proteomics) and LC time-of-flight (TOF) MS (metabolomics)
Cell stress responses to accumulation of toxic metabolites	Rutherford *et al.*[20]	LC quadrupole TOF (proteomics)
Cell stress responses to accumulation of toxic metabolites	Hasunuma *et al.*[21]	CE/GC-MS (metabolomics)
N-acetylglucosamine	Liu *et al.*[13]	UHPLC-ESI-QqQ (metabolomics)
Artemisinin	Pitera *et al.*;[27] Fuentes *et al.*[32]	GC-TOF MS and LC-MS (metabolomics)

fine-tuning of the synthetic modules to successfully enhance the production of desired chemicals. Examples that will be discussed are summarized in Table 6.1.

6.2 MS as Emerging Tool for Synthetic Biology

6.2.1 Prospecting for Target Molecules

Plants are exceptional chemical systems, which efficiently convert photochemical energy into carbohydrates that serve as the carbon skeleton for several building blocks. As sessile organisms, plants have evolved the biosynthesis of a plethora of small molecules to survive and cope with environmental cues.[6] Although plants contain a cornucopia of novel chemicals, only 15% of plant species have been explored for their chemical composition, suggesting that only a limited percentage of the chemicals have been investigated. Similarly, it is estimated that only 2% of the chemicals produced by microorganisms to cope with interspecies competition and communication have been discovered.

Despite the great potential of biodiversity to unravel the novel chemicals and biochemical pathways that could be exploited for synthetic biology, the discovery of those metabolites relies on the capability of the emerging technologies in temporal and spatial detection in complex mixtures as well as chemical structural elucidation from small quantities of extracted samples.[6,7] MS has emerged as a viable option for sensitive mass specific detection within complex matrices. Traditionally, target molecule identification relied on bioactivity assays, in which extracts were screened for certain biological activity, followed by extract fractionation and structure elucidation. The drawback of this workflow was the rediscovering of the same natural

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

products. An illustration of this statement is the difficulty of discovering new antibiotics to fight against emerging pathogens that acquire multiple resistance to available drugs. Soil is the main source of most of the compounds with antimicrobial activity that have been so far discovered and thus new ones can still potentially be discovered in this environment. However, current estimates show that less than 1% of soil microorganisms are cultivable, which results in the low efficiency of traditional prospecting techniques.[8] As only few microorganisms are cultivable, most drugs are rediscovered several times before a new compound is found. As such, reports of compounds with new chemical structures that operate in the elimination of pathogens through mechanisms different from those already reported are rather uncommon.[9]

Aware of this scenario, Gillespie *et al.*[9] reported the isolation of two new organic polycyclic aromatic cations (triaryl cations) from the screening of a metagenomic soil library containing 24 546 clones of Escherichia coli carrying bacterial artificial chromosome (BAC) vectors. This work started with the selection of three clones inside this library that produced brown and orange coloured compounds in a Luria-Bertani medium. The brown material suggested the production of bacterial melanins, which led the authors to further investigate these clones. After steps of acid precipitation, extraction using methanol, and chromatographic separation, two compounds, one red and one orange, were isolated from the supernatant culture of three clones named P57G4, P89C8 and P214D2. Due to the absence of information on the red and orange compounds, a complete set of chemical characterization techniques, including MS, were employed in order to determine the elemental composition of these compounds and confirm their structure. The MS platform of choice was high-resolution fast atom bombardment MS (HRFABMS), which has as its principle the bombardment of the analyzed sample with a stream of primary noble gas atoms, so that secondary ions are ejected from the sample and their mass determined. HRFABMS first allowed the identification of the molecular formula $C_{25}H_{18}N_3$ and $C_{23}H_{17}N_2$ for the compounds turbomycin A (orange compound) and turbomycin B (red compound), respectively. While turbomycin A had previously been characterized from fungi but never reported as a bacterial metabolite, turbomycin B had not been found before. Both compounds were tested against several genera of Gram-positive and negative bacteria, exhibiting a broad spectrum of antimicrobial activity. This work in metagenomics is already considered to be exceptional because of the discovery of new compounds from a relatively small experimental universe, considering that other traditional techniques are comparatively much more laborious and limited. However, the work of Gillespie *et al.* also stands out for the way clones were sequenced and mutated using transposons, which interrupt genes, eliminating their function. The authors concluded that a single open reading frame (ORF) present in the P57G4 clone, which belongs to a bacterial locus involved in the tyrosine degradation pathway, was necessary and sufficient to supplement the *E. coli* genome, leading to the production of the triaryl cations. The sequence identified in this study had low identity with those deposited in databases at the time of the work, then representing a new

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

"part" that could be useful for the creation of chimeric synthetic routes. MS was inserted in this context as an auxiliary yet indispensable tool for drawing useful conclusions from experimental observations, contributing to leading the authors to the genic level of the observed phenotype.

In addition to the identification of the chemical structure and molecular formula of a given target molecule, MS-based approaches can be used to develop economically viable synthetic organisms in high-throughput screenings for production optimization.[10] Despite the accuracy and sensitivity of MS for screening studies, most MS assays required a chromatographic separation prior to MS detection to reduce sample complexity. To overcome this limitation, new technologies, such as Rapid Firesystem, combining solid phase extraction cartridges and direct infusion in a tandem MS, offer high throughput and robustness by processing a 96-well plate in less than 11 min,[11] opening up a new avenue for target molecule detection in metabolic engineering. In addition to its high throughput, this system does not require chromatographic separation and benefits from its runtimes.

6.2.2 Pathway Design and Optimization

Most of the high-value small molecules cannot be synthetized on large scales using the native metabolic pathway. For the efficient formation of the target metabolite, innovative strategies are required to engineer the synthetic pathway. The evolution of next-generation sequencing technologies has permitted an exponential increase in genome and metagenome sequences, unravelling a wealth of novel methods for exploiting biosynthetic routes[12] and querying target genes. However, in most cases, solely introducing genes into a host organism is insufficient to generate huge quantities of a molecule of interest. It can be challenging to properly match the different modules of the system as, usually, the target pathway encompasses several committed conversion steps involving a range of intermediates and enzymes with different catalytic activities. Some of these intermediates can be toxic to the cell and their accumulation is not desirable. Therefore, how is it possible to ensure the concerted action of the enzymes in a synthetic pathway in order to avoid intermediate accumulation and guarantee a perfect flux along the path?

The engineering of workhorse strains to receive new biosynthetic pathways goes far beyond the simple introduction of new genes. In recent years, the cheapness of next-generation sequencing and the invention of Cas9-mediated genome editing have truly revolutionized the lineage construction, which can now be engineered in a more precise way than ever before.[13] However, several metabolic aspects must be considered for overall path optimization. Usually, the first item to be verified is the efficiency of the heterologous expression in a given organism, that is, whether the gene is being transcribed into RNA. For example, currently the mRNA abundance can be readily verified using next-generation sequencing or reverse transcription polymerase chain reaction (RT-PCR). While necessary, the analysis of the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

transcripts does not provide all the desired information, once it does not directly correlate with the protein levels that are actually present in a cell at a given moment.

The process in which mRNA is translated into proteins is regulated by factors that have not yet been fully elucidated. For example, it is known that each microorganism has a preference for a particular codon set.[14] Thus, the introduction of genes coding proteins derived from different biological systems into a new host to constitute a synthetic pathway can be translated into different levels, leading to the accumulation of intermediates. In addition to the codon usage, other factors that influence the abundance of proteins in a given cell time are: the mRNA secondary structure, the ribosome binding site strength, the transcript ribosome occupancy, the half-life of the mRNA and of the protein produced, the promoter strength, spacing between genetic elements, and the position of genes in the operon.[10–15] Therefore, development of methods to accurately monitor the protein expression can boost the engineering to succeed. Although immunoblotting techniques are widely used for protein detection and quantification, assessing the expression levels of different proteins from the same pathway can be demanding.

A great method for targeted proteomics is the tandem-based method named selected-reaction monitoring MS (SRM-MS), whereby a precursor ion is initially isolated in a first MS stage and then a second ion product of its fragmentation is selected for detection in a second stage.[16,17] A successful example of SRM-MS application in synthetic biology was the mensuration of proteins involved in the sesquiterpene amorpha-4,11-diene pathway, a precursor of artemisinic acid.[15] Sesquiterpenes are hydrocarbons made up of three isoprene units that may or not be ramified. Previously, Martin and co-workers[18] had modified the strain *E. coli* DH1 with nine gene manipulations, introducing five genes of the mevalonate biosynthesis pathway from *Saccharomyces cerevisiae*, the amorpha-4,11-diene synthase gene from *Artemisia annua* and extra copies of three endogenous *E. coli* genes. In this context, SRM-MS aimed at improving the production of the amorpha-4,11-diene, and understanding the bottlenecks of the artificially constructed pathway.

From the observation of SRM transitions for proteins and the selection of the most abundant peptides that provide unique identification for each one of the nine proteins, the authors found that the bottleneck for producing amorpha-4,11-diene was a flaw in the translation of two proteins, the mevalonate kinase (MK) and the phosphomevalonate kinase (PMK). Such a flaw was not evident from the measurements of the mRNAs through RT-PCR, which were present for these enzymes at similar levels as for other enzymes belonging to the pathway. The authors then optimized the MK and PMK codons for expression in *E. coli*. They also changed the control of the expression to a strong promoter. These modifications made sesquiterpene production increase by more than 300%.

A similar approach was used for optimizing the production of isopentenol, a potential biofuel that can also be produced by including

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

modifications in the mevalonate pathway.[19] A dozen different plasmids containing different combinations of seven enzymes leading from acetyl-CoA to isopentenol along the mevalonate pathway were constructed. Within the plasmid modifications included the type of promoter, the order of ORFs in operons and the sequence of each protein, thereby changing the codon usage. The isopentenol was then produced for each strain, and proteins and metabolites were quantified for each experiment. From the measurements, the authors established correlations between levels of proteins and metabolites, creating a model to reveal the critical bottlenecks of the path inserted in *E. coli*. Two proteins were defined as critical points in the isopentenol production, one responsible for an irreversible reaction that traps the carbon flux in the mevalonate pathway, the 3-hydroxy-3-methylglutaryl-CoA synthase (HMGS), and again the MK. An interesting finding of this work is the fact that MK is only a critical step in the pathway when the HMGS is being expressed at high levels. Such a pattern is due to the inhibition of MK by the accumulation of its substrate, the mevalonate, when HMGS directs the carbon flux towards the mevalonate pathway. The strain engineered in this work by balancing the factors that impact the model produced 1.5 g L^{-1} of isopentenol, reaching 46% of the theoretical yield. Despite being minor in absolute value, this number is five times greater than that previously reached in a proof of concept, which was 8%.[19]

In an attempt to look beyond to protein expression, once the expression reaches suitable levels, what else impacts product formation? Native pathways have evolved to avoid the formation of toxic intermediates. Efforts to further optimize synthetic biological systems have taken metabolomics- and proteomics-based approaches to identify toxicity factors. The industrial production of alternative biofuels such as n-butanol using microbes has gained popularity even though this short chain alcohol is more toxic to cells than ethanol. Shotgun liquid chromatography (LC)-MS/MS-based proteomics studies have enabled the characterization of the cell stress responses to high levels of this molecule, allowing the identification of the key proteins recruited to alleviate the stress response and consequently to improve the synthetic production of n-butanol.[20]

Another example of an MS application for detecting the toxic effect of pathway intermediates in engineered cells was the analysis of the metabolome of a recombinant xylose-fermenting strain of *Saccharomyces cerevisiae* by capillary electrophoresis gas chromatography (CE-GC)-MS.[21] The main drawback to the use of *S. cerevisiae* for ethanol production from lignocellulosic hydrolysates is the lack of metabolic enzymes that can use xylose, the most common pentose sugar in those fractions, as a substrate for fermentation. Efforts to reconstruct an efficient xylose assimilation pathway in *S. cerevisiae* have been mainly through the introduction of key enzymes from other organisms such as *Scheffersomyces stipitis* and *Piromyces*. An issue to be considered together with xylose fermentation is that lignocellulosic hydrolysates have a range of toxic compounds, including acetic acid and phenolics.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

A kinetics analysis of the metabolome of one xylose-fermenting strain of *S. cerevisiae* exposed to acetic acid treatment revealed the accumulation of several intermediates of the non-oxidative pentose phosphate pathway (PPP), such as sedoheptulose-7-phosphate and ribulose-5-phosphate, indicating a reduction in xylose consumption that could be overcome by targeting PPP enzymes.[21] This example illustrates well the power of MS-based metabolomics in developing rational strategies to achieve stress tolerance through metabolic engineering.

Although the efficiency of a given synthetic pathway can usually be assessed by the final product titre, factors contributing to the pathway performance, such as consumption of metabolites by native pathways and metabolic flux, cannot only be evidenced by monitoring productivity. Overexpression of native pathways may lead to an increase in flux, promoting imbalance in enzymatic activity levels and higher fluctuations in metabolite concentration.[5] In addition, it is necessary to highlight that intermediate metabolites of a newly introduced pathway in a given host can interact with its endogenous pathways. In other words, the intermediate metabolite flow may suffer scavenging through internal pathways of the workhorse strain that hosts the heterologous expression. This depletion is detrimental, first because it hides the real bottlenecks to optimizing the route of synthesis and also because it also decreases the production yield. In the breakdown of the costs of the fermentative process, the raw material, that is, the sugars consumed by the microorganism, is often the element of greatest impact in the final cost. Any capture of intermediates by the endogenous metabolism that diverts the carbon flow then increases the final cost of the process. When this diversion occurs, the microorganism consumes the carbon source but it turns it into other compounds rather than into the target product.

The detection of precursors and intermediates in a single statistic point of the metabolism, as described in the two examples mentioned above, is usually not enough to elucidate the imbalance between native and heterologous pathways, since the metabolism is dynamic. Thus, an evaluation of the metabolic flux in a particular stage of growth or production might not explain the changes in microbial physiology of a native host after the introduction of a synthetic pathway.[22] Research groups have been working on the improvement of techniques of analysis to cover problems that are not solved by experiments conducted in metabolic steady states. For example, Liu *et al.*[13] presented a dynamic metabolomics approach to improve a synthetic *N*-acetylglucosamine (GlcNAc) pathway in *Bacillus subtilis*. *N*-acetylglucosamine is an amino sugar that has nutraceutical applications and hence is a target of the pharmaceutical industry. GlcNAc and glucosamine are precursors of glycosaminoglycans such as hyaluronic acid and chondroitin sulphate.[23]

Production of GlcNAc in a *B. subtilis* strain[23] was achieved by the overexpression of two enzymes, an endogenous glucosamine-6-P synthase and a heterologous *N*-glucosamine-6-P *N*-acetyltransferase from *S. cerevisiae*. In addition, the authors deleted enzymes of the GlcNAc catabolic pathway

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

and *via* a module engineering approach, performed a first improvement in the synthetically constructed *B. subtilis* strain. Nevertheless, this lineage did not show an adequate performance for scaling towards industrial production because, in a minimal media, adding glucose as a carbon source greatly reduced production and growth. In the first investigations aimed at clarifying what was happening in the metabolism, Liu *et al.*[13] modified the flow of substrates in the pathway by replacing the native phosphofructokinase (PFK) with an Arg252Ala-mutated PFK with reduced activity as well as overexpressing the glutamine synthase. These experiments led them to discard the hypothesis that the problems regarding the pathway could be related to the supply of the three precursors (fructose-6-P, acetyl-CoA and glutamine), which were present in adequate concentrations. As a next step, the metabolomics analysis of the cell at a stationary point of the metabolism (mid-exponential growth) revealed that the concentration of *N*-acetylglucosamine-6-P widely varied, by almost six hundred times, in the modified strain compared to the wild type. This notorious change suggested that some metabolic disturbance in the pathways surrounding the *N*-acetylglucosamine-6-P, in which this molecule could be an intermediary, could be the cause of the poor performance of the mutant. The authors reported that it would be rather impossible to judge whether this flux occurred due to low enzymatic activity of a dephosphorylating enzyme or due to transport limitations by analysing only the metabolomics data of steady state.

In order to understand the dynamics of metabolites around the *N*-acetylglucosamine-6-P, Liu *et al.*[13] constructed models including the four linear metabolites of the *N*-acetylglucosamine production pathway and the secretion step. The variations in the metabolite concentration were simulated in various scenarios, including withno limitations in the pathway, with feedback inhibition, with limitations in enzyme abundance, with a futile cycle inside the pathway, and with a futile cycle at the end of the pathway. In addition, an experiment monitored the changes in the metabolite concentration of the pathway. The results led the authors to the conclusion that the changes in metabolite concentration over time fitted the simulation case in which there was a futile cycle in the pathway or a limiting reaction. Finally, the authors proved experimentally, using labelled glucose [U–^{13}C], that a futile cycle occurred inside the pathway. Here the use of a high-resolution ultra-high performance LC, electrospray ionization, triple quadrupole (UHPLC-ESI-QqQ) system was crucial for quantitatively detecting the mass isotopomers of GlcNAc. The existence of a dissipation node in the phosphorylation/dephosphorylation between GlcNAc6P and GlcNAc was strongly responsible for the low titres of GlcNAc production. Finally, the authors identified in the *B. subtilis* genome a putative glucokinase that had not previously been annotated. The deletion of this gene greatly reduced the futile cycle and more than doubled the GlcNAc production in the engineered strain in a minimal media consuming glucose as carbon source, as was the initial goal of the work.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

6.3 MS Contribution to the Classic Example of the Semi-synthesis of the Anti-malarial Drug Artemisinin

One of the best examples of how MS can be powerful in prospecting novel biochemical pathways for drug discovery, and make use of this knowledge to engineer organisms to efficiently produce those high-value small molecules, is the case of the antimalarial drug artemisinin. The empirical formula of this C_{15} isoprenoid (sesquiterpene) was first determined by MS analysis after screening plant extracts used for combating the chill and fever symptoms of malaria.[24] Although artemisinin is the main component of the only effective treatment for malaria, its only natural source of production is the sweet wormwood plant *Artemisia annua*. Attempts to produce this sesquiterpene by agricultural supply or chemical synthesis were very costly and/or difficult to obtain in large scale,[25] making the manufacture of this drug unaffordable for use in developing countries.[26] The commercial production of this drug was only achieved by developments in synthetic biology and metabolic engineering, which have enabled the heterologous production of an artemisinin precursor at high titres in microorganisms.[25,26] The success of this technology was only achieved by reprogramming the chassis organism and introducing the genes encoding the key components of this biosynthetic pathway.

MS-based techniques were crucial in identifying the shortcomings and optimizing the system at several levels (Figure 6.2). To build the synthetic version of the artemisinic acid route, two isoprenoid biosynthetic pathways were used as a basis: the mevalonate and 1-deoxy-D-xylulose-5-phosphate pathways. Initially, *E. coli* was used as a chassis organism to generate high levels of amorphadiene, a metabolite obtained from the conversion of farnesyl diphosphate by amorphadiene synthase, the first committed step of artemisinin biosynthesis. To this aim, one of the strategies used was to introduce the heterologous expression of the yeast mevalonate pathway in *E. coli*, along with a codon-optimized version of amorphadiene synthase from *A. annua*. Despite the production of amorphadiene using this strain, the imbalance of the three enzymes needed for the conversion of acetyl-CoA to mevalonate led to growth inhibition due to the accumulation of an unknown intermediate.[18] GC-MS and LC-MS analysis revealed that the toxicity was caused by an imbalance in carbon flux, resulting in the accumulation of the intermediate 3-hydroxy-3-methyl-glutaryl-coenzyme A (HMG-CoA),[27] which inhibits fatty acid biosynthesis in the host, and, therefore, triggers membrane stress.[28] This limitation could be counteracted by the addition of palmitic and oleic acids into the growth media, which exemplifies the contribution of MS to identifying bottlenecks in strain engineering.

Although the heterologous production of artemisinic acid in yeast had a huge impact on the large-scale production of this anti-malarial drug, large volumes of sterile synthetic cultures in massive capacity bioreactors are still very costly. To meet the growing demand for this drug and its derivatives, an alternative could be production using a high-biomass crop growing in

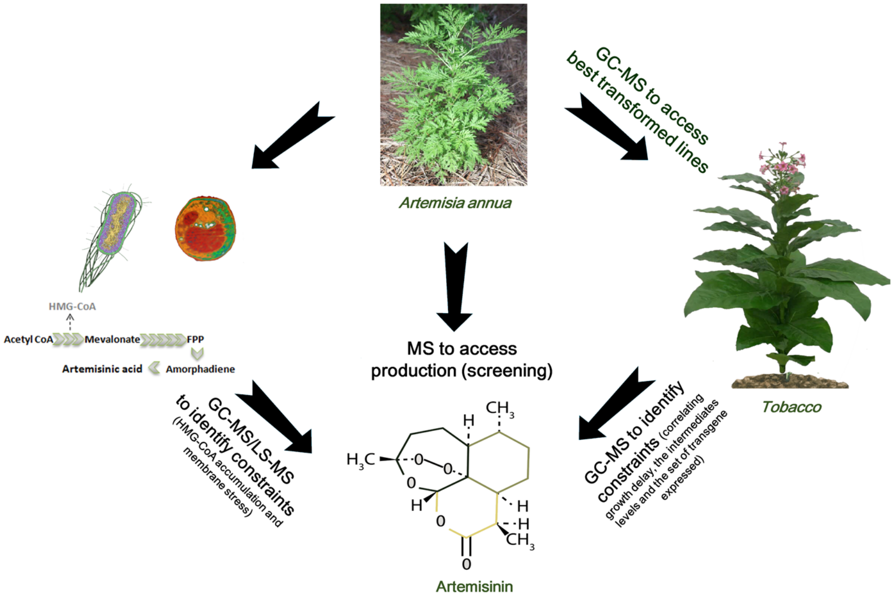
GC-MS to access best transformed lines
Artemisia annua
HMG-CoA
Acetyl CoA
Mevalonate
FPP
Artemisinic acid
Amorphadiene
MS to access production (screening)
GC-MS/LS-MS to identify constraints (HMG-CoA accumulation and membrane stress)
GC-MS to identify constraints (correlating growth delay, the intermediates levels and the set of transgene expressed)
Tobacco
H
CH3
H3C
O
O
O
H
O
H
H
CH3
O
Artemisinin

Figure 6.2 MS contribution to optimizing the biosynthetic route for the production of the anti-malarial drug artemisinin using recombinant microbes or engineered plants. MS-based techniques were crucial to identifying the shortcomings and optimizing the system at several levels (from screening of artemisinin production to identifying bottlenecks of engineered strains). Image of *A. annua*: credit to Lloyd Crothers. Creative Commons Attribution CC BY 2.0 (https://creativecommons.org/licenses/by/2.0/).

large territories. However, in order to integrate complete canonical pathways into a plant host there are a range of factors, such as codon optimization, genomic position effects and genetic instability, that make this process very challenging.[29] The prokaryotic nature of the plastid genome, as well as high expression with the absence of multiple promoters and simple stacking of multiple transgenes in synthetic operon, make it amenable to building synthetic pathways in plants for metabolic engineering.[29–31] Regarding biosafety, the maternal organelle inheritance in most crops presents an additional advantage for the genetic manipulation of plastids. Recently, a synthetic approach enabled the introduction of an artemisinic acid biosynthesis pathway into the tobacco chloroplast genome, allowing the production of over 120 mg kg^{-1} fresh weight of artemisinic acid in this fast-growing crop at very low cost.[32] In this approach, GC-MS-based metabolite profiling enabled not only the screening of a large population of supertransformed tobacco lines containing the entire biochemical pathway for artemisinic acid production, but also identifying constraints for the synthesis of this metabolite. A tremendous challenge in establishing assays for each pathway intermediate is the absence of chemical standards, the presence of isomers and rapid turnover.[10] In order to accurately target all the intermediates of the artemisinic acid pathway, different strategies, including extraction and analytical methods were used according to the chemical nature of the metabolite.[32] Volatile compounds such as amorpha-4,11-diene, artemisinin and its degradation products were isolated with the use of headspace, whereas lipophilic saponification compounds were extracted using a KOH:methanol/hexane buffer. Metabolites were identified based on the mass spectral intensity of specific and selective mass fragments based on reference substances obtained from genetically engineered yeast strain cultures.

The generation of the combinatorial supertransformation of transplastomic recipient lines (COSTREL) was performed in multi steps in which the selection of best-performing lines was assessed by GC-MS. First, four homoplasmic transplastomic lines were transformed with the synthetic artemisinic acid operon constructs containing different arrangements of the four core enzymes (FPS, ADS, CYP and CPR). MS screening allowed not only the selection of the best transplastomic lines producing higher levels of artemisinic acid, but also the unravelling that the pale-green and attenuated growth delay phenotype of two of the transplastomic lines was attributed to a higher conversion rate of amorpha-4,11-diene to downstream metabolites. Interestingly, investigation of the molecular mechanism underlying this phenotype revealed that the relative orientation of two operons led to a higher CYP/CPR ratio, resulting in a more efficient use of the redox power for artemisinic acid synthesis.

To maximize the artemisinic acid production, enzymes known to enhance the flux through the pathways (CYB5, ADH1, ALDH1, DBR2 and DXR) in *A. annua* were introduced into the best-performing transplastomic plants by COSTREL. Apart from the identification of a line with a 77-fold increase in artemisinic acid levels compared to the best performing transplastomic line, the MS-based method allowed the detection of the limiting steps for this drug production by correlating the levels of intermediates, artemisinic acid and the set of transgenes expressed in the nucleus. The results revealed that ALDH1 causes the greatest effect in enhancing the pathway flux as well as the efficiency of oxidation of artemisinic alcohol, which serve as a key bottleneck for this drug production, enabling the maximization of the metabolic output.

6.4 Conclusion

There are an increasing number of studies that use MS at some stage for improving routes of synthesis, with widely diversified goals such as the production of isobutyl acetate,[33] advances in the production of isopentenol and other alcohols of five carbons,[34] and the use of MS to improve polyketide production, the adipic acid.[35] It is evident that the limits of the contribution that MS can give to synthetic biology are far from being reached. 13% of current world trade (US$2 trillion) is based in biology or biosciences, and includes forestry, food, bioenergy, biotechnology and green chemistry products. More than 40 countries have committed to boosting the fraction of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00159

their economies that is based in bio-economy.[36] In this context of the integration of omic sciences, MS presents itself as a great resource to assist the development of science.

Acknowledgements

The work in CC laboratory was partly supported by Max Planck Society and grant #2012/19561-0, São Paulo Research Foundation (FAPESP). VCHS was supported by CAPES Scholarship Number (CAPES/CNPEM n° 24/2013).

References

1. A. A. Cheng and T. K. Lu, *Annu. Rev. Biomed. Eng.*, 2012, **14**, 155–178.
2. S. Rollié, M. Mangold and K. Sundmacher, *Chem. Eng. Sci.*, 2012, **69**, 1–29.
3. S. A. Benner and A. M. Sismour, *Nat. Rev. Genet.*, 2005, **6**, 533–543.
4. R. Breitling and E. Takano, *Curr. Opin. Biotechnol.*, 2015, **35**, 46–51.
5. V. Chubukov, A. Mukhopadhyay, C. J. Petzold, J. D. Keasling and H. G. Martín, *NPJ Syst. Biol. Appl.*, 2016, **2**, 16009.
6. E. Wurtzel and T. Kutchan, *Science*, 2016, **353**, 1232–1236.
7. M. T. Henke and N. L. Kelleher, *Nat. Prod. Rep.*, 2016, **33**, 942–950.
8. H. Zaehner and H. Fiedler, in *Fifty Years of Antimicrobials: Past Perspectives and Future Trends*, ed. N. J. Russell, Cambridge University Press, 1995, pp. 67–84.
9. D. E. Gillespie, S. F. Brady, A. D. Bettermann, N. P. Cianciotto, M. R. Liles, M. R. Rondon, J. Clardy, R. M. Goodman and J. Handelsman, *Appl. Environ. Microbiol.*, 2002, **68**, 4301–4306.
10. C. J. Petzold, L. J. G. Chan, M. Nhan and P. D. Adams, *Front. Bioeng. Biotechnol.*, 2015, **3**, 135.
11. D. Grote-Koska, S. Czajkowski and K. Brand, *Ther. Drug Monit.*, 2015, **37**, 400–404.
12. M. C. Wilson and J. Piel, *Chem. Biol.*, 2013, **20**, 636–647.
13. Y. Liu, H. Link, L. Liu, G. Du, J. Chen and U. Sauer, *Nat. Commun.*, 2016, **7**, 11933.
14. J. E. Krebs, B. Lewin, E. S. Goldstein and S. T. Kilpatrick, *GENES XI*, Jones & Bartlett Learning, 2014.
15. A. M. Redding-Johanson, T. S. Batth, R. Chan, R. Krupa, H. L. Szmidt, P. D. Adams, J. D. Keasling, T. Soon Lee, A. Mukhopadhyay and C. J. Petzold, *Metab. Eng.*, 2011, **13**, 194–203.
16. P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon and R. Aebersold, *Cell*, 2009, **138**, 795–806.
17. V. Lange, P. Picotti, B. Domon and R. Aebersold, *Mol. Syst. Biol.*, 2008, **4**, 222.
18. V. J. Martin, D. J. Pitera, S. T. Withers, J. D. Newman and J. D. Keasling, *Nat. Biotechnol.*, 2003, **21**, 796–802.

19. K. W. George, A. Chen, A. Jain, T. S. Batth, E. E. Baidoo, G. Wang, P. D. Adams, C. J. Petzold, J. D. Keasling and T. S. Lee, *Biotechnol. Bioeng.*, 2014, **111**, 1648–1658.
20. B. J. Rutherford, R. H. Dahl, R. E. Price, H. L. Szmidt, P. I. Benke, A. Mukhopadhyay and J. D. Keasling, *Appl. Environ. Microbiol.*, 2010, **76**, 1935–1945.
21. T. Hasunuma, T. Sanda, R. Yamada, K. Yoshimura, J. Ishii and A. Kondo, *Microb. Cell Fact.*, 2011, **10**, 2.
22. H. L. Meng, Z. Q. Xiong, S. J. Song, J. Wang and Y. Wang, *Biotechnol. J.*, 2016, **11**, 530–541.
23. Y. Liu, Y. Zhu, J. Li, H.-d. Shin, R. R. Chen, G. Du, L. Liu and J. Chen, *Metab. Eng.*, 2014, **23**, 42–52.
24. D. Klayman, *Science*, 1985, **228**, 1049–1055.
25. C. J. Paddon and J. D. Keasling, *Nat. Rev. Microbiol.*, 2014, **12**, 355–367.
26. M. Peplow, *Nature*, 2016, **530**, 389–390.
27. D. J. Pitera, C. J. Paddon, J. D. Newman and J. D. Keasling, *Metab. Eng.*, 2007, **9**, 193–207.
28. L. Kizer, D. J. Pitera, B. F. Pfleger and J. D. Keasling, *Appl. Environ. Microbiol.*, 2008, **74**, 3229–3241.
29. W. Liu and C. N. Stewart Jr, *Trends Plant Sci.*, 2015, **20**, 309–317.
30. L. B. Scharff and R. Bock, *Plant J. Cell Mol. Biol.*, 2014, **78**, 783–798.
31. Y. Lu, H. Rijzaani, D. Karcher, S. Ruf and R. Bock, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E623–E632.
32. P. Fuentes, F. Zhou, A. Erban, D. Karcher, J. Kopka and R. Bock, *eLife*, 2016, **5**, e13664.
33. Y. Tashiro, S. H. Desai and S. Atsumi, *Nat. Commun.*, 2015, **6**, 7488.
34. K. W. George, M. G. Thompson, A. Kang, E. Baidoo, G. Wang, L. J. G. Chan, P. D. Adams, C. J. Petzold, J. D. Keasling and T. Soon Lee, *Sci. Rep.*, 2015, **5**, 11128.
35. A. Hagen, S. Poust, T. d. Rond, J. L. Fortman, L. Katz, C. J. Petzold and J. D. Keasling, *ACS Synth. Biol.*, 2016, **5**, 21–27.
36. B. El-chichakli, J. von Braun, C. Lang, D. Barben and J. Philip, *Nature*, 2016, **535**, 221–223.

CHAPTER 7

Studying Enzyme Mechanisms Using Mass Spectrometry, Part 1: Introduction

CRISTINA LENTO[a], PETER LIUNI[a] AND DEREK J. WILSON*[a,b,c]

[a]Department of Chemistry, York University, Toronto, M3J 1P3, Canada; [b]Centre for Research of Biomolecular Interactions, York University, Toronto, M3J 1P3, Canada; [c]Centre for Research in Mass Spectrometry, York University, Toronto, M3J 1P3, Canada
*E-mail: dkwilson@yorku.ca

7.1 Introduction

Enzymes are biological molecules (proteins) responsible for the catalysis of reactions in all living organisms. Their remarkable ability to enhance the rate of biological reactions by many orders of magnitude allows for select kinetically controlled processes to occur on biologically relevant time scales.[1,2] Often the altered or absent activity of an enzyme is used for diagnostic purposes. For example, the presence of certain enzymes in the blood might indicate tissue-specific damage in the liver, pancreas, or cardiac muscles.[3,4] In addition to enabling life, enzymes are of growing importance in the food, agricultural and pharmaceutical industries.[5] Enzyme mechanisms can range from relatively simple two-step processes to complex multi-step reactions. Although great strides have been taken since the start of the

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

nineteenth century to better understand the function and mechanisms by which enzymes operate, limitations to the most widely used analytical methods must be overcome.

Enzymes are proteins, composed of amino acid chains that fold into biologically functional 'native' three-dimensional structures.[6] Enzymes can vary substantially in size; one of the smallest known subunits is a bacterial 4-oxalocrotonate tautomerase, composed of only 62 amino acid residues,[7,8] while catalase is one of the largest and in humans is composed of four 60 kDa subunits.[9] Some enzymes require the binding of non-protein components, called cofactors, in order to become active. Cofactors are small molecules that can range from inorganic ions to more complex organic or metallo-organic molecules. When a cofactor is tightly or covalently bound to the enzyme it is referred to as a prosthetic group. An active enzyme with a bound cofactor is termed a holoenzyme, whilst an enzyme lacking its cofactor is an apoenzyme.[10]

Typically, only a few amino acids within an enzyme are responsible for carrying out catalysis through direct interactions with the substrate in the active site. The particular structure and chemistry of the active site is what confers the ability to act on specific substrates in a noisy biochemical background. Several models have been proposed to explain this specificity. It was first believed that substrates would fit perfectly into the active site of an enzyme without having to undergo a change in structure, this is referred to as the 'lock and key' model.[11] However, the induced fit model proposes that conformational changes occur upon binding of the substrate, which has led to the recognition that inherent dynamics play a critical role for substrate recognition and catalysis (Figure 7.1).[11] Proteins are not static structures, and can populate various conformational ensembles. Conceptualizing ligand binding as affecting distal sites of the protein (and *vice versa*) through conformational change also allows for allosteric regulation of enzymatic activity.[12] Arguably the dominant current model for substrate recognition and binding is the 'conformer selection' model, where resting and active-state dynamics are identical to the substrate selecting the appropriate conformer for binding. In this model, the rate limiting step of binding is the adoption by the enzyme of a particular 'substrate ready' structure within the ensemble of structures populated in the native state.[13,14]

During the first few milliseconds of an enzymatic reaction, the enzyme–substrate complex is formed producing a significant population of intermediates while product concentrations are low; this is referred to as the pre-steady state.[15] Directly studying pre-steady state kinetics is quite challenging, as millisecond-or-better time scale detection is required. Following binding of the substrate, the reactant molecules must have sufficient thermal energy to cross a dominant free energy barrier along the reaction coordinate. This energy barrier is referred to as the transition state and the fundamental role of enzymes is either to directly lower the energy of the transition state or to provide an alternate reaction coordinate with a lower barrier.[16] The enzyme–substrate complex undergoes turnover to yield

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

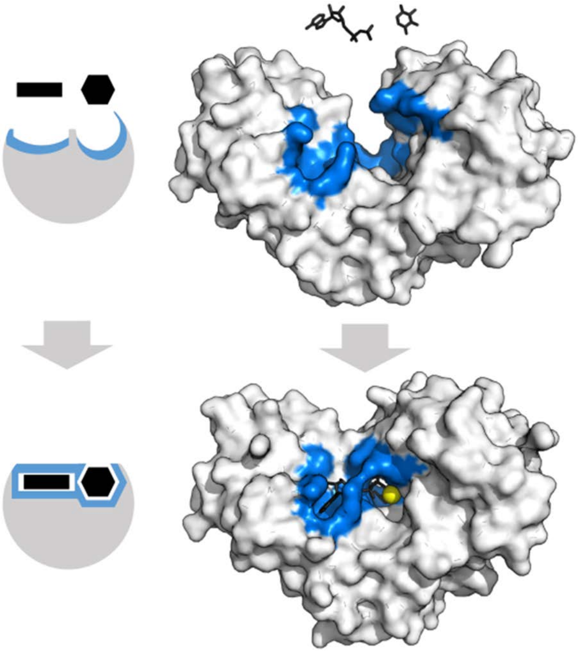

Figure 7.1 Induced fit model upon substrate binding. When substrates (black) adenosine triphosphate (ATP) and xylose attach to the binding sites (blue), the hexokinase enzyme undergoes a large induced fit motion that closes over the substrates. The Mg^{2+} cofactor (yellow) is required to produce the active form of ATP (PDB IDs: 2E2N, 2E2Q). Image by Thomas Shafee, CC BY.

enzyme–product complexes and, ultimately, free enzyme and product to complete the cycle. It is a common practice to supply an excess amount of substrate to the enzyme, allowing the reaction to quickly achieve a pseudo-equilibrium of enzyme–substrate and enzyme–product complexes. This is referred to as the steady state and is more easily studied because it can be made to persist for an indefinite period, provided the substrate is not significantly depleted (Figure 7.2).[15]

Several factors affect catalytic activity, including pH, temperature, enzyme concentration, substrate concentration and the presence of activators or inhibitors.[17] As the names suggest, an activator increases the activity of an enzyme while an inhibitor decreases its activity. Many drug treatments involve the use of competitive inhibitors to prevent substrate access, for example to abolish the activity of pathogenic enzymes responsible for breaking down antibiotics leading to multi-drug resistance, or for the adjustment of metabolic imbalances (Figure 7.3).[18,19] Enzyme-targeted inhibitor design is a long-standing area of research in the pharmaceutical industry and requires a detailed understanding of how catalysis occurs at the enzyme's active site.[20]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

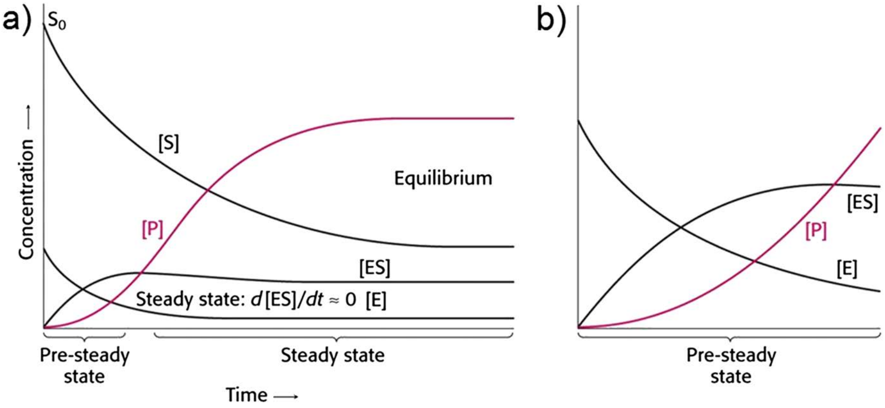

Figure 7.2 Changes in concentration of reaction participants of an enzyme-catalyzed reaction with time. (A) The pre-steady state is relatively short-lived and occurs at the initial stages of the reaction. At equilibrium, or steady state conditions, there is no net change of enzyme–substrate [ES], free enzyme [E], substrate [S], and product [P]. (B) A closer look at the pre-steady state where the [ES] begins to form, and there is a significant population of intermediates leading up to the formation of [P]. Adapted from ref. 15.

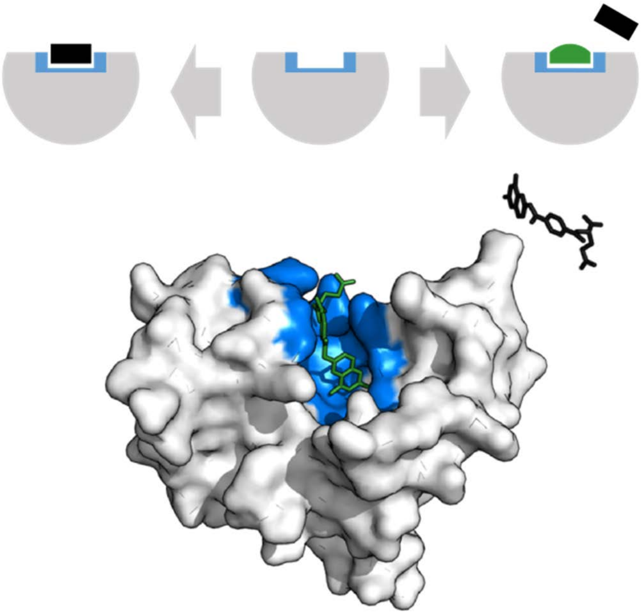

Figure 7.3 An enzyme binding site (blue), which would normally bind the substrate (black) can alternatively bind a competitive inhibitor (green) to prevent substrate access. In this example, dihydrofolate reductase is inhibited by methotrexate, which prevents binding of its substrate, folic acid (PDB ID: 4QI9). Image by Thomas Shafee, CC BY.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

7.2 Methods for Studying Enzyme Mechanisms

Any single technique capable of the detailed exploration of enzyme mechanisms must be exceptionally versatile. Firstly, the simultaneous detection and identification of reaction intermediates is essential for a detailed characterization of reaction pathways. As mentioned earlier, characterization of enzymatic pre-steady states has proven challenging over the years due to their transient nature, which often persists for less than one second. Therefore, time-resolved techniques are needed to detect intermediates under pre-equilibrium conditions. This implies that the analysis begins within milliseconds (or microseconds) following the start of the reaction. The capacity to meet this requirement has improved over the last several decades thanks to the development of rapid mixing devices. Kinetic information is also useful in elucidating possible mechanisms for a reaction, which can be used to understand a variety of disease states.[21] Monitoring the structural changes that occur at the protein level during catalysis is also of great importance, revealing critical insights such as allosteric effects in substrate and/or inhibitor binding.[22] Ideally, molecular snapshots of the protein converting substrate to product must be obtained in order to gain a full mechanistic understanding. We will discuss the most used analytical techniques for the study of enzymatic reactions and the advancements that have been made in the field of mass spectrometry (MS) over the past several decades, making it suitable for the study of various unique systems.

7.2.1 X-Ray Crystallography

To understand how enzymes function it is important to know their three-dimensional ground-state structures. One of the two ‘classical’ techniques for high-resolution structure determination is X-ray crystallography. Briefly, this technique involves the diffraction of X-ray beams passing through a protein crystal. The diffraction pattern produced by a given protein can be interpreted using Bragg’s law to produce an electron density map, which is then used to map the most thermodynamically favoured conformation.[23] This technique has the ability to offer atomic-level resolution of an enzyme allowing for the identification of active sites and binding pockets. In addition to crystallization of the enzyme in isolation, co-crystallization of the enzyme along with its substrate (or more often, a competitive inhibitor) can be used to study the geometry of binding.[24] In more recent years, the use of ‘time-resolved’ studies have helped elucidate the details of catalytic mechanisms and understand the sequence of steps and residues involved in each step.[25] It has been observed that some enzymes are capable of undergoing catalytic turnover in the crystalline state, even in conditions that sometimes consist of high pH, viscosity and ionic strength.[26]

To trap reaction intermediates for crystallography, both physical and chemical techniques are used. This often involves changing the pH,[27] lowering the temperature in order to slow down intermediate turnover,[28] or the addition of

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

an inhibitor.[29] These methods have been applied to various systems including several serine-protease acyl–enzyme complexes.[30,31] High-resolution crystal structures have the ability to shed light on many chemical parameters, including planarity of peptide bonds, bond lengths, bond angles, and torsion angles.[32] However, there are several limitations that must be considered. First and foremost, high quantities of the target enzyme (milligrams) are required for the crystallization process to successfully occur. Secondly, incubation with a substrate often does not yield co-crystallization of intermediate (or even substrate bound) species. In the event that crystallization successfully occurs, the way in which X-ray diffraction data is collected and refined has a strong impact on resolution.[32] Atomic resolution (typically <2.0 Å) is required to identify subtle changes occurring in the substrate. Many proteins simply do not diffract due to high water content, which decreases protein–protein interactions and increases the overall mobility of the protein. Therefore, it is often extremely challenging to produce crystals of sufficient quality for substrate co-crystallization studies, even with extensive optimization. Once suitable crystals are obtained, there is no guarantee that the induced packing forces have not distorted the natural structure of the enzyme and/or its substrate. In addition, the diffraction process itself can be relatively harsh, and might cause low-barrier intermediates to decay, resulting in structures that are not biologically relevant. There is therefore a growing need to verify what data has been obtained from X-ray crystallography with other experimental methods.

7.2.2 Site-directed Mutagenesis

Protein engineering comprises the modification of an enzyme's primary sequence in order to study the effects on structure and function. This often involves site-directed mutagenesis of one or more amino acid residues and is most widely used in catalysis to identify the critical active site residues.[33] The premise is that exchange of a single catalytically important amino acid residue with another of differing chemical properties should abolish some, if not all, of the enzyme's activity, whereas similar modification of a non-essential amino acid should have no effect. This method can be used to identify the amino acids responsible for binding of the substrate, and is often used to complement the knowledge obtained from X-ray crystallography, particularly in the (common) case where co-crystallization proves difficult or uninformative.[34] Considerable caution must be taken in the interpretation of mutagenesis analyses, however, as in some cases even a single point mutation outside of the active site can alter enzymatic activity through disruption of structure or dynamics.

7.2.3 Optical Methods

One of the first, and still widely-used, approaches for studying enzyme mechanisms is UV–visible spectrophotometry, in which enzymatic activity can be monitored by the appearance of a chromophoric product (or the

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

disappearance of a chromophoric substrate). In rare cases, enzymatic intermediates may have distinctive absorbance properties, making it possible to detect pre-steady state species. Fluorescence spectroscopy is also a commonly used method, with the advantage of greatly enhanced sensitivity, and even the ability to perform single-molecule measurements or distance measurements using Forster resonance energy transfer.[35] Of course, the major drawback of both of these approaches is that they require a chromophore (or fluorophore), which, if not endemic to the system, must be artificially introduced. Nonetheless, optical methods have a major advantage in terms of duty cycle, with individual measurements requiring microseconds or less of acquisition. This is of particular use in rapid time-course measurements, such as are required for pre-steady state studies.

Where sub-second reaction times are to be monitored, a rapid mixing apparatus is implemented to efficiently mix the solutions of interest, initiating the chemical reaction and mimicking the biological process that occurs *in vivo*. Rapid mixers coupled to optical detection methods can be classified as either stopped-flow or continuous-flow.[36,37] Classically, stopped-flow mixers flush the reactants at high flow rates through the mixer and observation cell. Once enough new volume has entered the cell, displacing the contents from previous experiments, the flow is stopped and observation begins.[38] In continuous-flow mixers, the solutions are pumped at high flow rates through two different channels, which then coalesce and mix within a small mixing chamber.[39–41] The stopped-flow technique is often preferred when coupled to spectroscopic methods such as UV–visible fluorescence absorbance.

Optical detection by UV–visible absorbance or fluorescence is suitable for the study of enzyme reactions thanks to its high sensitivity (though this depends ultimately on the chromo/fluorophore being detected). In addition, the high flow rates used to flush reactants between experiments allows for a time resolution of microseconds under ideal conditions.[40] However, there are several limitations to this method that must be considered. Perhaps the biggest disadvantage is low selectivity, that is, the inability to study more than a few species simultaneously. Most optical studies are limited to detecting one or two species with the additional assumption that chromophores are available for the system being analyzed. This is not ideal for unravelling complex, multistep reactions with a large number of reactive species. In addition, the attachment of a chromophore to produce artificial substrate analogues raises concerns, as it might alter kinetics compared to the natural substrate.[42]

7.2.4 Isothermal Titration Calorimetry (ITC)

ITC measures the heat change over time of the binding between an enzyme incubated with its small molecule substrate.[43,44] It can be used to measure kinetic parameters such as the catalytic rate constant (k_{cat}) and the Michaelis constant (K_{M}).[45] The ITC instrument is a heat-flux calorimeter, and measures the amount of power required to keep a constant temperature difference

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

between the sample and reference cell. In the context of enzymes, it is used to measure enzymatic efficiency and allows for the kinetic characterization of enzyme inhibition.[46] It is also an important tool used for drug screening as it reveals the thermodynamics of binding, which can help elucidate binding affinities.[47,48] Whilst being a powerful method for the characterization of biomolecular interactions in the solution-phase, one of its biggest drawbacks is the need for large sample volumes.

7.2.5 Two-dimensional Nuclear Magnetic Resonance (2D NMR) Spectroscopy

In contrast to crystallography, solution NMR spectroscopy offers the ability to measure high resolution protein structures at physiological ionic strengths and concentrations. Relatively low concentrations are needed for analysis on modern instruments, and thus can be used to study proteins at low μM concentrations. A key advantage of NMR is its versatility; there are a host of experiments, each covering structure and dynamics analysis, across a broad range of biologically relevant timescales, from seconds to hours. 2D NMR spreads severely overlapping spectra into two orthogonal frequency dimensions, allowing for easier analysis of protein samples.[49] In the context of enzymes, 2D measurements can reveal changes in local chemical structure that accompany ligand or substrate binding, which can be used for the elucidation of active sites, among other applications.[50] In many cases, NMR has sufficient sensitivity and selectivity to detect minor equilibrium species, allowing for the study of protein motions during catalytic turnover such as those related to allostery.[51,52] Of particular importance are Carr–Purcell–Meiboom–Gill relaxation dispersion experiments, which allow for determination of the rate of interconversion between ‘ground-state’ species and conformations linked to the catalytic reaction coordinate.

NMR can also be combined with hydrogen–deuterium exchange (HDX), a solution-phase, structure-dependent labeling technique that involves the substitution of labile amide hydrogens on the backbone of proteins for deuterium from solvent. This method can be used to study enzyme structure and dynamics during catalytic turnover, providing a site-specific map of exchange rates (corresponding to the extent of secondary and tertiary structure) for individual amino acids.[53] However, NMR measurements do suffer from a set of inherent drawbacks, not the least of which is obtaining isotopically labeled protein and the high cost of purchasing and maintaining the equipment. Acquisition times are exceptionally long, with a typical 2D spectrum requiring several hours of measurements (a number of methods exist to reduce this time considerably, but most negatively affect sensitivity and/or resolution).[54] This can make it impossible to perform real-time measurements, and effectively rules out direct pre-steady state studies in the vast majority of enzymes. There is also a fundamental limit on analyte size, generally precluding complex analyses on biomolecules with masses >75 kDa.[55]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

7.2.6 Mass Spectrometry (MS)

Over the past three decades, MS has become a powerful tool for the structural characterization of biomolecules. The use of this technique was initiated by J. J. Thomson's work in 1897, who demonstrated the existence of the electron and measured the first ever charge-to-mass ratio (now usually expressed as the mass-to-charge ratio m/z).[56] MS has the ability to precisely determine the m/z of an ionized analyte *via* manipulation with electric and/or magnetic fields. Over the past 100 years, improvements have been made to ionization techniques, allowing for the study of ever larger molecules, leading ultimately to the development of 'soft' ionization through which even non-covalent complexes can be transferred intact into the gas phase.[57] The two dominant soft ionization methods, electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) have greatly expanded the analytical domain of MS to include direct measurements of biomacromolecules in action.[58] MS's greatest strength is its inherent ability to simultaneously distinguish (and sometimes quantify) a large number of species from complex mixtures.[59] This is a great advantage when studying enzyme mechanisms as it allows for the possibility of detecting multiple intermediates in substrate turnover.

The ESI process occurs in two distinct steps. First, the liquid sample forms charged droplets as it is ejected from a narrow capillary held at high electric potential (typically 3–5 kV).[60,61] These droplets then progressively decompose *via* solvent evaporation (assisted by N_2 nebulizing gas) and coulombic fission to form multiply charged ions.[62] As ESI-MS measures analytes in the gas phase, a key concern is if it is reflective of what occurs in solution-phase, especially since the distribution of charges acquired by proteins during ESI, called the charge state distribution, is widely accepted to reflect the global protein structure in solution (*i.e.*, unfolded proteins acquire more charge and a wider distribution of charges than folded proteins when they are transferred to the gas phase by ESI). For non-covalent protein complexes, the direct connection to solution is less clear. Solution-phase droplets shrink when undergoing the ESI process, thus changing the concentration of the analytes and possibly causing a shift in equilibrium. Of even greater concern is the formation of non-specific complexes when ESI droplets contain two or more protein molecules just prior to transfer to the gas phase. On the other hand, solution protein complexes might be disrupted if ESI conditions are harsh. Many studies have addressed this question, with the general consensus being that ESI-based detection of complexes (or lack thereof) should be taken with a degree of caution, but that with careful controls, even accurate quantitation of complexes is often possible.[63–65]

While already regarded as a soft ionization technique, advancements have been made over the last several decades that have increased the probability of preserving solution non-covalent protein complexes through ESI. One straightforward approach is to lower the flow rate. Typically, the ESI process usually requires flow rates that are in the μL min^{-1} range, with initial droplets

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

having diameters in the μm range.[66] The development of nanoESI, essentially a minimized-flow ESI with flow rates in the nL min^{-1} range and initial droplets having diameters in the nm range, allows for extremely low sample consumption and a shorter ion generation process, allowing for improved intact analysis of non-covalent complexes.[67–70] More recently, Takáts *et al.*[71] introduced electrosonic spray ionization (ESSI), a microESI source with a supersonic nebulizing gas at high linear velocities. Compared to ESI, the study found that the narrow charge state distributions observed with ESSI are more likely to correspond to native-like solution-phase structures, thus preserving protein complex structures.[71] ESSI was successfully applied to study the formation of enzyme–substrate and enzyme–substrate–inhibitor complexes in the gas phase.[72] A study by Jecklin *et al.*[73] compared the three methods of ESI, nanoESI and ESSI of the model system hen egg white lysozyme binding to *N*,*N*′,*N*″-triacetylchitotriose (NAG3) and found that all three yielded K_D (equilibrium dissociation constant, a measure of how tightly a target binds its substrate or ligand) values that were in agreement with solution-phase literature values. Deciding which of these methods is preferable will depend on the system under study, as some complexes require 'softer' conditions than others.

Collectively, MS has several unique advantages for the investigation of enzymatic reactions. By measuring *m*/*z*, a property of any ionizable species, the need for chromophoric labeling or immobilization of one of the binding partners, a common practice in other biochemical methods, can be bypassesed.[74] Since the vast majority of enzymatic reactions involve a change in mass, natural substrates can be used for the independent, simultaneous detection of intermediates and products during catalytic turnover. As mentioned previously, this is of great importance as there can be significant changes in kinetics between natural and chromophoric substrates.[75] In addition, the large *m*/*z* range allows for the simultaneous study of changes occurring in both enzyme and substrate during the course of the reaction. The low sample consumption for mass spectrometric analysis also allows for the characterization of proteins or ligands available at extremely low concentrations. In the following sections, we will discuss the development of some MS-related techniques for the study of enzymes.

7.2.6.1 Time-resolved ESI MS (TRESI-MS)

Stopped-flow and continuous-flow rapid mixing have been combined with the power of MS in order to obtain 'time-resolved' measurements of (bio) chemical reactions. TRESI-MS allows for kinetic experiments to be carried out with a dead time ranging from a few milliseconds to seconds.[76,77] Stopped-flow ESI-MS was first implemented by Kolakowski *et al.*, where the optical cell was replaced by a reaction capillary. However the open system caused considerable pressure when switching from the initial high-flow-rate phase to the low-flow-rate injection into the ESI source, ultimately having a negative impact on time resolution.[78] This quickly shifted the focus to continuous-flow capillary-based devices, which allowed for decreased dead times and increased time resolution. Coupling of a time-resolved continuous-flow capillary mixer to MS has made it possible to overcome many of the challenges commonly faced with other analytical techniques for the study of pre-steady state enzyme kinetics.

One of the earliest continuous-flow rapid mixers consisted of a static mixing tee followed by a reaction capillary whose fixed internal volume regulated the reaction time (Figure 7.4A).[79] Different time points were obtained by 'switching out' the reaction capillary to differing lengths or varying the flow rates, allowing one to track a reaction over time. This continuous-flow setup was subsequently improved with the development

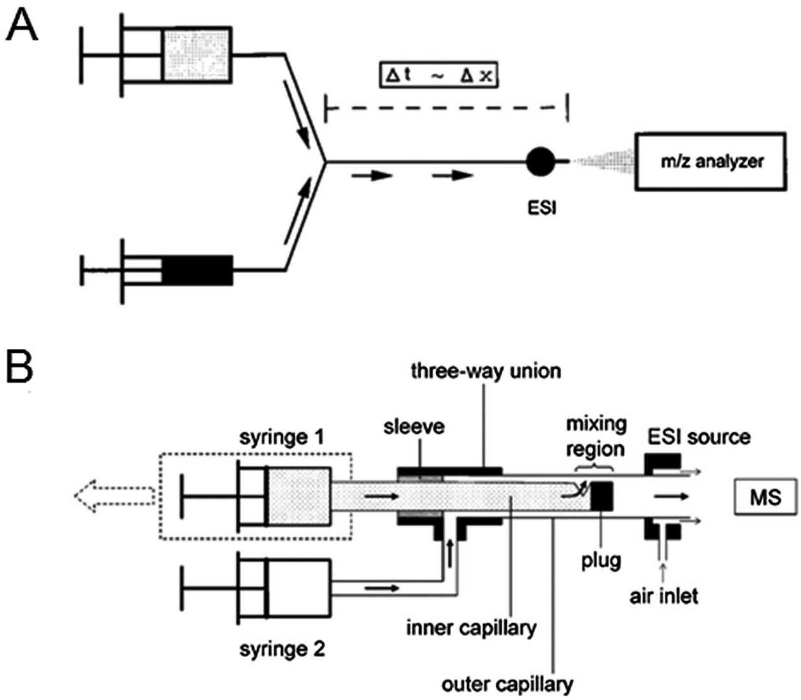

Figure 7.4 Schematic depictions of continuous-flow setups used for TRESI-MS. (A) A fixed mixing tee where the reaction time (Δ*t*) is determined by the length of the reaction capillary (Δ*x*). Arrows indicate the direction of flow from syringes carrying the desired reactants.[79] Reprinted with permission from L. Konermann, B. A. Collings and D. J. Douglas, *Biochemistry*, 1997, **36**, 5554–5559. Copyright 1997 American Chemical Society. (B) A capillary mixer with adjustable reaction volume. Two concentric capillaries are injected with reactants from syringes 1 and 2. A notch is made 2 mm from the end of the inner capillary allowing the two solutions to mix. Changing the position of the inner capillary within the outer capillary changes the volume between mixing and ESI detection allowing for the continuous tracking of a reaction over various time points. Reprinted with permission from D. J. Wilson and L. A. Konermann, *Anal. Chem.*, 2003, **75**, 6408–6414. Copyright (2003) American Chemical Society.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

of an adjustable reaction chamber able to measure multiple time points in a single experiment (Figure 7.4B).[80] In this set-up, the reactants flow through concentric capillaries, allowing for rapid mixing from a notch cut 2 mm from the distal end of the inner capillary. When the inner capillary is lined up with the outer capillary, observable reaction times can reach as low as 8 ms. Later reaction times can be monitored by withdrawing the inner capillary within the outer capillary, increasing the volume between the mixing point and ESI detection. Reactions can be monitored in two modes: (1) spectral mode, where the reaction volume is held at specific points and allows for longer acquisitions at specific time points, and (2) continuous mode, where the reaction volume is increased at a constant rate.[81]

While many of the previously discussed analytical techniques focus on studying steady state parameters such as k_{cat} and K_M, they do not provide information regarding sub-states that contribute to the mechanism. TRESI-MS is uniquely suited to monitoring enzyme reactions because it is able to detect subtle changes in substrate, intermediate, and product populations as a function of reaction time and m/z ratio.[82] Revealing the mechanistic details of catalysis also depends on knowing the positions of key functional groups in the active site during the transition state.[83] Consequently, TRESI-MS has been applied for the elucidation of kinetic isotope effects (KIEs). A KIE is the change in rate of a chemical reaction following the substitution of one atom in the reactant for one of its isotopes.[84,85] KIEs can give insights into the rate-limiting steps[86,87] and transition state structures[16,88,89] from quantifiable changes in reaction rates. Differences in vibrational frequencies of the heavy and light bonds alter their zero point energies between the ground and transition state, with the heavy isotope having a lower energy compared to the light. As a result, their activation energies will be different and are reflected in their measured rates where transition-state bond lengths and geometry may be inferred. The ability of MS to distinguish multiple isotopes simultaneously and with high sensitivity makes it particularly suitable for the study of isotope effects. In addition, the study of KIEs in the pre-steady state using TRESI-MS can provide further mechanistic insights into catalytically active functional groups.[85,90]

7.2.6.2 *Determination of Binding Constants and Allostery in Multimeric Enzymes*

All processes within a cell must be carefully controlled in order to carry out essential physiological processes. Many enzymes operate *via* allosteric regulation, whereby the binding of an effector molecule or ligand away from the active site induces conformational change allowing for regulation of catalytic activity by altering k_{cat}, K_M, or both.[91] Cooperative binding often operates through an allosteric mechanism, and occurs in enzymes with multiple binding sites where the affinity towards a second ligand increases (or decreases) upon binding of the first ligand.[92] Although not an enzyme, hemoglobin's oxygen binding ability was the first observed example of cooperativity.[11,93] It was this model system that helped reinforce the understanding that conformational changes within proteins, especially those that are multimeric in nature, could be responsible for the precise regulation of function.

Classically, allosteric regulation of protein–ligand complexes have been studied *via* X-ray crystallography or NMR spectroscopy.[94] However, ESI-MS has become an increasingly utilized tool for the characterization of non-covalent protein–ligand interactions including the determination of dissociation constants, and will be discussed in the next chapter.[74,95–99] As mentioned previously, these parameters are of great importance for small molecule drug development and the understanding of complex biochemical pathway regulation.

7.2.6.3 Hydrogen–Deuterium Exchange Coupled to MS for Studying Catalysis-linked Dynamics

The use of HDX to study protein conformations has classically been applied in 2D NMR. However, coupling of this method to MS has a number of advantages, in particular that its sensitivity allows for the analysis of protein at concentrations in the nM–μM range, and has a virtually unlimited protein size range.[100] The HDX reaction can be acid (H_3O^+), base (OH^-), and water (H_2O) catalyzed, however, effects from water catalysis is minimal.[101,102] The chemical HDX rate constant (K_{ch}) of an amide in an unstructured peptide is as follows:

$$K_{ch} = k_{int,H}[H^+] + k_{int,OH}[OH^-] + k_{int,H_2O}[H_2O]. \tag{7.1}$$

At physiological pH, HDX is base-catalyzed, where abstraction of the amide proton by deuteroxide is followed by deuteration of the amide nitrogen by D_2O.[103] The availability of labile amide hydrogen to undergo exchange is determined by several factors: hydrogen bonding, solvent accessibility, pH and temperature. When the pH and temperature during analysis of a sample is kept constant, the rate of exchange is dependent on the former two factors. Stable intramolecular hydrogen bonding networks of backbone amides are found in secondary structural elements such as alpha helices and beta sheets, and these typically exhibit drastically reduced HDX rates compared to non-hydrogen bonded regions such as loops and disordered regions that engage in transient hydrogen bonds with the solvent.[104] Solvent accessibility is also a key factor, as amide protons buried deep within hydrophobic cores are not available for exchange.[105]

As alluded to earlier, proteins are not static structures, but can populate a variety of different conformations in solution through transient thermal fluctuations, known as conformational dynamics. These motions require the transient breaking of backbone hydrogen bonds, making the freed amides

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

available for exchange with solvent D_2O, thus HDX is effectively a measure of conformational dynamics.[54] Exchange kinetics can be expressed as follows:

$$\text{cl (H)} \underset{k_{cl}}{\overset{k_{op}}{\rightleftarrows}} \text{op (H)} \underset{D_2O}{\overset{k_{ch}}{\rightarrow}} \text{op (D)} \underset{k_{op}}{\overset{k_{cl}}{\rightleftarrows}} \text{cl (D)} \tag{7.2}$$

where cl and op refer to closed (unavailable for exchange) and open (available for exchange) states, with k_{cl} and k_{op} corresponding to their rate constants. H and D represent protonated and deuterated states. Proteins in solution are in equilibrium between closed and open states, with the rate of exchange affected by unfolding and re-folding events, as well as the intrinsic exchange rate, k_{ch}.[106,107] Therefore, the HDX rate constant can be expressed as follows:

$$k_{HDX} = \frac{k_{op} \times k_{ch}}{k_{op} + k_{cl} + k_{ch}}. \tag{7.3}$$

Typically, under native conditions, the frequency of protein folding is much greater than that of unfolding, where $k_{cl} \gg k_{op}$, reducing eqn (7.3) to:

$$k_{HDX} = \frac{k_{op} \times k_{ch}}{k_{cl} + k_{ch}}. \tag{7.4}$$

Depending on the relative timescale of the opening and closing events, two scenarios can arise. In the first scenario $k_{cl} \ll k_{ch}$, where protein refolding is much slower than the intrinsic exchange rate. In this case, a single unfolding event can allow for all amide hydrogens to exchange, making k_{HDX} extremely dependent on k_{op}, reducing eqn (7.4) to:

$$k_{HDX} = k_{op(k_{cl} \ll k_{ch},\ \text{EX1})}. \tag{7.5}$$

This mode is referred to as EX1 kinetics, where the observed rate of exchange is directly correlated to the rate constant of the opening reaction k_{op}. In the second scenario $k_{cl} \gg k_{ch}$, where protein refolding occurs rapidly following transient unfolding states. In this case, exchange occurs during several unfolding events, reducing eqn (7.4) to:

$$k_{HDX} = \frac{k_{op} \times k_{ch}}{k_{cl}} = K_{op} \times k_{ch\left(k_{cl} \gg k_{ch},\ \text{EX2}\right)} \tag{7.6}$$

where K_{op} is the equilibrium constant defined by $K_{op} = k_{op}/k_{cl}$. This mode is referred to as EX2 kinetics, and occurs for most proteins under physiological conditions.[108,109]

Coupling of TRESI-MS to HDX provides a measure of protein dynamics by monitoring the exchange reaction over time. This is primarily conducted using a continuous-flow device, where the protein of interest is in a native pH buffer (ammonium acetate) that is diluted with D_2O to initiate the labeling

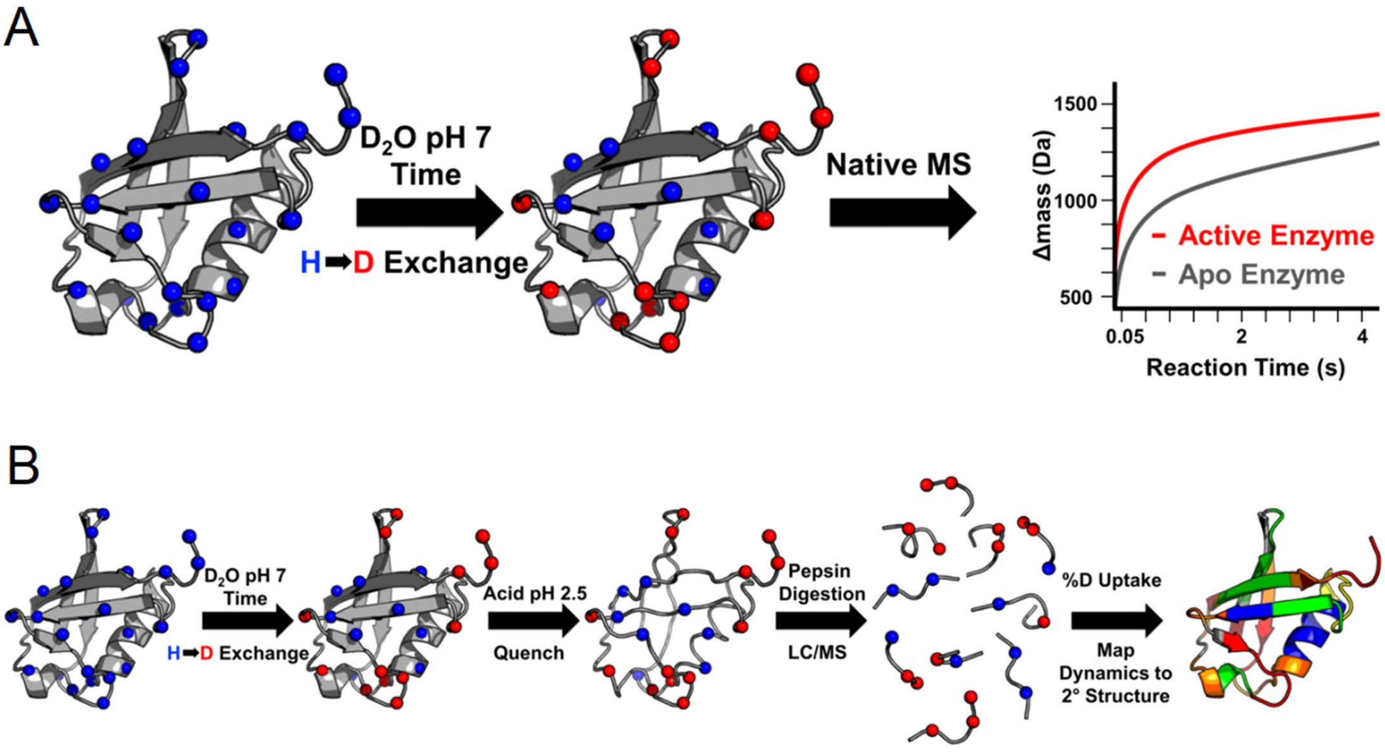

Figure 7.5 Coupling of HDX to MS. (A) Global HDX: mixing of the native protein with D_2O occurs at various time scales. As the reaction time increases, backbone amide hydrogens exchange causing the mass of the protein to increase, resulting in a mass shift to higher *m*/*z* when monitored by ESI-MS. (B) Local HDX: following incubation with D_2O, the mixture is quenched to pH ~ 2.5 to lock the deuterium in place. Pepsin is typically used, sometimes in combination with other acid proteases, for digestion into peptides. This is followed by liquid chromatography MS separation and detection of deuterated peptides. The dynamic regions of the protein are typically displayed on X-ray crystal structures.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

reaction (Figure 7.4B). Deuterium is often supplied in excess (greater than 50%) in order to favour the labelling process, which is then quenched by lowering the pH to ~2.5. The solution is then sprayed into the mass spectrometer where mass-shifts of the deuterated protein can be monitored. Monitoring the rate of uptake for an intact protein is referred to as global HDX, and can characterize global changes in the protein structure due to ligand binding or other events.[110,111] In the context of enzymes, this method has been used to study small-molecule inhibition[112] and catalysis-linked dynamics[113] (Figure 7.5A). Alternatively, for spatial resolution, the labeled, quenched protein can be subjected to pepsin digestion and MS analysis of the resulting peptides (Figure 7.5B). This is referred to as local HDX using a 'bottom up' workflow, which allows one to determine regions of substantial change, with recent developments nearing single amino acid resolution.[114]

7.3 Conclusion and Future Directions

An overview of the advantages and disadvantages of the methods presented in this chapter are described in Table 7.1. Briefly, MS avoids many of the challenges of other structural methods such as the use of bulky

Table 7.1 Comparison of analytical methods used for the study of enzymes.

Method	Advantages	Disadvantages
UV–visible fluorescence spectroscopy	Timescale: microseconds (μs); high sensitivity	Low selectivity; use of bulky chromophores (if available) that can alter kinetics
X-ray crystallography	Label-free analysis; can offer atomic-level (>1 Å) resolution (torsion angles, bond angles, *etc.*); no size limit for the protein of interest	Milligram (mg) quantities needed; extensive optimization required to produce highly quality crystals
2D NMR	Label-free (no chromophoric substrate required); amino acid resolution	μM–mM concentrations needed; cannot study proteins ~75 kDa or greater; lengthy acquisition times
Isothermal titration calorimetry	Label-free analysis; allows for thermodynamic characterization	Large sample volumes needed
MS	Label-free (no chromophoric substrate required); timescale: as low as milliseconds when coupled to TRESI; highly sensitive; nM–μM concentrations needed; no size limit for the protein of interest; can detect a large number of species from complex mixtures; nearing amino acid resolution when coupled to H–D exchange	Solution-phase structure not always retained in the gas phase if harsh conditions are used

chromophores, the need for high protein/substrate concentrations, and no limitation on protein size or the number of observable species. In addition, coupling to rapid mixing devices allows for the analysis of many reactions otherwise inaccessible by classical methods, including pre-steady state measurements, the detection of transient reaction intermediates, and the characterization of rapid conformational changes during catalysis.[113,115] Recent studies that have applied MS to several enzyme-based systems will be summarized in the following chapter. With continuous advancements in both instrumentation and data processing software, it is expected that the use of MS-based techniques will continue to rise for the study of challenging enzymatic systems, allowing for a more detailed understanding of vital biological processes.

References

1. L. T. Troland, Biological Enigmas and the Theory of Enzyme Action, *Am. Nat.*, 1917, **51**, 321–350.
2. R. Wolfenden and M. J. Snider, The depth of chemical time and the power of enzymes as catalysts, *Acc. Chem. Res.*, 2001, **34**, 938–945.
3. U. Schlattner, M. Tokarska-Schlattner and T. Wallimann, Mitochondrial creatine kinase in human health and disease, *Biochim. Biophys. Acta, Mol. Basis Dis.*, 2006, **1762**, 164–180.
4. X.-J. Huang, *et al.*, Aspartate Aminotransferase (AST/GOT) and Alanine Aminotransferase (ALT/GPT) Detection Techniques, *Sensors*, 2006, **6**, 756–782.
5. *Industrial Enzymology: The Application of Enzymes in Industry*, ed. T. Godfrey and J. Reichelt, 1982.
6. C. B. Anfinsen, Principles that Govern the Folding of Protein Chains, *Science*, 1973, **181**, 223–230.
7. L. H. Chen, *et al.*, 4-Oxalocrotonate tautomerase, an enzyme composed of 62 amino acid residues per monomer, *J. Biol. Chem.*, 1992, **267**, 17716–17721.
8. C. P. Whitman, The 4-oxalocrotonate tautomerase family of enzymes: how nature makes new enzymes using a β–α–β structural motif, *Arch. Biochem. Biophys.*, 2002, **402**, 1–13.
9. P. Chelikani, I. Fita and P. C. Loewen, Diversity of structures and properties among catalases, *Cell. Mol. Life Sci.*, 2004, **61**, 192–208.
10. R. Rej, Review: the role of coenzymes in clinical enzymology, *Ann. Clin. Lab. Sci.*, 1977, **7**, 455–468.
11. D. E. Koshland, Application of a Theory of Enzyme Specificity to Protein Synthesis, *Proc. Natl. Acad. Sci. U. S. A.*, 1958, **44**, 98–104.
12. H. N. Motlagh, J. O. Wrabl, J. Li and V. J. Hilser, The ensemble nature of allostery, *Nature*, 2014, **508**, 331–339.
13. Y. Savir and T. Tlusty, Conformational Proofreading: The Impact of Conformational Changes on the Specificity of Molecular Recognition, *PLoS One*, 2007, **2**, e468.

14. S. Kumar, B. Ma, C. J. Tsai, N. Sinha and R. Nussinov, Folding and binding cascades: dynamic landscapes and population shifts, *Protein Sci.*, 2000, **9**, 10–19.
15. J. M. Berg, *et al.*, *Biochemistry*, W H Freeman, 2002.
16. V. L. Schramm, Enzymatic transition states and transition state analogues, *Curr. Opin. Struct. Biol.*, 2005, **15**, 604–613.
17. M. A. Crook, *Clinical Biochemistry and Metabolic Medicine Eighth Edition*, CRC Press, 2013.
18. R. A. Bonomo and D. Szabo, Mechanisms of Multidrug Resistance in Acinetobacter Species and Pseudomonas aeruginosa, *Clin. Infect. Dis.*, 2006, **43**, S49–S56.
19. N. S. Datta, G. N. Wilson and A. K. Hajra, Deficiency of Enzymes Catalyzing the Biosynthesis of Glycerol-Ether Lipids in Zellweger Syndrome, *N. Engl. J. Med.*, 1984, **311**, 1080–1083.
20. J. G. Robertson, Mechanistic Basis of Enzyme-targeted Drugs, *Biochemistry*, 2005, **44**, 5561–5571.
21. R. L. Stein, *Kinetics of Enzyme Action: Essential Principles for Drug Hunters*, John Wiley & Sons, 2011.
22. N. Shkriabai, *et al.*, A Critical Role of the C-terminal Segment for Allosteric Inhibitor-induced Aberrant Multimerization of HIV-1 Integrase,*J. Biol. Chem.*, 2014, **289**, 26430–26440.
23. M. S. Smyth and J. H. J. Martin, x Ray crystallography, *Mol. Pathol.*, 2000, **53**, 8–14.
24. A. M. Hassell, *et al.*, Crystallization of protein–ligand complexes, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2007, **63**, 72–79.
25. J. Hajdu, *et al.*, Analyzing protein functions in four dimensions, *Nat. Struct. Mol. Biol.*, 2000, **7**, 1006–1012.
26. M. W. Makinen and A. L. Fink, Reactivity and Cryoenzymology of Enzymes in the Crystalline State, *Annu. Rev. Biophys. Bioeng.*, 1977, **6**, 301–343.
27. R. C. Wilmouth, *et al.*, X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate, *Nat. Struct. Mol. Biol.*, 2001, **8**, 689–694.
28. A. R. Pearson and R. L. Owen, Combining X-ray crystallography and single-crystal spectroscopy to probe enzyme mechanisms, *Biochem. Soc. Trans.*, 2009, **37**, 378–381.
29. M. T. Miller, B. O. Bachmann, C. A. Townsend and A. C. Rosenzweig, The catalytic cycle of β-lactam synthetase observed by x-ray crystallographic snapshots, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 14752–14757.
30. G. Katona, *et al.*, X-ray Structure of a Serine Protease Acyl-Enzyme Complex at 0.95-Å Resolution,*J. Biol. Chem.*, 2002, **277**, 21962–21970.
31. X. Ding, B. F. Rasmussen, G. A. Petsko and D. Ringe, Direct Crystallographic Observation of an Acyl-Enzyme Intermediate in the Elastase-Catalyzed Hydrolysis of a Peptidyl Ester Substrate: Exploiting the 'Glass Transition' in Protein Dynamics, *Bioorg. Chem.*, 2006, **34**, 410–423.
32. K. R. Acharya and M. D. Lloyd, The advantages and limitations of protein crystal structures, *Trends Pharmacol. Sci.*, 2005, **26**, 10–14.

33. D. A. Kraut, K. S. Carroll and D. Herschlag, Challenges in Enzyme Mechanism and Energetics, *Annu. Rev. Biochem.*, 2003, **72**, 517–571.
34. C. R. Wagner and S. J. Benkovic, Site directed mutagenesis: a tool for enzyme mechanism dissection, *Trends Biotechnol.*, 1990, **8**, 263–270.
35. A. K. Carmona, M. A. Juliano and L. Juliano, The use of Fluorescence Resonance Energy Transfer (FRET) peptides for measurement of clinically important proteolytic enzymes, *An. Acad. Bras. Cienc.*, 2009, **81**, 381–392.
36. H. Hartridge and F. J. W. A. Roughton, Method of Measuring the Velocity of Very Rapid Chemical Reactions, *Proc. R. Soc. London, Ser. A*, 1923, **104**, 376–394.
37. Q. H. Gibson and L. Milnes, Apparatus for rapid and sensitive spectrophotometry, *Biochem. J.*, 1964, **91**, 161–171.
38. A. Gomez-Hens and D. Perez-Bendito, The stopped-flow technique in analytical chemistry, *Anal. Chim. Acta*, 1991, **242**, 147–177.
39. P. Regenfuss, R. M. Clegg, M. J. Fulwyler, F. J. Barrantes and T. M. Jovin, Mixing liquids in microseconds, *Rev. Sci. Instrum.*, 1985, **56**, 283–290.
40. C.-K. Chan, *et al.*, Submillisecond protein folding kinetics studied by ultrarapid mixing, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 1779–1784.
41. M. C. R. Shastry, S. D. Luck and H. Roder, A Continuous-Flow Capillary Mixing Method to Monitor Reactions on the Microsecond Time Scale, *Biophys. J.*, 1998, **74**, 2714–2721.
42. D. B. Northrop and F. B. Simpson, New concepts in bioorganic chemistry beyond enzyme kinetics: Direct determination of mechanisms by stopped-flow mass spectrometry, *Bioorg. Med. Chem.*, 1997, **5**, 641–644.
43. G. Holdgate, in *Ligand-macromolecular Interactions in Drug Discovery*, ed. A. C. A. Roque, Humana Press, 2010, pp. 101–133.
44. M. W. Freyer and E. A. Lewis, Isothermal titration calorimetry: experimental design, data analysis, and probing macromolecule/ligand binding and kinetic interactions, *Methods Cell Biol.*, 2008, 79–113.
45. L. Mazzei, S. Ciurli and B. Zambelli, Hot biological catalysis: isothermal titration calorimetry to characterize enzymatic reactions, *J. Visualized Exp.*, 2014, **86**, e51487.
46. M. J. Todd and J. Gomez, Enzyme Kinetics Determined Using Calorimetry: A General Assay for Enzyme Activity? *Anal. Biochem.*, 2001, **296**, 179–187.
47. J. B. Chaires, Calorimetry and Thermodynamics in Drug Design, *Annu. Rev. Biophys.*, 2008, **37**, 135–151.
48. S. Leavitt and E. Freire, Direct measurement of protein binding energetics by isothermal titration calorimetry, *Curr. Opin. Struct. Biol.*, 2001, **11**, 560–566.
49. A. Bax, Two-Dimensional NMR and Protein Structure, *Annu. Rev. Biochem.*, 1989, **58**, 223–256.
50. X. Wang, H. Tachikawa, X. Yi, K. M. Manoj and L. P. Hager, Two-dimensional NMR Study of the Heme Active Site Structure of Chloroperoxidase, *J. Biol. Chem.*, 2003, **278**, 7765–7774.

51. G. P. Lisi and J. P. Loria, Solution NMR Spectroscopy for the Study of Enzyme Allostery, *Chem. Rev.*, 2016, **116**, 6323–6369.
52. S. Grutsch, S. Brüschweiler and M. Tollinger, NMR Methods to Study Dynamic Allostery, *PLoS Comput. Biol.*, 2016, **12**, e1004620.
53. S. W. Englander and L. Mayne, Protein Folding Studied Using Hydrogen-exchange Labeling and Two-dimensional NMR, *Annu. Rev. Biophys. Biomol. Struct.*, 1992, **21**, 243–265.
54. L. Konermann, J. Pan and Y.-H. Liu, Hydrogen exchange mass spectrometry for studying protein structure and dynamics, *Chem. Soc. Rev.*, 2011, **40**, 1224–1234.
55. G. M. Clore and A. M. Gronenborn, New methods of structure refinement for macromolecular structure determination by NMR, *Proc. Natl. Acad. Sci.*, 1998, **95**, 5891–5898.
56. I. W. Griffiths, J. J. Thomson — the Centenary of His Discovery of the Electron and of His Invention of Mass Spectrometry, *Rapid Commun. Mass Spectrom.*, 1997, **11**, 2–16.
57. J. B. Fenn, Electrospray Wings for Molecular Elephants (Nobel Lecture), *Angew. Chem., Int. Ed.*, 2003, **42**, 3871–3894.
58. P. Liuni and D. J. Wilson, Understanding and optimizing electrospray ionization techniques for proteomic analysis, *Expert Rev. Proteomics*, 2011, **8**, 197–209.
59. C. A. Hughey, R. P. Rodgers and A. G. Marshall, Resolution of 11 000 Compositionally Distinct Components in a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Crude Oil, *Anal. Chem.*, 2002, **74**, 4145–4149.
60. P. Kebarle and U. H. Verkerk, in *Electrospray and MALDI Mass Spectrometry*, ed. R. B. Cole, John Wiley & Sons, Inc., 2010, pp. 1–48.
61. S. Banerjee and S. Mazumdar, Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte, *Int. J. Anal. Chem.*, 2012, **2012**, e282574.
62. A. Gomez and K. Tang, Charge and fission of droplets in electrostatic sprays, *Phys. Fluids*, 1994, **6**, 404–414.
63. J. Liu and L. Konermann, Protein–Protein Binding Affinities in Solution Determined by Electrospray Mass Spectrometry, *J. Am. Soc. Mass Spectrom.*, 2011, **22**, 408–417.
64. W. Wang, E. N. Kitova and J. S. Klassen, Influence of solution and gas phase processes on protein-carbohydrate binding affinities determined by nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry, *Anal. Chem.*, 2003, **75**, 4945–4955.
65. T. J. D. Jørgensen, P. Roepstorff and A. J. R. Heck, Direct Determination of Solution Binding Constants for Noncovalent Complexes between Bacterial Cell Wall Peptide Analogues and Vancomycin Group Antibiotics by Electrospray Ionization Mass Spectrometry, *Anal. Chem.*, 1998, **70**, 4427–4432.
66. J. B. Fenn, Ion formation from charged droplets: roles of geometry, energy, and time, *J. Am. Soc. Mass Spectrom.*, 1993, **4**, 524–535.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

67. M. Karas, U. Bahr and T. Dülcks, Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine, *Fresenius' J. Anal. Chem.*, 2000, **366**, 669–676.
68. R. Juraschek, T. Dülcks and M. Karas, Nanoelectrospray—more than just a minimized-flow electrospray ionization source, *J. Am. Soc. Mass Spectrom.*, 1999, **10**, 300–308.
69. J. L. P. Benesch, B. T. Ruotolo, D. A. Simmons and C. V. Robinson, Protein complexes in the gas phase: technology for structural genomics and proteomics, *Chem. Rev.*, 2007, **107**, 3544–3567.
70. M. Wilm and M. Mann, Analytical properties of the nanoelectrospray ion source, *Anal. Chem.*, 1996, **68**, 1–8.
71. Z. Takáts, J. M. Wiseman, B. Gologan and R. G. Cooks, Electrosonic spray ionization. A gentle technique for generating folded proteins and protein complexes in the gas phase and for studying ion-molecule reactions at atmospheric pressure, *Anal. Chem.*, 2004, **76**, 4050–4058.
72. J. M. Wiseman, Z. Takáts, B. Gologan, V. J. Davisson and R. G. Cooks, Direct characterization of enzyme-substrate complexes by using electrosonic spray ionization mass spectrometry, *Angew. Chem., Int. Ed. Engl.*, 2005, **44**, 913–916.
73. M. C. Jecklin, D. Touboul, C. Bovet, A. Wortmann and R. Zenobi, Which Electrospray-Based Ionization Method Best Reflects Protein-Ligand Interactions Found in Solution? A Comparison of ESI, nanoESI, and ESSI for the Determination of Dissociation Constants with Mass Spectrometry, *J. Am. Soc. Mass Spectrom.*, 2008, **19**, 332–343.
74. S. A. Hofstadler and K. A. Sannes-Lowery, Applications of ESI-MS in drug discovery: interrogation of noncovalent complexes, *Nat. Rev. Drug Discovery*, 2006, **5**, 585–595.
75. B. Bothner, *et al.*, Monitoring Enzyme Catalysis with Mass Spectrometry, *J. Biol. Chem.*, 2000, **275**, 13455–13459.
76. L. Konermann, J. Pan and D. J. Wilson, Protein folding mechanisms studied by time-resolved electrospray mass spectrometry, *BioTechniques*, 2006, **40**, 135, 137, 139 passim.
77. T. Rob and D. J. Wilson, Time-resolved mass spectrometry for monitoring millisecond time-scale solution-phase processes, *Eur. J. Mass Spectrom. (Chichester)*, 2012, **18**, 205–214.
78. B. M. Kolakowski, D. A. Simmons and L. Konermann, Stopped-flow electrospray ionization mass spectrometry: a new method for studying chemical reaction kinetics in solution, *Rapid Commun. Mass Spectrom.*, 2000, **14**, 772–776.
79. L. Konermann, B. A. Collings and D. J. Douglas, Cytochrome c Folding Kinetics Studied by Time-Resolved Electrospray Ionization Mass Spectrometry, *Biochemistry*, 1997, **36**, 5554–5559.
80. D. J. Wilson and L. Konermann, A Capillary Mixer with Adjustable Reaction Chamber Volume for Millisecond Time-Resolved Studies by Electrospray Mass Spectrometry, *Anal. Chem.*, 2003, **75**, 6408–6414.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00173

81. C. Lento, G. F. Audette and D. J. Wilson, Time-resolved electrospray mass spectrometry — a brief history, *Can. J. Chem.*, 2014, **93**, 7–12.
82. D. J. Wilson and L. Konermann, Mechanistic studies on enzymatic reactions by electrospray ionization MS using a capillary mixer with adjustable reaction chamber volume for time-resolved measurements, *Anal. Chem.*, 2004, **76**, 2537–2543.
83. J. F. Kirsch, Enzyme kinetics and mechanism, by Paul F. Cook and W. W. Cleland, *Protein Sci.*, 2008, **17**, 380–381.
84. D. B. Northrop, Steady-state analysis of kinetic isotope effects in enzymic reactions, *Biochemistry*, 1975, **14**, 2644–2651.
85. W. W. Cleland, The use of isotope effects to determine enzyme mechanisms, *Arch. Biochem. Biophys.*, 2005, **433**, 2–12.
86. J. C. Nesheim and J. D. Lipscomb, Large Kinetic Isotope Effects in Methane Oxidation Catalyzed by Methane Monooxygenase: Evidence for C–H Bond Cleavage in a Reaction Cycle Intermediate, *Biochemistry*, 1996, **35**, 10240–10247.
87. K.-H. Kim, E. M. Isin, C.-H. Yun, D.-H. Kim and F. Guengerich, Kinetic deuterium isotope effects for 7-alkoxycoumarin O-dealkylation reactions catalyzed by human cytochromes P450 and in liver microsomes, *FEBS J.*, 2006, **273**, 2223–2231.
88. P. J. Berti, Determining transition states from kinetic isotope effects, *Methods Enzymol.*, 1999, **308**, 355–397.
89. X. Du and S. R. Sprang, Transition State Structures and the Roles of Catalytic Residues in GAP-Facilitated GTPase of Ras As Elucidated by 18O Kinetic Isotope Effects, *Biochemistry*, 2009, **48,** 4538–4547.
90. J. Rodgers, D. A. Femec and R. L. Schowen, Isotopic mapping of transition-state structural features associated with enzymic catalysis of methyl transfer, *J. Am. Chem. Soc.*, 1982, **104**, 3263–3268.
91. J. Monod, J. Wyman and J.-P. Changeux, On the nature of allosteric transitions: A plausible model, *J. Mol. Biol.*, 1965, **12**, 88–118.
92. K. Gunasekaran, B. Ma and R. Nussinov, Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 2004, **57**, 433–443.
93. J. Wyman and S. J. Gill, *Binding and Linkage: Functional Chemistry of Biological Macromolecules*, University Science Books, 1990.
94. R. A. Laskowski, F. Gerick and J. M. Thornton, The structural basis of allosteric regulation in proteins, *FEBS Lett.*, 2009, **583**, 1692–1698.
95. J. M. Daniel, S. D. Friess, S. Rajagopalan, S. Wendt and R. Zenobi, Quantitative determination of noncovalent binding interactions using soft ionization mass spectrometry, *Int. J. Mass Spectrom.*, 2002, **216**, 1–27.
96. A. Tjernberg, *et al.*, Determination of dissociation constants for protein-ligand complexes by electrospray ionization mass spectrometry, *Anal. Chem.*, 2004, **76**, 4325–4331.
97. J. A. Loo, Studying noncovalent protein complexes by electrospray ionization mass spectrometry, *Mass Spectrom. Rev.*, 1997, **16**, 1–23.

98. A. Ayed, A. N. Krutchinsky, W. Ens, K. G. Standing and H. W. Duckworth, Quantitative evaluation of protein-protein and ligand-protein equilibria of a large allosteric enzyme by electrospray ionization time-of-flight mass spectrometry, *Rapid Commun. Mass Spectrom.*, 1998, **12**, 339–344.
99. M. C. Jecklin, S. Schauer, C. E. Dumelin and R. Zenobi, Label-free determination of protein-ligand binding constants using mass spectrometry and validation using surface plasmon resonance and isothermal titration calorimetry, *J. Mol. Recognit.*, 2009, **22**, 319–329.
100. Z. Zhang and D. L. Smith, Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation, *Protein Sci.*, 1993, **2**, 522–531.
101. Y. Bai, J. S. Milne, L. Mayne and S. W. Englander, Primary structure effects on peptide group hydrogen exchange, *Proteins*, 1993, **17**, 75–86.
102. D. L. Smith, Y. Deng and Z. Zhang, Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry, *J. Mass Spectrom.*, 1997, **32**, 135–146.
103. C. L. Perrin, Proton exchange in amides: Surprises from simple systems, *Acc. Chem. Res.*, 1989, **22**, 268–275.
104. V. Katta, B. T. Chait and S. Carr, Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry, *Rapid Commun. Mass Spectrom.*, 1991, **5**, 214–217.
105. R. Li and C. Woodward, The hydrogen exchange core and protein folding, *Protein Sci.*, 1999, **8**, 1571–1590.
106. A. Hvidt and S. O. Nielsen, Hydrogen exchange in proteins, *Adv. Protein Chem.*, 1966, **21**, 287–386.
107. L. Konermann, X. Tong and Y. Pan, Protein structure and dynamics studied by mass spectrometry: H/D exchange, hydroxyl radical labeling, and related approaches, *J. Mass Spectrom.*, 2008, **43**, 1021–1036.
108. D. M. Ferraro, N. D. Lazo and A. D. Robertson, EX1 hydrogen exchange and protein folding, *Biochemistry*, 2004, **43**, 587–594.
109. D. D. Weis, T. E. Wales, J. R. Engen, M. Hotchko and L. F. Eyck, Identification and characterization of EX1 kinetics in H/D exchange mass spectrometry by peak width analysis, *J. Am. Soc. Mass Spectrom.*, 2006, **17**, 1498–1509.
110. D. Houde, J. Arndt, W. Domeier, S. Berkowitz and J. R. Engen, Rapid characterization of IgG1 conformation and conformational dynamics by hydrogen/deuterium exchange mass spectrometry, *Anal. Chem.*, 2009, **81**, 2644–2651.
111. M. M. Zhu, D. L. Rempel, Z. Du and M. L. Gross, Quantification of Protein–Ligand Interactions by Mass Spectrometry, Titration, and H/D Exchange: PLIMSTEX, *J. Am. Chem. Soc.*, 2003, **125**, 5252–5253.
112. S. R. Marcsisin and J. R. Engen, Hydrogen exchange mass spectrometry: what is it and what can it tell us? *Anal. Bioanal. Chem.*, 2010, **397**, 967–972.

113. P. Liuni, A. Jeganathan and D. J. Wilson, Conformer Selection and Intensified Dynamics During Catalytic Turnover in Chymotrypsin, *Angew. Chem., Int. Ed.*, 2012, **51**, 9666–9669.
114. Z.-Y. Kan, B. T. Walters, L. Mayne and S. W. Englander, Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 16438–16443.
115. D. J. Clarke, A. A. Stokes, P. Langridge-Smith and C. L. Mackay, Online quench-flow electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for elucidating kinetic and chemical enzymatic reaction mechanisms, *Anal. Chem.*, 2010, **82**, 1897–1904.

CHAPTER 8

Studying Enzyme Mechanisms Using Mass Spectrometry, Part 2: Applications

PETER LIUNI[a], CRISTINA LENTO[b] AND DEREK J. WILSON*[a,b,c]

[a]Department of Chemistry, York University, Toronto, M3J 1P3, Canada; [b]Centre for Research of Biomolecular Interactions, York University, Toronto, M3J 1P3, Canada; [c]Centre for Research in Mass Spectrometry, York University, Toronto, M3J 1P3, Canada
*E-mail: dkwilson@yorku.ca

8.1 Introduction

Chapter 7 introduced the important role of enzymes in biological systems and provided a summary of analytical methods classically used for their study. Several key advantages in using mass spectrometry (MS) for observing structural changes at both the protein and substrate level during enzymatic turnover were also outlined. We will now discuss how MS has been successfully applied to study complex enzyme mechanisms (both steady-state and pre-steady-state), binding constants, allosteric regulation, and catalysis-linked dynamics.

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

8.2 Enzyme Mechanisms

8.2.1 Complex, Multistep Enzymatic Mechanisms

With the ability to simultaneously monitor many species, MS offers the unparalleled capability of unraveling complex enzymatic mechanisms. A prime example of this is Dovala and coworkers' use of X-ray crystallography and MS to characterize the structure and function of LpxM from *Acinetobacter baumannii*, a lysophospholipid acyltransferase that is responsible for lipid A synthesis.[1] Lipid A is an immune system activator and barrier to xenobiotics for Gram-negative bacteria, making its biosynthetic pathway a lucrative target for antibiotics.[2] After initiating the reaction on an Agilent RapidFire 300 high-throughput solid-phase extraction system, the conversion of lauryl-ACP to holo-ACP by wild-type LpxM and two mutants ($\text{LpxM}_{\text{E127A}}$ and $\text{LpxM}_{\text{R159A}}$) was monitored by MS/MS of intact ACP based on two lipid prosthetic group product ions at 443.294 *m*/*z* (acyl-ACP) and 261.127 *m*/*z* (holo-ACP) (Figure 8.1). Transfer of the lipid chain from lauryl-ACP to lipid IV_A was also monitored by liquid chromatography (LC)-MS (Figure 8.1). Using this unique approach, it was found that LpxM produced holo-ACP and free lauric acid in the absence of lipid IV_A, providing evidence for novel acyl-protein thioesterase activity. Substrate inhibition in the presence and absence of lipid IV_A was also observed at lauryl-ACP concentrations of 7 µM or higher, pointing to two acyl-ACP binding sites per LpxM. The authors proposed an ordered binding mechanism for Lipid IV_A and two lauryl-ACPs to LpxM, as well as a reset mechanism for high concentrations of lauryl-ACP (Figure 8.1). This robust MS-assay, which is based on a novel characteristic loss of lipid moieties in ACP, can be extended to a diverse set of lysophospholipid acyltransferases, as well as enabling high-throughput inhibitor screening of potential LpxM inhibitors (Table 8.1).

8.2.2 Time-resolved Electrospray Ionization (TRESI) for the Detection of Enzymatic Intermediates

Another example of mechanistic insights acquired by direct MS measurement comes from the Anderson group, who were among the first researchers to combine rapid mixing with ESI-MS analysis. Anderson and co-workers applied a time-resolved ESI-MS approach to studying the catalytic complexes of the tetrameric 4-hydroxybenzoyl–coenzyme A (4-HBA-CoA) thioesterase from *Arthrobacter* sp. strain SU,[3] which catalyzes the hydrolysis of 4-HBA–CoA to 4-HBA and CoA.[4] By mixing 4-HBA-CoA substrate (5–30 µM) with thioesterase for 7 ms and monitoring by MS, all non-covalent ligand-bound species were observed (Table 8.2). Time-resolved catalysis experiments were run for reaction times of 6–160 ms after mixing 40 µM thioesterase and 30 µM 4-HBA–CoA (10 mM ammonium acetate, pH 7.5)

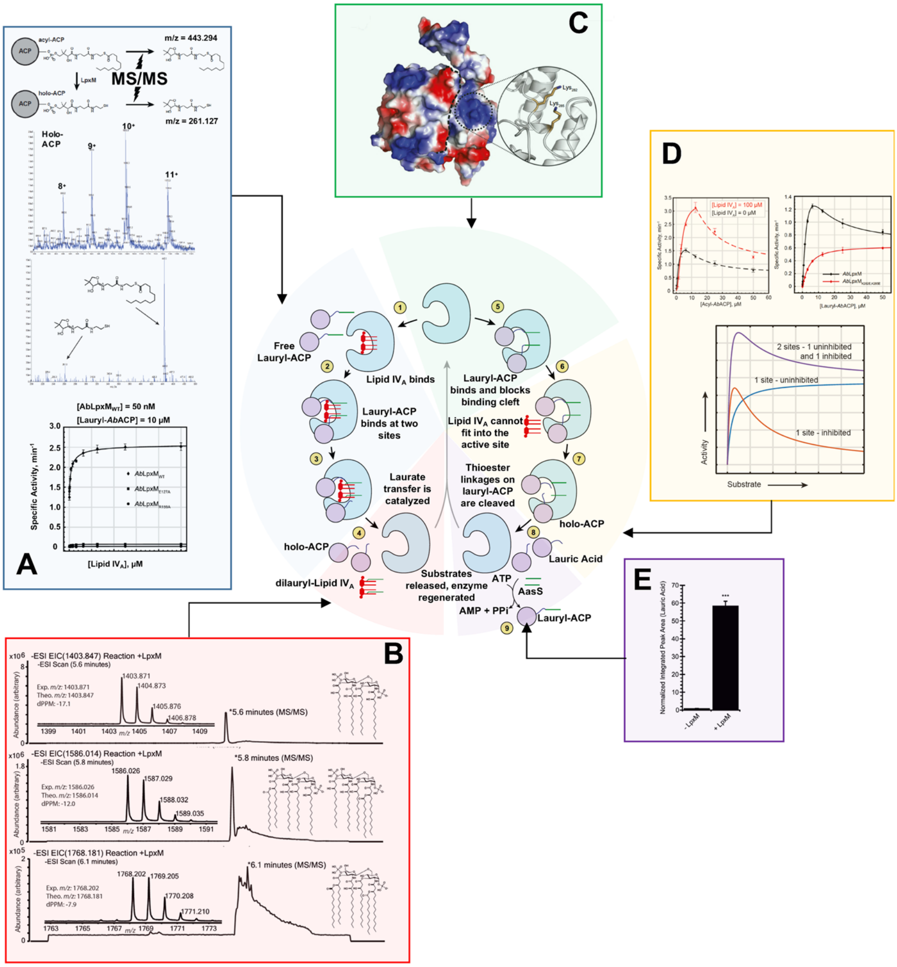

Figure 8.1 Proposed mechanism for AbLpxM. (A) LpxM activity assay from the formation of holo-ACP with increasing lipid IV$_A$ concentration. MS/MS of intact lauryl-ACP yields two phosphopantetheine product ions. (B) LC chromatograms and MS spectra for the addition of lauryl groups to lipid IVA. (C) Electrostatic surface of LpxM crystal structure identifying a second acyl-ACP binding site. (D) Left: substrate inhibition of AbLpxM in the presence (red) and absence (black) of lipid IVA. Right: substrate inhibition of AbLpxM with increasing lauryl-ACP concentration. Bottom: modeled kinetics for uninhibited LpxM (blue), inhibited LpxM (orange), and a combined inhibited/uninhibited LpxM. (E) Production of lauric acid in the presence and absence of LpxM determined by LC-MS.

Table 8.1 Analysis of acyl-ACP acyl chain length specificity on AbLpxM thioesterase activity as determined by high-throughput MS.[a]

Acyl chain donor	K_M^{App} (μM)	Specific activity (min^{-1})	Hill coefficient
Capryl-ACP	9.887 ± 0.481	0.238 ± 0.004	1.134 ± 0.041
Lauryl-ACP[b]	1.845 ± 0.238	1.483 ± 0.099	1.392 ± 0.134
Myristyl-ACP	10.35 ± 0.943	0.346 ± 0.013	1.328 ± 0.115
Palmityl-ACP	30.47 ± 10.3	0.057 ± 0.009	1.128 ± 0.196

[a] Values were calculated from data shown in Figure 8.4A using Matlab; error shown as 95% confidence interval. All experiments were repeated in triplicate.

[b] Parameters for lauryl-ACP were generated using data in the non-inhibited regime only.

Table 8.2 Experimental and theoretical masses of complexes of thioesterase with its substrate and/or products during catalysis.[a]

Complex of enzyme with its substrate and/or products	Experimental mass (Da)[b]	Theoretical mass (Da)
4(E + S)	69 123 ± 6	69 127.70
(3(E + S)+(E + CoA))	68 001 ± 4	69 008.01
(2(E + S)+2(E + CoA))	68 883 ± 5	68 887.74
((E + S)+3(E + CoA))	68 762 ± 6	68 767.65
(4(E + CoA))	68 641 ± 4	68 647.76
(3(E + S)+E)	68 238 ± 3	68 240.50
(2(E + S)+(E + CoA)+E)	68 116 ± 5	68 120.41
((E + S)+2(E + CoA)+E)	67 996 ± 6	68 000.32
(3(E + CoA)+E)	67 874 ± 6	67 880.23
(2(E + S)+2E)	67 347 ± 3	67 352.88
((E + S)+(E + CoA)+2E)	67 231 ± 4	67 232.79
(2(E + CoA)+2E)	67 106 ± 4	67 112.70
((E + CoA)+3E)	66 461 ± 4	66 465.26
((E + S)+3E)	66 338 ± 5	66 345.17
4E	65 572 ± 6	65 577.64

[a] Note: the theoretical mass of one sub-unit E, substrate S, CoA, and 4-HBA are 16 394.38, 887.62, 767.57 and 138.12 Da, respectively.

[b] Represents the average date of three experiments.

(Figure 8.2). The [(E + S)] complexes, as well as several [(E + P)] complexes, and several [(E + S) + (E + P)] complexes were identified based on their unique *m*/*z* values. The time course results show that at 80 ms CoA stays bound to the enzyme and dissociates slowly, with a substantial amount remaining bound to the enzyme at 160 ms. Catalysis tends to also occur more rapidly with higher occupancy, with the [4(E + S)] and [(4E+3S)] species falling to 0 intensity 80 ms into the reaction, suggesting some form of positive cooperativity. The authors conclude that the reaction pathway for the 4-HBA-CoA thioesterase follows release of product governed by the opening and closing of an N-terminal loop region in the hydroxybezoyl binding pocket. Based on the kinetic data, product release remains ordered where 4-HBA is released first, followed by CoA (Figure 8.2).

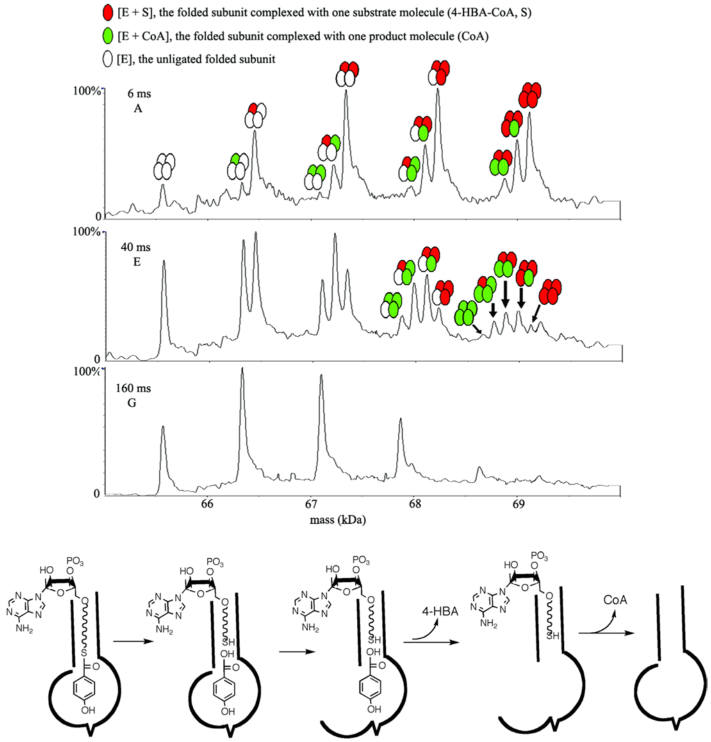

Figure 8.2 The proposed kinetic mechanism of thioesterase catalysis and order of product release.

8.2.3 Combination With Isotopic Labeling

MS is also particularly well suited to be combined with classical heavy isotope labeling techniques for mechanism elucidation, particularly when ultra-high resolution instruments are available. In a work by Bandarian and coworkers, native high resolution MS was used to characterize the mechanism of toyocamycin nitrile hydratase (TNH) under single and multiple turnover conditions.[5] The alpha sub-unit of TNH (ToyJ) was selected as it exhibits reduced catalytic activity compared to the full complex for the substrate toyocamycin (TNH: k_{cat} = 159 ± 2 s^{-1}, $K_M = 2.8 \times 10^{-2} \pm 1 \times 10^{-3}$ mM, *versus* ToyJ: k_{cat} = 0.44 ± 0.04 s^{-1}, K_M = 15 ± 2 mM). Reactions were carried out for 2 h in ^{18}O water for a range of toyocamycin concentrations (1 μM–10 mM) and fixed ToyJ (30 μM). Conversion from toyocamycin (S) to sangivamycin (P) was analysed by high-performance LC (HPLC) coupled to a Thermo Fisher LTQ Orbitrap XL, and ToyJ from the reaction was analysed by direct infusion into an Exactive Plus EMR MS. In multiple turnover conditions, a high excess of substrate (10 mM toyocamycin) to active sites displayed dominant formation of an [^{18}O]-sangivamycin peak at 312.1189 *m*/*z*. In single turnover conditions, a high excess of active sites relative to substrate (1 μM toyocamycin) produced an [^{16}O]-sangivamycin product peak at 310.1146 *m*/*z*, with a small 3.5% abundance of the ^{18}O-labeled product.

Deconvoluted spectra for the 8+ charge state of ToyJ in the absence and presence of excess substrate shows an increase in mass from 23 517.30 Da to 23 519.29 Da on multiple turnovers. Experimental isotopic distributions for a series of 8+ ToyJ mixed with 1 µM–10 mM toyocamycin were fitted to theoretical distributions to determine the percentage of total enzyme containing ^{18}O. The percentage of ^{18}O incorporation in ToyJ mirrored ^{18}O incorporation into sangivamycin. From this, the authors were able to conclude that a protein-based nucleophilic oxygen, and not a solvent-based nucleophilic oxygen, was responsible for the hydration of nitriles to product amides in nitrile hydratases (Figure 8.3).

8.2.4 Pre-steady-state Kinetic Isotope Effects (KIEs) Using TRESI-MS

KIEs can be used to infer transition state geometries, and have historically played a large role in elucidating enzyme mechanisms.[6,7] Liuni *et al.* reported a method for measuring KIEs in both the steady-state and pre-steady-state using a millisecond timescale time-resolved ESI-MS method.[8]

Kinetic measurements for 2 µM yeast alcohol dehydrogenase were conducted using a series of equimolar ethanol and 1,1-D_2 ethanol solutions (10–200 mM), each containing 400 µM NAD^+ prepared in 10 mM ammonium acetate (pH 8.4). An internal competition assay followed the release of cofactor NADH at 666 *m*/*z* and NADD at 667 *m*/*z*. An 'observed' kinetic isotope effect KIE_{obs} was reported as V_0(NADH)/V_0(NADD) = 1.8 ± 0.4. A classic Michaelis–Menton style experiment for ethanol concentrations (10–200 mM) yielded a KIE_{spec} of 2.19 ± 0.05 (R^2 NADH = 0.82, R^2 NADD = 0.88, Figure 8.4), which was in line with values reported by Mahler and Douglas (2.00)[9] and by Park *et al.* (2.9 ± 0.6).[10] The authors remark that the differences between KIE_{obs} and KIE_{spec} suggests a substantial shift in one or more microscopic equilibria upon isotopic substitution, but due to the complexity of the ADH mechanism, they were unable to pinpoint the exact chemical step(s) involved.

For chymotrypsin, a 25 µM chymotrypsin solution (pH 8.4) was mixed with 40% methanol containing both 2.5 mM ^{12}C–p-nitrophenyl acetate (NPA) and 2.5 mM ^{13}C–p-NPA (pH 7.0). A 10% methanol 6 mM HCl makeup solvent was added to quench the reaction and enhance the electrospray. Spectra acquired for a total of 30 min (n = 5) showed free enzymes at 25 446 ± 2 Da, and both acyl-enzyme at 25 488 ± 3 Da for the unlabeled substrate and 25 489 ± 3 Da for the labeled substrate (Figure 8.5). Extracted ion currents on each side of the 10+ acyl-enzyme peak give average KIE (n = 5) for acylation of 1.09 ± 0.02. Previous work by Hess and co-workers[11] report a steady state ^{13}(V/K) KIE of 1.030 ± 0.002, however, the measurement represents a convolution of primary KIEs for acylation and deacylation. The authors were ultimately able to conclude that the reaction proceeds through a late transition state for loss of paranitrophenolate from the tetrahedral acyl-enzyme complex, consistent with strong stabilization of the oxyanion.

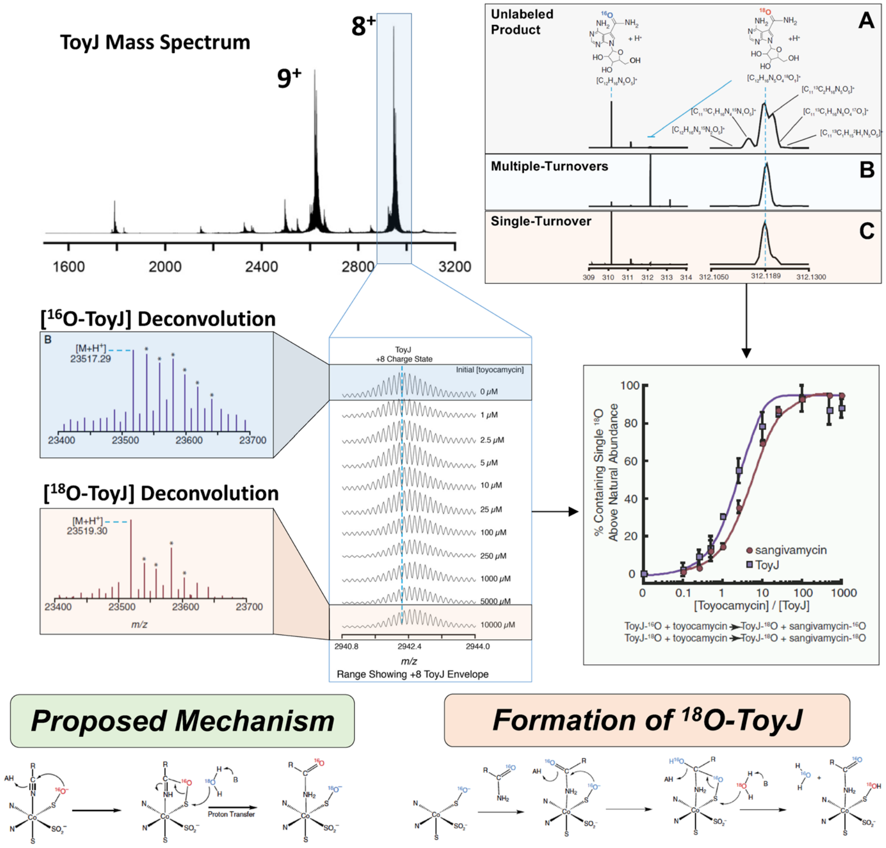

Figure 8.3 Mass spectrum of the 8+ and 9+ charge states of ToyJ acquired on the Exactive Plus EMR. The isotopic envelope of the 8+ charge state increases in *m*/*z* with increasing addition of toyocamycin substrate. Deconvolution of the native ToyJ mass spectrum at 0 μM substrate gives a nominal mass of 23 517.30 Da. At 10 mM, the nominal mass increases with the addition of an ^{18}O to 23 519.29 Da. High resolution LC-MS of the sangivamycin product acquired on the LTQ Orbitrap XL shows changes in the isotopic distribution from (A) unlabelled product, (B) product after multiple turnovers from 10 mM toyocamycin in ^{18}O labeled water, and (C) product under single turnover conditions with 1 μM toyocamycin in ^{18}O labeled water. A plot of the percentage containing single ^{18}O from fitting theoretical and experimental isotopic envelopes of ToyJ, and from intensities of the ^{18}O peak for sangivamycin, is given. Fits to the data are determined from the schemes and produced in KinTek Explorer software. The proposed mechanism of amide hydration (green) and the formation of ^{18}O-ToyJ from sangivamycin under multiple turnover conditions are also given.

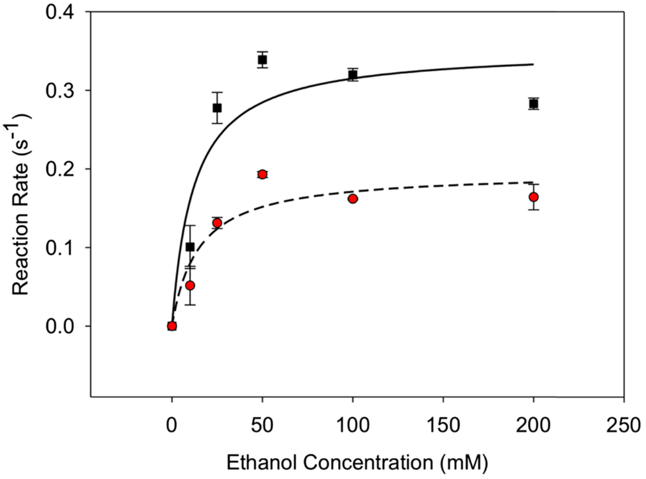

Figure 8.4 Steady-state analysis of YADH catalyzed oxidation of ethanol. Initial velocities of NADH production (black squares), and NADD production (red circles) are shown. After fitting to the Michaelis–Menten equation, the data yield a KIE_{spec} of 2.19 ± 0.05. Error bars are from 5 replicates.

8.3 Steady-state Kinetics

8.3.1 Steady-state Kinetics for Drug Development Assays

The development of new drugs to combat resistant strains of tuberculosis has focused on the shikimate pathway, which is not present in mammals and is unique to fungi, higher plants, and bacteria.[12] Simithy and coworkers developed an LC-ESI-MS method for monitoring the conversion of ATP and shikimate to ADP and shikimate-3-phosphate in shikimate kinase from *Micobacterium tuberculosis* (*Mt*Sk).[13] In this method, reactions are initiated with a 0.2 μM addition of *Mt*Sk to saturating conditions of ATP (1.2 mM) and shikimate (5 mM), incubated for 30 s and then quenched by the addition of 98% formic acid. For determining K_M and k_{cat}, initial velocities were measured across a range of concentrations (ATP: 56–1120 μM, shikimate: 325–4875 μM) corresponding to 0.5–10 times K_M (Figure 8.6). The apparent K_M for ATP and shikimate were 0.20 mM and 0.53 mM, respectively, with good agreement with previously reported values.[14] The k_{cat} values for ATP (68 s^{-1}) and shikimate (65 s^{-1}) were also in good agreement with values obtained from spectrophotometric methods.[14,15] To validate their method, the authors also conducted classical UV-coupled spectrophotometric assays, yielding similar values to those found by LC-MS. Inhibition constants were also generated for a well–studied inhibitor, which was also used to validate the current LC-MS assay (Table 8.3). Overall, the authors demonstrated a robust and methodical technique for steady-state kinetic characterization of enzymatic reactions by LC-MS/MS, showcasing MS as a reliable alternative to spectrophotometric assays when no chromophore is present.

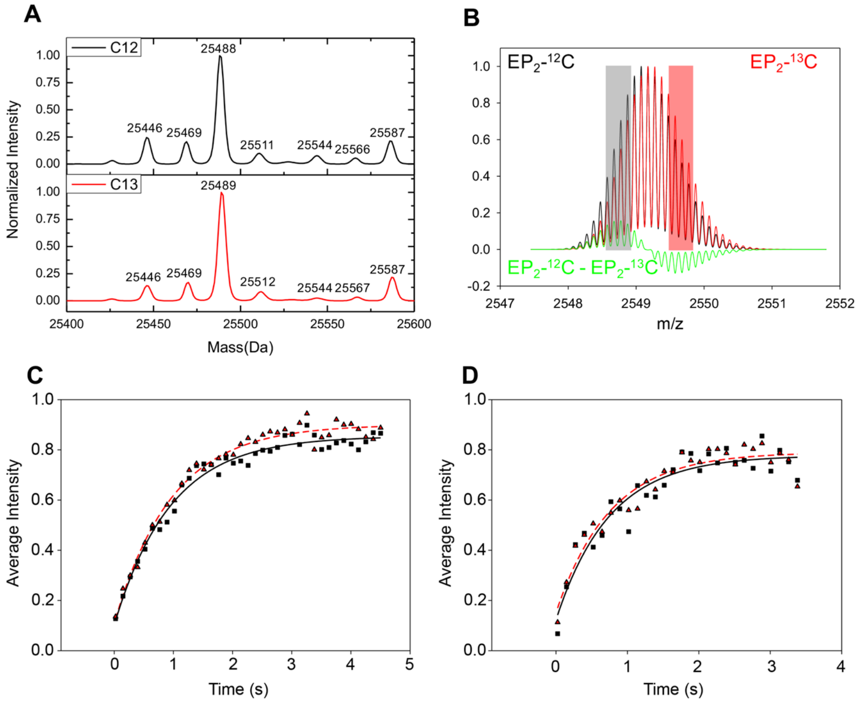

Figure 8.5 (A) Matched pair of deconvoluted mass distributions for unlabeled (^{12}C) and labeled (^{13}C) acyl-chymotrypsin. The peak at 25 446 Da corresponds to free δ′-chymotrypsin, 25 488 Da corresponds to the unlabeled acyl-enzyme, and 25 489 Da corresponds to the labeled acyl-enzyme. (B) A model of the 10$^+$ *m*/*z* peak for a 50/50 mixture of unlabeled (^{12}C, black) and labeled (^{13}C, grey) acyl-chymotrypsin. The smaller peaks appearing above and below the *m*/*z* axis are difference peaks (^{12}C–^{13}C). Coloured bars correspond to the ion current extraction window for ^{12}C acyl-enzyme (left) and ^{13}C acyl-enzyme (right), resulting in a 60% contribution from the desired species. (C) Intensity-time profile drawn from acyl-chymotrypsin and (D) chymotrypsin ion currents. Filled squares represent intensities taken from the right-hand side of the peak, and red triangles represent intensities taken from the left-hand side. (C) Typical ^{12}C/^{13}C profile for KIE measurements. (D) A ^{12}C/^{12}C negative control to ensure that KIE measurements are not an artifact of time-dependent changes in the peak shape.

8.3.2 Quantitative Assays on Challenging Analytes

Lysine acetyltransferases catalyze histone acetylation through transfer of an acetyl-group to a lysine residue from acetyl-CoA. Andrews and coworkers developed a label-free quantitative MS method for site-specific identification and steady-state analysis of acetylation events on histones H3 and H4 by Gcn5.[17] Steady-state kinetic assays were initiated by titrating histone concentrations (0.15–10.3 μM) against fixed concentrations of acetyl-CoA (200 μM) and Gcn5 (180 nM). Sample times and Gcn5 concentration were

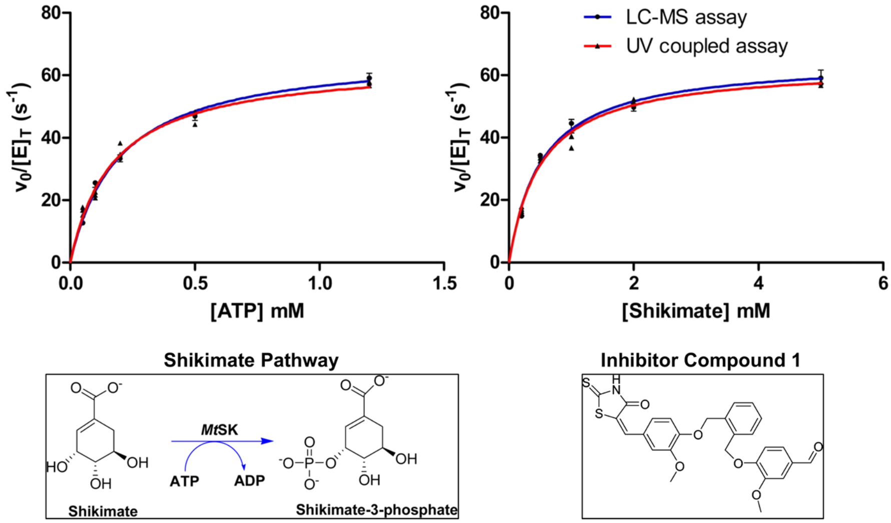

Figure 8.6 Michaelis–Menten kinetic plots for *Mt*Sk with respect to ATP and shikimate concentration. The blue lines are fits to initial velocity data obtained from LC-MS with a 0.2 μM fixed *Mt*SK concentration, and red lines are UV-coupled assays for 20 nM *Mt*SK. Concentrations for substrates were 5 mM shikimate, and 1.2 mM ATP. Table 8.3 summarizes inhibitory constants obtained from similar experiments conducted and associated literature values,[16] with a well-studied inhibitor (bottom right).

Table 8.3 Inhibition constants for compound 1 against *Mt*SK established from LC-MS and a UV-coupled assay.

	LC-MS assay	UV-assay	Literature UV-assay
K_i (μM)	10.90	9.87	9.84
αK_{iATP} (μM)	6.94	7.82	7.18
$\alpha K_{\mathrm{iShikimate}}$ (μM)	10.09	8.07	10.67

adjusted in order to obtain initial acetylation rates for individual modified lysines. Reactions were quenched, and free lysines on histones were modified with propionic anhydride to prevent over-digestion by trypsin. 47 tryptic peptides were separated by ultra-HPLC (UHPLC, Waters Acquity H-class) and analysed by selected-reaction monitoring on a Thermo TSQ Quantum Access triple quadrupole MS. Peptide fractions were quantitated by eqn (1) from integrated peak areas of all possible acetylated and propionylated states for an individual peptide, and validated for any ionization efficiency biases (Figure 8.7).

$$F_s = \frac{I_{\mathrm{peptide}}}{I_{\mathrm{all\ peptide\ states}}}. \qquad (8.1)$$

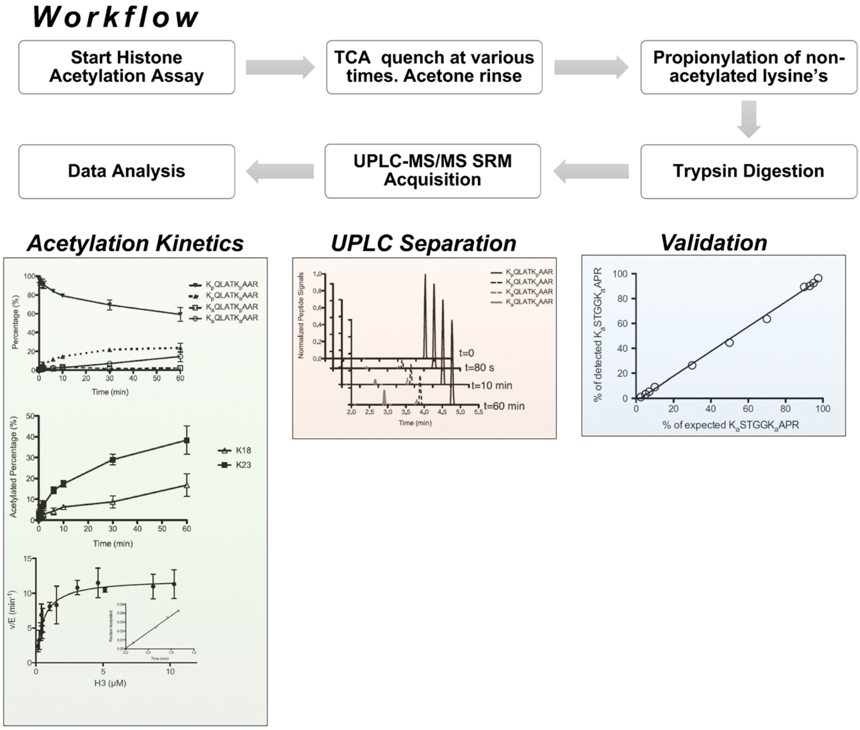

Figure 8.7 Experimental workflow of the multiplexed MS-based assay. Comparison of the detected percentage of K_aSTGGK_aAPR *versus* the expected percentage of K_aSTGGK_aAPR in the presence of propionylated peptides (R^2 = 0.998) shows no significant differences in ionization efficiency. A UHPLC chromatogram of the time-dependent acetylation kinetics for all species of peptide ^{18}KQLATKAAR^{26} is shown, as are the percentages of each state calculated for each state as a function of time. The steady-state kinetic parameters were determined from fits to the Michaelis–Menten equation for acetylation of Gcn5 on histone H3.

As an example, the appearance of $^{18}K_pQLATK_aAAR^{26}$ prior to $^{18}K_aQLATK_p$-AAR^{26} was quantitated in a time-dependant manner by summing the fractions of all peptides containing the modification over the reaction time (Figure 8.7). Steady-state kinetic parameters were determined by fitting fractional percentage acetylation as a function of time using the Michaelis–Menton equation,

$$\frac{v}{[E]} = k_{\text{cat}} \frac{[S]^{nH}}{\left([S]^{nH} + K_{(\text{app})}^{nH}\right)} \tag{8.2}$$

where v is the initial rate, S is the concentration of histone or acetyl-CoA (depending on titration), E is the concentration of Gcn5, and nH is the Hill coefficient. The Hill coefficient is applied to peptides exhibiting sigmoidal kinetics in histone acetylation, suggesting that a single acetylation accelerates subsequent acetylation in histones. This combination of an accurate, quantitative label-free MS method for steady-state analysis of histone acetylation kinetics is highly applicable to broad-based *in vitro* studies of histone post-translational modifications.

8.4 Pre-steady-state Kinetics

Clarke and coworkers developed an online quench-flow reactor interfaced to a Fourier transform ion cyclotron resonance (FTICR) MS for high mass resolution and high temporal resolution enzyme kinetic experiments.[18] The hydrolysis of *p*-NPA by α-chymotrypsin was chosen as a model system because of its well-characterized pre-steady-state kinetics in the literature.[19] Enzyme reactions were initiated by pumping 55 μM α-chymotrypsin with a series of *p*-NPA concentration (1–15 mM) through a nanomixer and into a fused-silica reaction capillary (50 μm i.d., 375 μm o.d, length 25 mm, 75 mm) (Figure 8.8). A 99 : 1 methanol/formic acid solution was used to quench the reaction immediately prior to ESI. Flow rates of 3–200 μL min^{-1} produced reactions times of 50–5000 ms, and were monitored on an Apex Ultra Qh FTICR MS equipped with a 12 T superconducting magnet and an electrospray ion source. The unfolded mass spectrum for the δ′-form of the enzyme was observed and calculated from isotopic fits to the $[\text{M} + 20\text{H}]^{20+}$ ion (Figure 8.8). Build-up of the acyl-enzyme intermediate was monitored over time for four different concentration of *p*-NPA and fit to eqn (3):

$$[ES]'(\text{t}) = C\{1 - \exp(-k_{\text{obs}}t)\}. \tag{8.3}$$

The values obtained for K_d, k_2, and k_3 were 1.6 ± 0.3 mM, 2.8 ± 0.2 s^{-1}, and 0.0 ± 0.2 s^{-1}, respectively. Because the enzyme intermediate could be isolated by *m*/*z*, it was possible to subject it to top-down electron capture dissociation (ECD) fragmentation, generating a total of 177 fragments (Figure 8.8). Based on these fragments, the covalently bound acyl-group of the enzyme intermediate was isolated to a five amino acid region, which included the catalytic Ser195 residue. This robust semi-automated workflow makes high-resolution FTICR broadly applicable for monitoring pre-steady-state enzyme kinetics, as well as making it uniquely suited for characterizing enzyme bound intermediates by ECD fragmentation.

8.5 Binding Constants

MS has enormous potential as a tool for direct, facile measurements of protein complexes, particularly with respect to stoichiometry and K_D. However, caution must be taken with the interpretation ESI-MS spectra of protein complexes in particular, since both false positives and false negatives are possible. Cubrilovic *et al.* reported a nanoESI MS method for determining

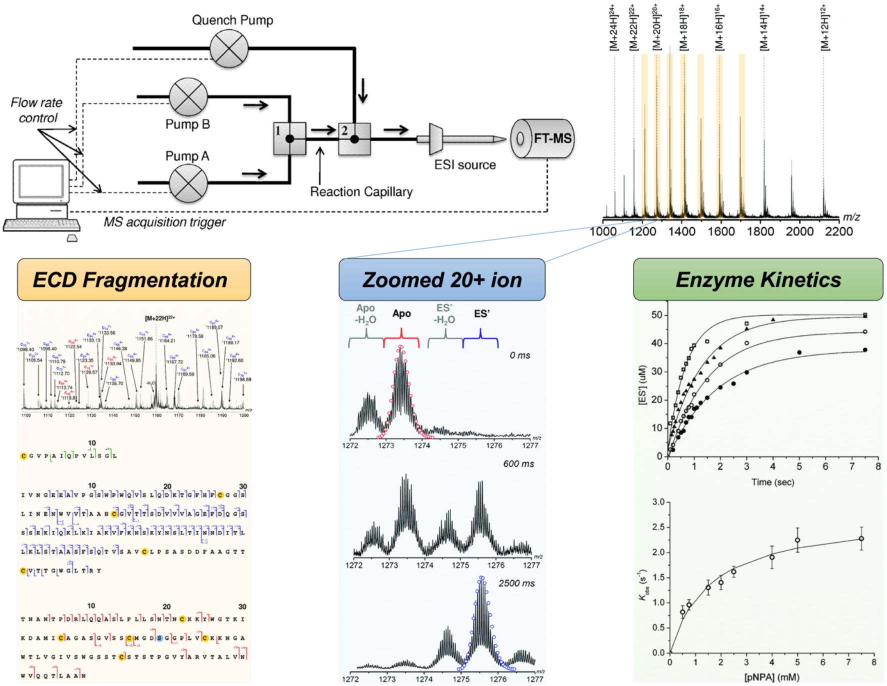

Figure 8.8 A schematic diagram of the quench-flow micro-reactor used for time-resolved ESI-FTICR MS studies. Pumping and acquisition were computer controlled, allowing for automated enzyme kinetic workflows. Also shown is an unfolded δ′-chymotrypsin mass spectrum with a charge state distribution ranging from 12+ to 24+: the orange windows surrounding the 17+ to 21+ ions of the electrospray ion species were subjected to ECD fragmentation; ECD cleavage maps for the A-chain (green), B-chain (blue), and C-chain (red) of δ′-chymotrypsin. High-resolution FTICR MS spectra for the 20+ ion depicts acylation of chymotrypsin at 0 ms, 600 ms, and 2500 ms. Pre-steady-state kinetics for the appearance of the electrospray ion species for 2.5 mM(●), 0.75 mM(○), 2 mM(▲), and 5 mM(□) *p*-NPA. The values of k_{obs} as a function of substrate concentration provide values for K_d, k_2, and k_3.

protein–ligand binding constants and cooperativity in the 146.8 kDa tetrameric enzyme fructose 1,6,bisphophotase (FBPase).[20a] A series of five sulfonylurea-class inhibitors that bind to an allosteric AMP binding site were incubated at varying concentrations (0.5–15 μM) with 2.8 μM FBPase, and free and bound FBPase complexes were differentiated using MS (Figure 8.9). Dissociation constants were reported by taking the ratio of the peak intensity between free and bound FBPase complexes at fixed concentrations of ligand and protein (5 μM and 2.8 μM, respectively) using eqn (4).

$$\frac{\text{bound}}{\text{total}} = \frac{[L]^n}{K_D + [L]^n}. \tag{8.4}$$

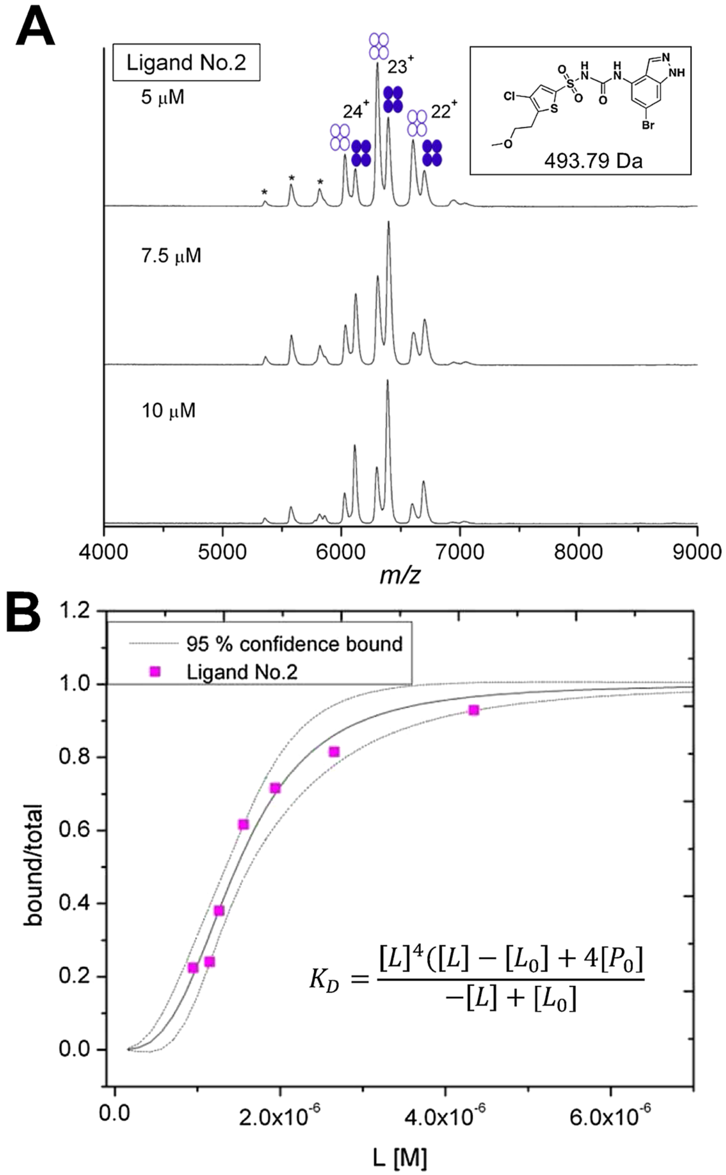

Figure 8.9 (A) FBPase–ligand 2 complex tetramer spectrum. Free FBPase (open circles) decreases while the FBPase–ligand 2 complex increases with increasing ligand 2 concentration. (B) Titration curve for the FBPase–ligand 2 complex. Dissociation constants (Table 8.4) were determined from the fits to the inset equation.

Table 8.4 Dissociation constants (K_D) and Hill coefficients (n) for five inhibitor-FBPase complexes determined by the nanoESI-MS titration method.

Inhibitor	IC_{50} (μM)[a]	K_D (μM)[b,d]	nH[b,d]	K_D (μM)	N[c,d]
1	0.33	3.55 ± 0.38	3.82 ± 1.65	6.66 ± 1.35	2.28 ± 1.0
2	0.14	1.47 ± 0.14	3.10 ± 1.02	6.0 ± 0.85	2.1 ± 0.6
3	0.05	0.60 ± 0.03	5.59 ± 1.60	3.75 ± 0.56	2.35 ± 0.93
4	1.66	4.49 ± 2.29	4.14 ± 1.37	8.38 ± 1.05	2.5 ± 1.0
5	0.53	4.24 ± 0.37	3.79 ± 1.50	7.6 ± 0.98	2.28 ± 0.78

[a] IC_{50} values determined in solution.[20b]
[b] Data for extracted L.
[c] Data for assumption $L = L_0$.
[d] The error is given as the standard deviation calculated from three different measurements and is based on a 95% confidence interval.

Ligands 2, 3, and 4 gave ratios of 0.74, 1.38, and 0.39 respectively, where a higher ratio denotes higher affinity. The binding affinity order determined by MS matched the order determined by IC_{50} measurements. The measured Hill coefficients (n) in eqn (2) range from 3.10 to 5.59, suggesting positive cooperativity (Table 8.4). Direct visualization of individual ligation states and the associated conformational changes that occur on binding can be quickly extracted from a mass spectrum, highlighting MS's unique capability of studying the cooperative mechanisms of ligand binding. This fast, label-free approach enables MSs to become even more valuable, especially for drug development, providing a framework for future studies on protein drug targets.

8.6 Allosteric Regulation

MS combined with biophysical techniques such as covalent labeling,[21] chemical crosslinking,[22] oxidative labeling,[23] and hydrogen/deuterium exchange[24] (HDX) provides a method for studying the structure of functioning enzyme systems.[25] Recent work by Sowole *et al.* investigated the degree of allosteric control in the 136 kDa enzyme dihydrodipicolinate synthase (DHDPS) by hydrogen deuterium exchange MS.[26]

HDX was initiated with a 9 : 1 dilution of D_2O and sample solutions. The final concentrations for allosteric inhibitors, Lys and bisLys, were 2.4 mM and 30 μM, respectively, in 2 μM protein. The reaction was quenched (HCl, pH 2.4) at various times (1–120 min) followed by online digestion with pepsin and peptide separation, achieved using a nanoAquity UHPLC connected to a Synapt G2 for mass analysis. Sequence coverage for 27 peptides totaled 96% of DHDPS. Global HDX on the intact protein level for DHDPS–Lys and DHDPS–bisLys showed significant reduction in deuterium incorporation when compared to DHDPS, which suggests a stabilization of conformational dynamics on inhibitor binding. Spatially resolved HDX difference maps, where deuterated DHDPS peptides at 60 mins are subtracted from DHDPS–Lys and DHDPS–bisLys peptides, revealed several regions of the protein that exhibit significantly lower deuterium exchange, specifically, peptides 249–260, and 270–288, which line the pyruvate substrate access channel but are distant from the allosteric binding site (Figure 8.10). The authors concluded that DHDPS allosteric regulation by Lys and bisLys occurs through restricted conformational dynamics of the substrate access channel, preventing access of the substrate to the active site. This finding shows that significant dynamic changes owing to allosteric regulation may not be captured by X-ray crystal structures, confirming the necessary role of HDX-MS as a valuable tool for revealing the conformational dynamic of proteins.

8.7 Catalysis-linked Dynamics

Recent work by Vahidi and co-workers provides a comprehensive look into the conformational dynamics associated with the 525 kDa catalytically active F_0F_1 ATP synthase complex.[27] The F_0F_1 ATP synthase is a membrane bound "rotor" enzyme that is responsible for both synthesis and hydrolysis of ATP. A large proton motive force (PMF) is responsible for driving ATP synthesis, while a small PMF is responsible for driving ATP hydrolysis.[28]

An *Escherichia coli* membrane mimic was used to embed ATPase and an ATP regeneration system using pyruvate kinase and phosphoenol pyruvate[29] with a sustained k_{cat} = 11 ± 1 s^{-1} for F_0F_1 ATP synthase. HDX experiments were conducted on a nanoAquity UHPLC and a Waters Synapt G2. Three catalytic conditions were studied: (1) an ADP-bound inhibited state, I_{ADP}. (2) A working, or catalytically active, state where the inner-membrane PMF is high, W_{PMF}. (3) A working state where carbonyl cyanide-4-(trifluoromethoxy) phenylhydrazone (FCCP) is used to prevent proton build-up in the membrane, W_{FCCP} (Figure 8.11). Deuterium incubation times ranging from 10 s to 45 min resulted in ~30 000 catalytic events per enzyme during the 45 min HDX time window. Three γ sub-unit peptides (Figure 8.12) for I_{ADP} and W_{FCCP} states exhibit similar deuteration kinetics across all incubation times, indicating similar dynamic behavior. The C-terminal helix peptide γ ^{260}LQLVYNKARQASITQEL^{276} in the W_{PMF} state, however, undergoes time-dependant destabilization during catalytic turnover. The authors concluded that the γ sub-unit exerts mechanical stress on the $\alpha_3\beta_3$ sub-unit, and when under a PMF load, the C-terminal portion of the γ sub-unit over-twists, causing hydrogen bonds to destabilize. Vahidi and co-workers provide an excellent mathematical explanation for this in their supporting information.[27]

8.8 Conclusions and Future Directions (MS: Is It One-size Fits All for Studying Enzyme Mechanisms?)

Across all biological systems, measurements of enzyme turnover rates show that for the most part, catalysis occurs on the millisecond timescale.[30] Characterizing the subset of reactions that occur much faster requires greater

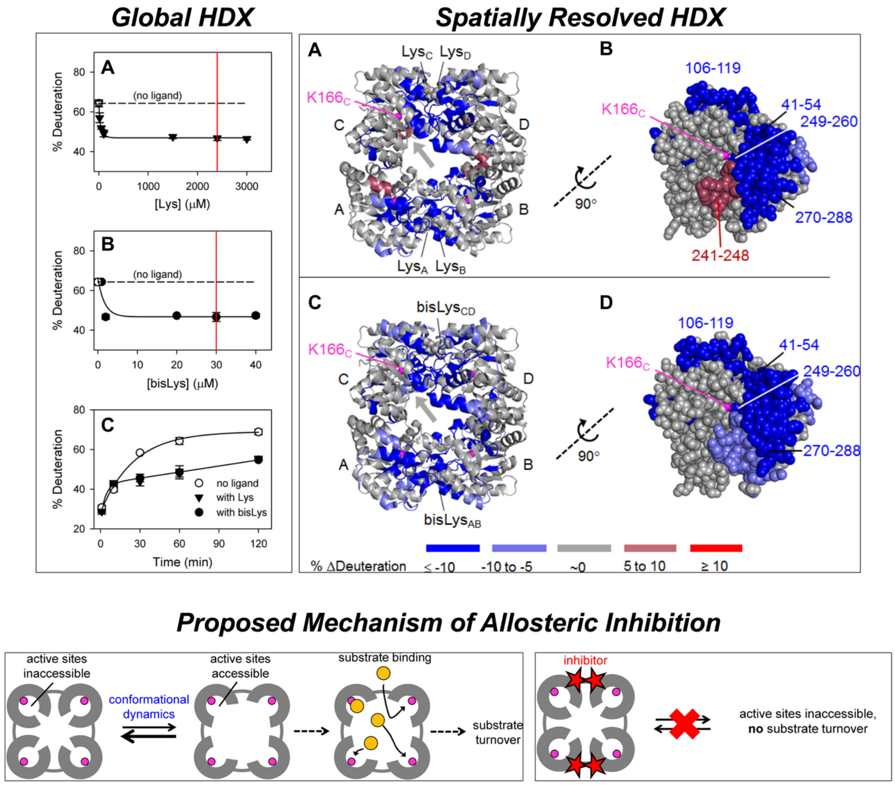

Figure 8.10 Global HDX: deuterium exchange on intact DHDPS with (A) an increasing concentration of Lys, and (B) an increasing concentration of bisLys. Red lines indicate the concentrations suitable for spatially resolved HDX measurements. (C) A plot of global hydrogen deuterium exchange as a function of incubation time for DHDPS, DHDPS–Lys and DHDPS–bisLys. Spatially resolved HDX: difference in HDX mapped onto the DHDPS crystal structures for (A) Lys and (C) bisLys. Space-filling structures for (B) Lys and (D) bisLys viewed from the central cavity of sub-unit C. The catalytic K166 is indicated in pink, while peptides surrounding the substrate cavity are also labeled. Proposed mechanism of allosteric inhibition: left: thermally active structural fluctuations in a small pore in DHDPS allowing access of the substrate to the active site and substrate turnover. Right: binding of Lys or bisLys to the allosteric site restricts conformational fluctuations, preventing the substrate from entering the access channel, inhibiting substrate turnover.

time resolution than current rapid mixing devices have to offer.[31,32] Ultrafast mixing on the microsecond timescale coupled to MS for monitoring protein folding was recently achieved by Mortensen and Williams using theta-glass emitters.[33] Microsecond mixing was also achieved using a fused droplet ESI-MS method monitoring hydrogen/deuterium exchange for the peptide bradykinin.[34] Advancements in rapid mixing will not only open the door to

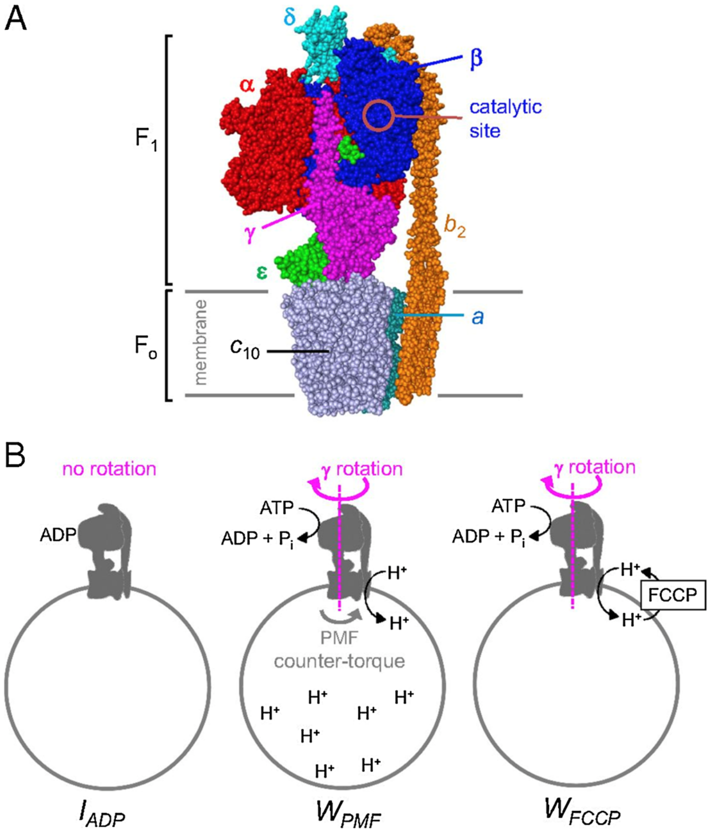

Figure 8.11 Sub-unit architecture of F_0F_1 ATP synthase from *E. coli*. (A) A model assembled from PDB codes 3OAA, 3J0J, 1C17, 2XOK, and 2WSS. (B) Schematic diagram of a membrane vesicle-bound F_0F_1 under different experimental conditions: an ADP-bound inhibited state (I_{ADP}), a working or catalytically active state where the inner-membrane PMF is high (W_{PMF}), and a working state where carbonyl FCCP is used to prevent proton build-up in the membrane (W_{FCCP}).

a greater understanding of enzyme mechanisms, it can also, in combination with structural MS techniques such as HDX, potentially shed light on catalytically coupled conformational dynamics, a technique recently applied using MS,[35] but previously limited to Carr–Purcell–Meiboom–Gill nuclear magnetic resonance (CPMG NMR) spectroscopy.[36]

Microfluidic systems are an attractive and inexpensive total-analysis solution for studying enzyme mechanisms using MS.[37,38] Recently, a microfluidic platform was able to monitor low microsecond to millisecond timescale reaction kinetics for the oxidative labeling of lysozyme.[31] A digital microfluidic system was used to monitor the pre-steady-state kinetics of protein tyrosine phosphatase by matrix-assisted laser desorption ionization time-of flight MS.[39] The Wilson group has developed a number of microfluidic system coupled to ESI-MS for protein folding,[40] hydrogen/deuterium exchange,[41]

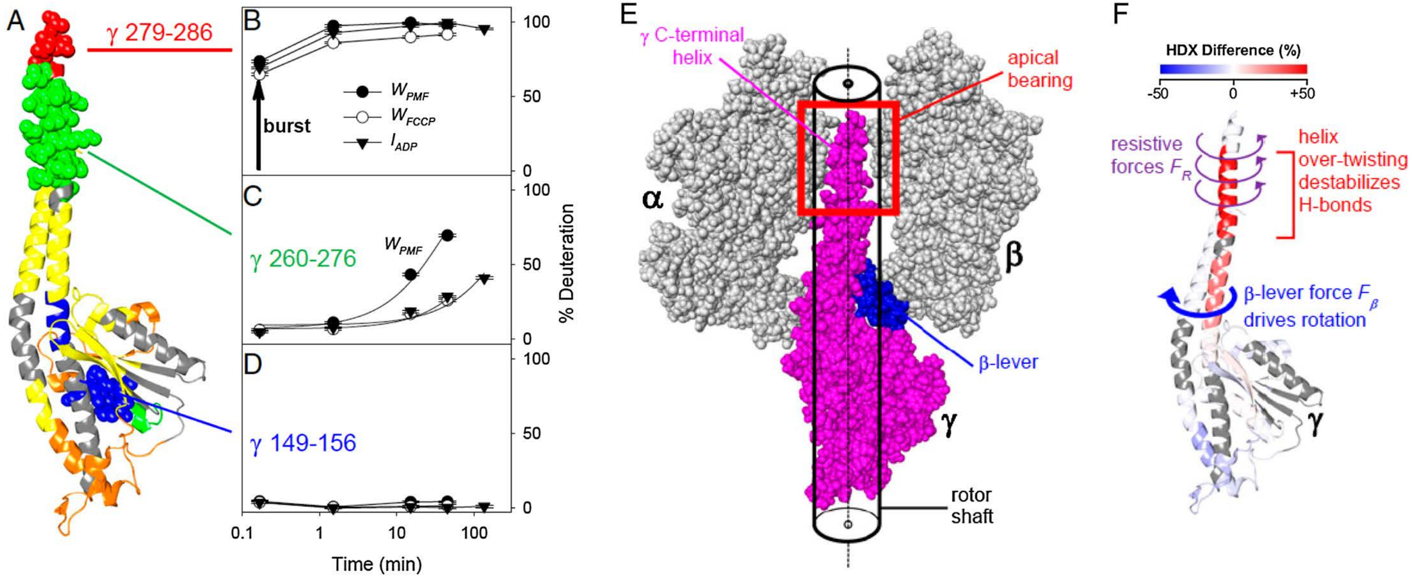

Figure 8.12 Deuterium uptake of selected peptides within the γ sub-unit. (A) X-ray structure of γ coloured by deuterium uptake percentages (blue 0–20%, green 20–40%, yellow 40–60%, orange 60–80%, red 80–100%). The three peptides highlighted as space-fill representations are ^{279}IVSGAAAV^{286}, ^{260}LQLVYNKARQASITQEL^{276}, and ^{149}IGPVKVML^{156}. (B)–(D) HDX kinetics of these peptides under W_{PMF}, W_{FCCP}, and I_{ADP} conditions, fit to single-exponential expressions. (E) Zoomed-in view of γ/$\alpha_3\beta_3$ contacts (PDB ID code 3OAA). Surface roughness in the apical bearing causes F_{R} during γ rotation. (F) Interplay between F_{R} and F_β induces torsional stress in the γ C-terminal helix during each power stroke. It is proposed that this torsional stress destabilizes H-bonds and accelerates HDX in this region.

and pulsed-labeling HDX.[42] The versatility of fully integrated microfluidic total-analysis platforms has recently been demonstrated for automated millisecond protein-ligand interactions[43] and capillary electrophoresis ESI for hydrogen/deuterium exchange.[44]

Multi-enzyme reaction monitoring in both cell-based and cell-free scenarios is becoming increasingly feasible with modern MSs and automated pumping equipment. Cell-free systems offer the advantage of using non-natural substrates, diminished mass transfer effects with an absent cell membrane, and improved control over reaction kinetics.[45] This was demonstrated recently by Hold *et al.* through a ten-enzyme cascade reaction for monosaccharide synthesis, where 17 compounds were monitored simultaneously as a function of time using a novel experimental apparatus coupled to a 4000 QTRAP MS.[46] In total, 11 mechanistic enzyme rate laws were derived with 60 parameters, such as affinities and Hill coefficients, as well as the kinetic characterization of the intermediate dihydroxyacetone phosphate. Shlomi *et al.* described a ^{13}C metabolic flux assay monitoring methionine metabolism in human fibrosarcoma cells by LC-MS.[47] Metabolic levels were quantified by measuring the ratio of endogenous metabolite to an internal standard, and then to the ratio of unlabelled standards. Real-time monitoring of the metabolic flux has also been achieved for *E. coli*, where over 300 compounds were tracked over several hours.[48]

Ultimately, MS offers enormous, and still largely untapped, potential for studying all aspects of enzyme function. The distinguishing feature of MS as an analytical tool remains its extreme selectivity, which, in the context of enzymes translates into the capacity to independently monitor multiple co-existing (unlabeled) species. This is ideal for investigating complex enzyme mechanisms, particularly when coupled to rapid mixing for the detection of transient intermediates. When combined with structurally sensitive techniques such as oxidative labeling, HDX or ion-mobility spectrometry, the opportunities for new insights into enzyme structure, dynamics and function expand still further. The transformation of MS into a multifaceted bioanalytical tool has been underway for over thirty years, and there is still much room for growth. MS is still at the early stages of development as a tool for studying enzyme mechanisms, but its trajectory is clearly to become one of the most powerful and commonly used tools in the enzymologist's toolbox.

References

1. D. Dovala, *et al.*, Structure-guided enzymology of the lipid A acyltransferase LpxM reveals a dual activity mechanism, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, E6064–E6071.
2. M. D. Figueiredo, M. L. Vandenplas, D. J. Hurley and J. N. Moore, Differential induction of MyD88- and TRIF-dependent pathways in equine monocytes by Toll-like receptor agonists, *Vet. Immunol. Immunopathol.*, 2009, **127**, 125–134; C. Whitfield and M. S. Trent, Biosynthesis and export of bacterial lipopolysaccharides, *Annu. Rev. Biochem.*, 2014, **83**, 99–128.

3. Z. Li, F. Song, Z. Zhuang, D. Dunaway-Mariano and K. S. Anderson, Monitoring Enzyme Catalysis in the Multimeric State: Direct Observation of Arthrobacter 4-Hydroxybenzoyl-Coenzyme A Thioesterase Catalytic Complexes using Time-resolved Electrospray Ionization Mass Spectrometry, *Anal. Biochem.*, 2009, **394**, 209–216.
4. Z. Zhuang, K.-H. Gartemann, R. Eichenlaub and D. Dunaway-Mariano, Characterization of the 4-Hydroxybenzoyl-Coenzyme A Thioesterase from Arthrobacter sp. Strain SU, *Appl. Environ. Microbiol.*, 2003, **69**, 2707–2711.
5. M. T. Nelp, Y. Song, V. H. Wysocki and V. A. Bandarian, A Protein-derived Oxygen Is the Source of the Amide Oxygen of Nitrile Hydratases, *J. Biol. Chem.*, 2016, **291**, 7822–7829.
6. P. Cook and W. W. Cleland, *Enzyme Kinetics and Mechanism*, Garland Science, 2007.
7. H. Gutfreund and J. M. Sturtevant, The mechanism of the reaction of chymotrypsin with p-nitrophenyl acetate, *Biochem. J.*, 1956, **63**, 656–661.
8. P. Liuni, E. Olkhov-Mitsel, A. Orellana and D. J. Wilson, Measuring Kinetic Isotope Effects in Enzyme Reactions Using Time-Resolved Electrospray Mass Spectrometry, *Anal. Chem.*, 2013, **85**, 3758–3764.
9. H. R. Mahler and J. Douglas, Mechanisms of Enzyme-catalyzed Oxidation-Reduction Reactions. I. An Investigation of the Yeast Alcohol Dehydrogenase Reaction by Means of the Isotope Rate Effect1,2, *J. Am. Chem. Soc.*, 1957, **79**, 1159–1166.
10. H. Park, G. Kidman and D. B. Northrop, Effects of pressure on deuterium isotope effects of yeast alcohol dehydrogenase using alternative substrates, *Arch. Biochem. Biophys.*, 2005, **433**, 335–340.
11. R. A. Hess, A. C. Hengge and W. W. Cleland, Isotope Effects on Enzyme-Catalyzed Acyl Transfer from p-Nitrophenyl Acetate: Concerted Mechanisms and Increased Hyperconjugation in the Transition State, *J. Am. Chem. Soc.*, 1998, **120**, 2703–2709.
12. R. Bentley, The shikimate pathway–a metabolic tree with many branches, *Crit. Rev. Biochem. Mol. Biol.*, 1990, **25**, 307–384.
13. J. Simithy, G. Gill, Y. Wang, D. C. Goodwin and A. I. Calderón, Development of an ESI-LC-MS-Based Assay for Kinetic Evaluation of Mycobacterium tuberculosis Shikimate Kinase Activity and Inhibition, *Anal. Chem.*, 2015, **87**, 2129–2136.
14. L. A. Rosado, *et al.*, The Mode of Action of Recombinant Mycobacterium tuberculosis Shikimate Kinase: Kinetics and Thermodynamics Analyses, *PLoS One*, 2013, **8**, e61918.
15. Y. Gu, *et al.*, Crystal structure of shikimate kinase from Mycobacterium tuberculosis reveals the dynamic role of the LID domain in catalysis, *J. Mol. Biol.*, 2002, **319**, 779–789.
16. C. Han, *et al.*, Discovery of Helicobacter pylori shikimate kinase inhibitors: bioassay and molecular modeling, *Bioorg. Med. Chem.*, 2007, **15**, 656–662.

17. Y.-M. Kuo, R. A. Henry and A. J. Andrews, A quantitative multiplexed mass spectrometry assay for studying the kinetic of residue-specific histone acetylation, *Methods*, 2014, **70**, 127–133.
18. D. J. Clarke, A. A. Stokes, P. Langridge-Smith and C. L. Mackay, Online Quench-Flow Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Elucidating Kinetic and Chemical Enzymatic Reaction Mechanisms, *Anal. Chem.*, 2010, **82**, 1897–1904.
19. D. J. Wilson and L. Konermann, Mechanistic studies on enzymatic reactions by electrospray ionization MS using a capillary mixer with adjustable reaction chamber volume for time-resolved measurements, *Anal. Chem.*, 2004, **76**, 2537–2543.
20. (a) D. Cubrilovic, *et al.*, Determination of Protein–Ligand Binding Constants of a Cooperatively Regulated Tetrameric Enzyme Using Electrospray Mass Spectrometry, *ACS Chem. Biol.*, 2014, **9**, 218–226; (b) P. Hebeisen, *et al.*, Orally active aminopyridines as inhibitors of tetrameric fructose-1,6-bisphosphatase, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 3237–3242.
21. V. L. Mendoza and R. W. Vachet, Probing protein structure by amino acid-specific covalent labeling and mass spectrometry, *Mass Spectrom. Rev.*, 2009, **28**, 785–815.
22. A. Sinz, Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions, *Mass Spectrom. Rev.*, 2006, **25**, 663–682.
23. P. Liuni, S. Zhu and D. J. Wilson, Oxidative Protein Labeling with Analysis by Mass Spectrometry for the Study of Structure, Folding, and Dynamics, *Antioxid. Redox Signaling*, 2014, **21**, 497–510.
24. L. Konermann, J. Pan and Y.-H. Liu, Hydrogen exchange mass spectrometry for studying protein structure and dynamics, *Chem. Soc. Rev.*, 2011, **40**, 1224–1234.
25. L. S. Busenlehner and R. N. Armstrong, Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry, *Arch. Biochem. Biophys.*, 2005, **433**, 34–46.
26. M. A. Sowole, S. Simpson, Y. V. Skovpen, D. R. J. Palmer and L. Konermann, Evidence of Allosteric Enzyme Regulation *via* Changes in Conformational Dynamics: A Hydrogen/Deuterium Exchange Investigation of Dihydrodipicolinate Synthase, *Biochemistry (Moscow)*, 2016, **55**, 5413–5422.
27. S. Vahidi, Y. Bi, S. D. Dunn and L. Konermann, Load-dependent destabilization of the γ-rotor shaft in FOF1 ATP synthase revealed by hydrogen/deuterium-exchange mass spectrometry, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 2412–2417.
28. W. Junge and N. Nelson, ATP Synthase, *Annu. Rev. Biochem.*, 2015, **84**, 631–657.
29. S. M. Resnick and A. J. B. Zehnder, *In Vitro* ATP Regeneration from Polyphosphate and AMP by Polyphosphate:AMP Phosphotransferase and Adenylate Kinase from Acinetobacter johnsonii 210A, *Appl. Environ. Microbiol.*, 2000, **66**, 2045–2051.

30. A. Bar-Even, *et al.*, The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters, *Biochemistry (Moscow)*, 2011, **50**, 4402–4410.
31. L. Wu and L. J. Lapidus, Combining Ultrarapid Mixing with Photochemical Oxidation to Probe Protein Folding, *Anal. Chem.*, 2013, **85**, 4920–4924.
32. N. Zinck, A.-K. Stark, D. J. Wilson and M. Sharon, An Improved Rapid Mixing Device for Time-Resolved Electrospray Mass Spectrometry Measurements, *ChemistryOpen*, 2014, **3**, 109–114.
33. D. N. Mortensen and E. R. Williams, Ultrafast (1 μs) Mixing and Fast Protein Folding in Nanodrops Monitored by Mass Spectrometry, *J. Am. Chem. Soc.*, 2016, **138**, 3453–3460.
34. J. K. Lee, S. Kim, H. G. Nam and R. N. Zare, Microdroplet fusion mass spectrometry for fast reaction kinetics, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 3898–3903.
35. P. Liuni, A. Jeganathan and D. J. Wilson, Conformer Selection and Intensified Dynamics During Catalytic Turnover in Chymotrypsin, *Angew. Chem., Int. Ed.*, 2012, **51**, 9666–9669.
36. E. Z. Eisenmesser, D. A. Bosco, M. Akke and D. Kern, Enzyme Dynamics During Catalysis, *Science*, 2002, **295**, 1520–1523.
37. A. Oedit, P. Vulto, R. Ramautar, P. W. Lindenburg and T. Hankemeier, Lab-on-a-Chip hyphenation with mass spectrometry: strategies for bioanalytical applications, *Curr. Opin. Biotechnol.*, 2015, **31**, 79–85.
38. D. E. W. Patabadige, *et al.*, Micro Total Analysis Systems: Fundamental Advances and Applications, *Anal. Chem.*, 2016, **88**, 320–338.
39. K. P. Nichols and J. G. E. A. Gardeniers, Digital Microfluidic System for the Investigation of Pre-Steady-State Enzyme Kinetics Using Rapid Quenching with MALDI-TOF Mass Spectrometry, *Anal. Chem.*, 2007, **79**, 8699–8704.
40. T. Rob and D. J. A. Wilson, Versatile Microfluidic Chip for Millisecond Time-Scale Kinetic Studies by Electrospray Mass Spectrometry, *J. Am. Soc. Mass Spectrom.*, 2009, **20**, 124–130.
41. T. Rob, *et al.*, Measuring Dynamics in Weakly Structured Regions of Proteins Using Microfluidics-Enabled Subsecond H/D Exchange Mass Spectrometry, *Anal. Chem.*, 2012, **84**, 3771–3779.
42. T. Rob, P. K. Gill, D. Golemi-Kotra and D. J. Wilson, An electrospray ms-coupled microfluidic device for sub-second hydrogen/deuterium exchange pulse-labelling reveals allosteric effects in enzyme inhibition, *Lab Chip*, 2013, **13**, 2528–2532.
43. Y. Cong, *et al.*, Mass spectrometry-based monitoring of millisecond protein–ligand binding dynamics using an automated microfluidic platform, *Lab Chip*, 2016, **16**, 1544–1548.
44. W. A. Black, B. B. Stocks, J. S. Mellors, J. R. Engen and J. M. Ramsey, Utilizing Microchip Capillary Electrophoresis Electrospray Ionization for Hydrogen Exchange Mass Spectrometry, *Anal. Chem.*, 2015, **87**, 6280–6287.

45. S. Billerbeck, J. Härle and S. Panke, The good of two worlds: increasing complexity in cell-free systems, *Curr. Opin. Biotechnol.*, 2013, **24**, 1037–1043.
46. C. Hold, S. Billerbeck and S. Panke, Forward design of a complex enzyme cascade reaction, *Nat. Commun.*, 2016, **7**, 12971.
47. T. Shlomi, J. Fan, B. Tang, W. D. Kruger and J. D. Rabinowitz, Quantitation of Cellular Metabolic Fluxes of Methionine, *Anal. Chem.*, 2014, **86**, 1583–1591.
48. H. Link, T. Fuhrer, L. Gerosa, N. Zamboni and U. Sauer, Real-time metabolome profiling of the metabolic switch between starvation and growth, *Nat. Methods*, 2015, **12**, 1091–1097.

CHAPTER 9

Chemical Biology Databases

ALAN C. PILON*[a], ANA PAEZ-GARCIA[b], DANIEL PETINATTI PAVARINI[a] AND MARCUS TULLIUS SCOTTI[c]

[a]Núcleo de Pesquisa em Produtos Naturais e Sintéticos (NPPNS), Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil; [b]Noble Research Institute LLC, Ardmore, Oklahoma, USA; [c]Departamento de Química, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba - Campus I, João Pessoa, Brazil
*E-mail: pilonac@gmail.com

9.1 Introduction

The pursuit of true knowledge is one of humanity's earliest desires. The tripartite theory found in Plato's dialogues asserted that knowledge was based on three principles: a truth, a belief and its justification.[1] Other philosophers have also described various ways of obtaining true knowledge, but even today there is no consensus as to what constitutes true knowledge in itself.

Throughout history, science has served as a lighthouse in the fog of ignorance and has revealed to us that the senses and consciousness are part of what we know as reality.[2,3] In this sense, knowledge can be understood as a result of the interaction between our minds and everything around us.[2–5] In quantum mechanics, this situation is quite common; for example, the paradox of the Schrödinger's cat experiment. When we open the box, that is, when we consciously look at the cat, is when the "magic happens", the collapse of the wave function occurs, the superposition of the cat's states

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

disappears, and finally we realize if the poor cat is dead or alive. Apart from the probabilistic issues broadly discussed in the Copenhagen interpretation of this hypothetical experiment, the most intriguing question to answer is how our consciousness can affect wave collapse and how it works in our daily lives, and, ultimately, in the way we build our knowledge.

Eastern culture has described life and true knowledge in terms of nonlinear complementary dynamic cycles, just like quantum mechanics has demonstrated concepts through elementary questions such as the wave–particle duality or Heisenberg's uncertainty principle.[3,6,7] If the understanding of the complementary functioning of individual atoms reunified concepts of matter and consciousness under a single domain and led to a complete transformation in physics, imagine when we apply this complementary approach to more complex problems, such as what is life and how does it work, or how the self-affirming and integrative parts are related in organisms and, in turn, how they establish our planetary ecosystem.

In life sciences the pace of these transitions between the reductionist and holistic viewpoints has been slower, and a longer time will be required to integrate complementary concepts such as "lock and key" and aspects of cell self-regulation. According to Capra,[6] knowledge of the cellular and molecular aspects of biological structures has led to remarkable advances in science. It is crucial, however, to have a full understanding of life, considering organisms as systems, and replacing the unilateral viewpoint in which organisms work through a set of gears by a concept that emphasizes a stratified organization, including simultaneous interactions and the mutual interdependence of its parts. It is important to note that these concepts are not mutually exclusive and, in fact, are the two sides of the same coin. Reductionism and holism represent the analysis and synthesis of the functioning of all things.

As occurred in physics, when several classical theories were abandoned, the application of a systemic approach in other areas required significant efforts to extract relevant information from older models, or even create new theories. Although this is difficult, chemistry and biology communities are well aware of this necessity, and efforts have been made to integrate their contents based on bioinformatics.[8] The idea is to establish a language among the distinct biological organization levels, including substances (metabolism), organisms (morphology), biomes (biogeography) and functionalities (bioactivities).

In this sense, chemical biology has emerged as a transdiscipline[9] that uses chemistry methods for advances in biology, which in turn favours chemistry with its advances.[10] Although this statement seems to be associated with an avant garde approach, its conception goes back to the science practiced in the 19th century. The urea synthesis developed by Wöhler helped biology to overcome some misconceptions, such as the beliefs about vital force theory and concepts about organic substances. Joseph Priestly, in turn, using rat behaviour as a response to chemical reaction, was able to discover more than ten different gases, including oxygen and nitrous oxide.[11]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

In fact, the great difference between the chemical biology performed in the 19th century and the one we know today is in the technological advances that have occurred over the last 50 years in analytical tools, especially in mass spectrometry (MS) and chromatography, as well as in the use of bioinformatics, the application of high throughput analysis and increased computational processing power.[8,12–14] These advances have allowed the development of disciplines such as synthetic biology, proteomics, chemical genetics, metabolic engineering and metabolomics.

During these 50 years, the low-resolution spectra obtained in the first magnetic-field deflection spectrometers became better and more complex. Ion trap systems, collision cells, time-of-flight (TOF) analysers and signal amplifiers have also dramatically increased spectral quality in terms of resolution, accuracy and sensitivity.[12] High-throughput approaches initially developed for pharmaceuticals were disseminated to several areas of chemistry and biology such as genomics, proteomics and natural product areas.

These unprecedented amounts of data, as well as their complexity, have led us back to the first question on how to obtain true knowledge. Are the human senses and minds capable of interpreting large amounts of data in a viable way? Maybe so, but it can take years of hard work and still bring us flaws and bias.

Within this scenario, databases have emerged as tools for collecting, storing and processing large amounts of data.[15–17] Technologies such as the Internet, database system management, increased secondary memory storage capacity, information systems, and especially the awareness of data sharing and standardization have accelerated the proliferation of hundreds of such databases in all scientific areas.

Since 1971, scientists have deposited Cartesian coordinates of proteins and enzymes in spreadsheet format in one of the oldest biological databases, the Protein Data Bank (PDB).[18] Initially, PDB-like databases had a function to share and provide free access to the experimental results in an easy way and thus ensure long-term availability of data.

Currently, these data repositories have been gradually transformed into "knowledge discovery in databases" (KDDs) or data mining databases. Unlike former repositories, these databases are characterized by distinct layers of non-trivial, interactive and iterative multistep processes to identify understandable, reliable and potentially useful patterns from large amounts of data. Websites such as PubChem, AlzGene, genome browsers and the BLAST sequence-search services are examples of KDDs that have played a key role in the development of chemical biology areas. However, the complete integration of metabolic, proteomic and genomic data is still far from being achieved, and much effort is needed before comprehensive studies involving a systems biology approach can be done.

In this chapter we will discuss how the KDD databases have helped the development of chemical biology in the areas of proteomics, metabolomics and natural products, and we also point out how database architecture has prevented efficient data integration.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

9.2 General Biological Databases

In this section we will discuss the function and utility of databases. Different types of databases will be described, and their actual and potential uses will be highlighted. One of the main objectives of building a database is to compile a large volume of systematized information that can be easily understood and accessed, and that is searchable. A useful database facilitates access from different disciplines to a broad amount of information for different usages. In the case of the Basic Local Alignment Search Tool[19] (BLAST), for instance, the same outcome (percentage of identity between a sample sequence and a sequence listed in the database) can be useful either to locate a gene in a specific organism's genome or to identify gene orthologues in different species. This tool uses the information compiled in the nucleotide and protein databases from the National Center for Biotechnological Information (NCBI).

With the development of genome sequencing techniques, complete genomes are now available for a larger number of organisms. With genetic information more accessible, the desire to know more about the function of the proteins that these genes encode has also increased. The ability to combine different "omics" techniques can fill this gap. Joining efforts to understand the functions of the proteins encoded by the newly discovered genes will lead researchers to greater knowledge of metabolic pathways and to the identification of the metabolites that are the result of these pathways.[20] The outcome of this will be a more holistic understanding of organisms, where they are considered as a whole, not only as the sum of their parts.

The biological response of an organism to a determined environmental condition is important for adaptation to different and constantly changing environments; this is determined by the organism's genetic background. In order to obtain a complete understanding of the biological behaviour of a complex organism, three levels of expression must be considered: RNAs, proteins and metabolites. It is important to understand the links between the functions of different molecules and to construct models to quantify these interactions in order to describe the dynamics of biological processes.[21] This is where functional genomics databases can be important for understanding the interactions between different genes that take part in the same physiological process, and later to link this knowledge to the proteins, and ultimately to the involved metabolic processes.

The bottleneck for a holistic analysis of complex organisms' metabolic pathways resides in the techniques available to acquire samples for further analysis. It is possible and simple to sample complete organisms or even different tissues of one organism. However, when one tries to acquire information at the cellular level, the methodological process becomes more complex. The current efforts to put information together and build metadata banks can be used in the future to better understand the metabolic processes at a cellular or intracellular level, when phenotyping techniques can be improved.

Biological databases can be arranged in many different ways, *i.e.*, as previously mentioned, sorted by the organism target of the study. For this method, the HUGO Gene Nomenclature Committee (http://www.genenames.

org/useful/genome-databases-and-browsers) has done a good job of listing genome databases and browsers in three groups: (1) human, (2) other vertebrates and (3) non-vertebrates.

The NCBI, on the other hand, has chosen to present its database classification based on scope and level of curation following the workflow shown in Figure 9.1.

All these approaches are listed in the databases issue that the journal *Nucleic Acids Research* publishes annually. According to this publication, biological databases can be organized according to the type of structure that is analysed. This way, four main groups of databases can be recognized.

(1) Nucleic acids databases: focused on DNA, RNA and gene expression.
(2) Amino acid/protein databases: protein sequence, protein structure and protein–protein interaction databases.

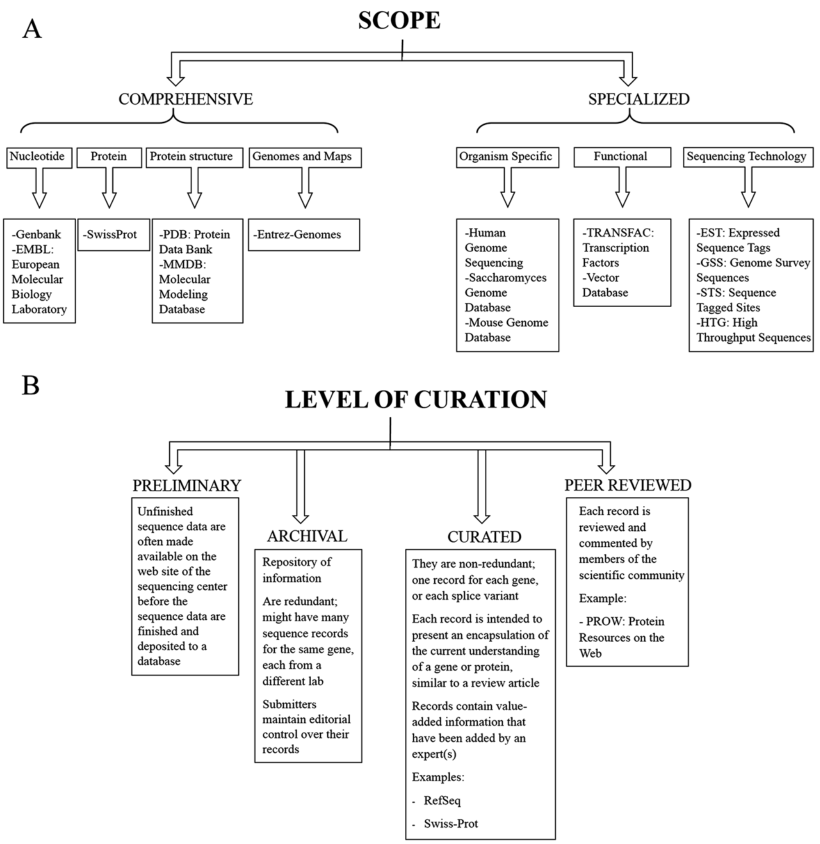
A
SCOPE
COMPREHENSIVE
SPECIALIZED
Nucleotide
Protein
Protein structure
Genomes and Maps
Organism Specific
Functional
Sequencing Technology
-Genbank
-EMBL: European Molecular Biology Laboratory
-SwissProt
-PDB: Protein Data Bank
-MMDB: Molecular Modeling Database
-Entrez-Genomes
-Human Genome Sequencing
-Saccharomyces Genome Database
-Mouse Genome Database
-TRANSFAC: Transcription Factors
-Vector Database
-EST: Expressed Sequence Tags
-GSS: Genome Survey Sequences
-STS: Sequence Tagged Sites
-HTG: High Throughput Sequences
B
LEVEL OF CURATION
PRELIMINARY
Unfinished sequence data are often made available on the web site of the sequencing center before the sequence data are finished and deposited to a database
ARCHIVAL
Repository of information
Are redundant; might have many sequence records for the same gene, each from a different lab
Submitters maintain editorial control over their records
CURATED
They are non-redundant; one record for each gene, or each splice variant
Each record is intended to present an encapsulation of the current understanding of a gene or protein, similar to a review article
Records contain value-added information that have been added by an expert(s)
Examples:
- RefSeq
- Swiss-Prot
PEER REVIEWED
Each record is reviewed and commented by members of the scientific community
Example:
- PROW: Protein Resources on the Web

Figure 9.1 Types of molecular biology database presented by NCBI. (A) Types of databases based on their scope. (B) Types of database based on the level of curation of their data. The databases listed in the figure are only representative examples. Data from the NCBI web page: https://www.ncbi.nlm.nih.gov.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

(3) Meta databases: this type collects relevant information about data. One example of this group is Enzyme Portal (http://www.ebi.ac.uk/enzymeportal/), where information about the biology of enzymes and other proteins with enzymatic activity can be found. It integrates publicly available information about enzymes from different individual databases such as UniProt Knowledgebase, Rhea, ChEBI and CheMBL.

(4) Additional databases: mathematical models, PCR primers, taxonomic, metabolic pathway and signal transduction databases.

It is not the objective of this chapter to list all the available databases, but to provide the reader with some useful tools to help narrow down the list of databases that can be of interest, depending on the objective of the research. In the case of biological databases, and especially those focused on molecular biology, updated and complete information can be found in the 24th annual *Nucleic Acids Research* database issue from 2017.[22] The issue is a compilation of 152 papers describing 54 new databases and updating 98 others. The databases cover a broad variety of molecular biology subjects, comprising genome structure, gene expression and regulation, proteins, protein domains and protein interactions. The description of the new and updated databases is done using visualization tools, with tables that contain the database name, its URL and a brief description. The authors of the issue present a list of databases that have succeeded and have been widely accessed over time. They also propose some rules to follow in order to make a database successful, for example, the selection of a good name that well represents the subject of the database or the provision of thorough data curation.[22] In order for a database to be effective, it is crucial that it is well maintained over time and validated with reliable references, as well to display a user-friendly interface. That is the case of GenBank and UniProt, for instance, which have been developed since the late 1980s and early 1990s. UniProt has multiple and frequent updates, and many citations for this database can be found in PubMed for the year 2017, which can be seen as a measure of success for a database.

In the next sections of this chapter, specific databases for proteomics, metabolomics, drug discovery and natural product studies will be highlighted.

9.3 Databases in Proteomics

Proteomics is the study of the entire diversity of proteins that a single organelle, cell or tissue can display at a given moment.[23] Such diversity includes gene-encoded proteins and products of post-translational modifications. In contrast to genomic analyses, proteome assessment can be used to examine the responses of living organisms to environmental biotic and abiotic stresses. In fact, a single genome, when facing different environmental conditions, can express several proteomes. For this reason, proteomics can be thought of as the primary basis of living organisms' phenotypical comparison.[24]

Traditionally, the identification of proteomes was based on two different approaches: (1) gel-based techniques and (2) shotgun analysis (see Chapter 4 for more information).

Gel-based techniques include several polyacrylamide gel electrophoreses, both mono- and bi-dimensional. The most important one is 2D polyacrylamide gel electrophoresis (PAGE).[25] For immunoblotting, which is gel-based separation followed by specific antibody detection, only known sequences of amino acids, of which proteins are made, or domains of the proteins, are suitable for examination. Alternative separations of complex mixtures of proteins include ion exchange chromatography and reversed-phase chromatography. Gel-based techniques are the most common protocols for targeted proteomics (*i.e.*, the search for a specific protein).

As mentioned above, the sequences of the amino acids that make up the proteins might be known, and so they are available in databases. However, these sequences might also be unknown. Gel-based techniques are a recommended, but not mandatory, separation step prior to MS data acquisition. For unknown sequences, this step is highly recommended.

The final raw data acquired after MS analysis, which will be explored later in this section, is useful for determining the sequences of amino acids that make up the primary structure of protein. Protein annotation refers to determination of the known sequences by the use of integrative online tools such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and BLAStoGO (B2G). This task requires a reference sequence to determine the final sequence by comparison of experimental data and reported data. Whenever this reference is missing, due to lack of interest in the target or lack of studies of the species, the *de novo* sequencing taking place is a quite challenging task. Unknown proteins or unknown sequences of amino acids require *de novo* proteomics protocols in which pieces of sequences might be superposed. Methods using a diverse array of scripts can determine how the fragmented pieces of information can connect to each other without knowing the final sequence. In this case, a previous separation step, such as gel-based or 1D and/or liquid chromatography (LC), are recommended as a step to decrease the complexity of the mixture.[26]

Every proteomics protocol, whether gel-based or gel-free, will rely on MS analysis of peptides from proteolytic digestion of proteins or even integral proteins. Each and every gel-free protocol has the harder task of assembling peptides as pieces of a whole protein once they have no previous separation of proteins.

Gel-free techniques are divided into two types: (1) bottom-up proteomics, which includes the shotgun technique and has a proteolytic step prior to MS injection; and (2) top-down proteomics, in which intact proteins are charged either by matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI) and then injected into the analyser (see Chapter 4 for more information).[27]

Bottom-up techniques have been successful in achieving *de novo* proteomics sequencing. Shotgun, which is the most widely spread bottom-up

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

approach, emerged as a promising time-saving methodology for proteomics in the early 2000s. Lysis of the proteins occurs while these proteins are inside complex mixtures. As it flows, proteins are allowed to directly react with chemo trypsin. A pool of peptides is expressed with no need for previous separation either by LC or PAGE. The strategy relies on the ability of 1D or 2D high-performance LC (HPLC) separation of peptides after the lysis and prior to MS injection. Afterwards, it can be performed as a new separation step in high resolution MS (HR-MS). HR-MS is a troubleshooting technique for unresolved mixtures because it allows the operator to select ions in the analyser stage.[28] Examples of HR-MS are vastly reported using MALDI-TOF, as well as MALDI-TOF/TOF.[28] Nonetheless, Orbitrap is being used as a powerful analyser stage in proteomics nowadays.[29]

Bottom-up approaches are more often used in protein identification, as well as for identification of post-translational modifications when the protein of reference is available.[27] At the end of the experiment, a set of raw MS data is yielded and must be subjected to data processing in order to generate database alignment of peptides and protein sequencing assembly.[30] Another upside of the shotgun approach is that all data generated is compatible with the majority of data available online such as in Protein Information Resource (http://pir.georgetown.edu/).

On the other hand, there is the top-down analysis. The most challenging part of top-down analysis is to generate fragments by using dissociation reactions inside the instrument instead of using a prior proteolysis step before injection.[31] Subsequently, the alignment of the peptides will afford a protein ladder that can be the subject of data searching using online tools. The limitation of top-down data processing is the lack of data resources devoted to it. Peptides afforded by dissociation into the spectrometer are different than the peptides yielded by proteolysis previous to injection on the spectrometer. Hence a new set of databanks is required for top-down experiments due to this difference. The Consortium for Top Down Proteomics (http://repository.topdownproteomics.org/) is currently leading efforts to provide public repositories regarding these new frontiers.

A few quantitative proteomics strategies have been proposed in the literature and are commercially available. Tagging the peptides within a complex mixture is the mechanism that allows quantification. The mechanisms were first proposed in the late 1990s and early 2000s with isotope-coded affinity tags and stable isotope labelling of amino acids in culture. However, the isobaric tag for relative and absolute quantitation (iTRAQ) is nowadays still the first choice for quantitative proteomics and widely applied in health sciences. Proteomics gave birth to new fields of study such as sub-proteomes and interactomes.

Since a proteome will cover the entire diversity of proteins of a living organism or cell, this is time consuming. In order to be more straightforward whenever a study addresses a specific set of proteins, the sub-proteomic or organellar proteomic emerged in the 2000s.[32]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

Interactome comprises all interactions of molecules related to cellular physiology from nucleic acid digestive proteins, such as RNAse and DNAse, to ubiquitination enzyme action. Protein-to-protein interaction is a theme within the interactome subject and is explored nowadays as a source of evolutionary insight.[33] The interactomes of model organisms such as *Arabidopsis thaliana* have already been reported.[34] We believe that semiochemicals are suitable to be studied with the knowledge of interactomes, based on examples such as *Helicobacter pylori* colonization of the gut[35] and more transient complex interactions.[36]

The final step of a proteomics or sub-proteomics protocol is the MS-data processing. Once the spectra are acquired, the software can generate the sequences of amino acids represented in each spectrum. This information can then go to the alignment step where the pieces of protein represented by each peptide can be properly tagged. An aligned sequence can be paired with the sequences available in the databanks for the ultimate information of percentage of similarity. Highly similar matches can be obtained in the pathway database KEGG and the Gene Ontology Consortium (http://www.geneontology.org/) to generate protein annotations on the context of metabolism or in general functional classes.

One of the most important databases in proteomics is PRoteomics IDEntifications (PRIDE). It is proposed as a "centralized, standards-compliant, public data repository for proteomics data". It is part of an effort to make all data standardized and available for research communities (https://www.ebi.ac.uk/pride/archive/).

MassIVE (http://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) is also worthy of note as a platform for exchanging MS-based proteomics data. Developed by NIH as repository and as a community resource for interactive, and analysis of, peptide information. It can offer solutions to end-to-end sequencing once datasets are not only shared but also joined together. Additional proteomics databases are available in Table 9.1.

9.3.1 Integrating the Omics Cascade from Transcripts to Proteins: A Successful Case in Plant Science

The combined use of proteomics and transcriptomics has been changing the experimental model of unravelling biosynthetic pathways.[37] As an example, in the field of plant science the whole workflow used to address metabolic questions was changed in the last 30 years by means of combining omics techniques.

The first choice for studying the enzymes behind the synthesis of terpenes, flavonoids and alkaloids used to be the evaluation of radioisotopic decay of marked precursors for tracking reaction steps of juvenile plants in *ex situ* conditions of culture.[38] Recently, the functional characterization of monoterpene, sesquiterpene and diterpene synthases, as well as oxide squalene synthase, was achieved in order to determine the biosynthetic

Table 9.1 Representative examples of databases used in proteomics.

Database	Data available	Uses (site description)	Link
Blast	Nucleotides and amino acid sequencing	"BLAST finds regions of similarity between biological sequences"	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Mascot	Amino acid sequencing	"Peptide Mass fingerprint, sequence query and MS/MS ions search"	http://www.matrixscience.com/search_form_select.html
Uniprot	Amino acid sequencing	"Comprehensive, high-quality and freely accessible resource of protein sequence and functional information"	http://www.uniprot.org/
Pride – EMBL	Proteomics (protein and peptides)	"Public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence"	https://www.ebi.ac.uk/pride/archive/
wwPDB	Crystallographic data of macromolecules	"Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies"	https://www.wwpdb.org/
PeptideAtlas	Proteomics	"PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments"	http://www.peptideatlas.org/
MassIVE	Proteomics	"MassIVE is a community resource developed by the NIH-funded center for computational mass spectrometry to promote the global, free exchange of mass spectrometry data"	http://massive.ucsd.edu/ProteoSAFe/static/massive.jsp
Jpost	MS proteomics editing and integration	"Researchers can now upload their proteome datasets to the jPOST repository where the raw MS data is re-processed using the jPOST standard protocol to automatically generate high-quality databases for data comparison and integration"	http://jpostdb.org/about/

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

pathway steps responsible for chemical diversity in the terpene group. A few bisabolene synthases can be cited as an example, such as the ones from *Abies grandis*[39] steam, *Arabidopsis thaliana* (Brassicaceae)[40] roots and *Picea abies* (Pinaceae)[41] plantulae. Cadinane synthase from *Eleutherococcus trifoliatus*[42] and cotton (*Gossypium* sp.),[43] caryophilene synthases and germacrene synthases, respectively from corn (*Zea mays*)[44] and lettuce (*Lactuca sativa* (Asteraceae))[45] are other examples of successful cloning as a functional characterization of encoding genes of semiochemicals involved in plant–insect direct and indirect signalling. Integrative omics research has also been reported to be successful in unravelling the dynamics of conifer forest networks.[46,47] By using iTRAQ, a quantitative level of proteome analysis combined with comparative transcriptomics has shown how sophisticated the signalling is of elicitor-induced Norway spruce (*Picea abies*) in the early conifer defence response.[48]

Databases that work in the intersection of hierarchical levels of cellular information are already available. NCBI makes available tools such as “blastx” to determine amino acid sequences from translated nucleotide sequences, while “tblastn” allows the opposite. KEGG can generate metabolic pathways by means of integrating proteomics and genomics data (http://www.genome.jp/kegg/). B2G (http://www.blast2go.com/) intends to provide an “all in one bioinformatics solutions for functional annotation of (novel) sequences and the analysis of annotation data” using both proteomics and genomics data. Representative proteomics databases are showed in Table 9.1.

9.4 Databases in Metabolomics

Understanding the cellular function, tissues and consequently the full functioning of living organisms requires an integrated view of the stratified network that involves the different affirming parts of self-regulation and self-organization.[6–8] Gene expression, transcripts, proteins and the whole set of metabolites integrate all these codified and post-translational events that modulate the complex web of interactions between the biological systems and the surrounding environment and, consequently, define the phenote.[49]

Metabolomics is one of these branches that compose the integrative approach of the systems biology and aims to analyse and quantify the multiparametric responses of endogenous and exogenous metabolites in living organisms when submitted to particular stimuli (see chapter 3 for more information).[50–52] Like other omics, the metabolomics pursuit through the molecular dynamics of sample sets to characterize biological events associated with diseases,[53–55] discovery of biomarkers,[51,56] ecological associations,[57,58] environmental disturbances,[59–61] crop improvement, drug discovery,[62–64] genetic manipulation and discovery of natural products.[57,65]

Comprehensive metabolite analysis in organisms is a major challenge for metabolomics. The lack of universal, sensitive and automated methods able to analyse the wide range of concentration, and structural and chemical

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

diversity of the metabolites has required the combined use of modern analytical techniques such as chromatography, nuclear magnetic resonance (NMR) and MS.[66–68] LC and gas chromatography (GC) are commonly used in metabolomics because of their separation efficiency, flexibility, compatibility with different types of samples (blood, urine, tissues, plants, microorganisms) and analytical platforms (NMR and MS).[69,70] NMR spectroscopy offers unique attributes such as non-destructive analysis and greater richness of structural, spatial and correlational information of metabolites. However, compared to MS-based techniques such as LC-MS and GC-MS, it has a huge disadvantage with respect to compromise between selectivity and sensitivity.[56,71]

A metabolomics experiment usually consists of a sequence of steps, starting with the preparation of the sample, including sample collection and processing, followed by an extraction process that can be performed for a specific class, target metabolite or the broadest possible profile of polar or non-polar metabolites.[72–74] Once the samples have been processed, the analytical tools must be configured to acquire all the information of hundreds of metabolites under a multitude of structures and a wide range of concentrations. Then all datasets are statistically processed to reduce complexity and eliminate unwanted variations, such as matrix effects and other undesirable effects, to finally obtain a set of information that answers a hypothetical question.

Analysing data from metabolomics experiments is as difficult as acquiring them. Depending on the properties of the sample and the type of study employed, specific settings are necessary. While this may initially seem beneficial for the expansion of metabolomics in many scientific fields, this versatility means a critical challenge for uniformity and data exchange. To support the integration of data from different formats, units and scales, metabolomics communities have recommended the use of standard protocols that include sharing experimental designs, instrument configurations and analysis conditions in a format easily interpreted by other research groups. Initiatives such as Metabolomics Workbench,[75] ArMet[76] (Architecture for Metabolomics) and MSI[77] (Metabolomics Standard Initiative) represent such efforts from scientific communities and funding agencies to develop public repositories for analysis, monitoring and dissemination of large volumes of data.

In this sense, public databases have played a key role in the organization, management and distribution of a vast amount of datasets generated from metabolomics experiments.[78,79] The purpose of a public database is to extract relevant information for the end user in easy-to-use platforms with extensive query possibilities that allow data entry, analysis, retrieval and visualization.

Metabolomics databases are commonly developed in relational format using Oracle, MySQL, SQLserver, MongoDB and other architectures that include programming languages such as Express/Note, Ruby, Perl, Python, PHP, Matlab and JavaScript.[80] Typically, such databases support spectral-, text- and structure-based searches, retrieving the results through graphs and tables. Advances in chemometrics, bioinformatics and computational tools have benefited the implementation of sophisticated statistical tools such as clustering, classification and correlation.[81–84]

According to Go[80] and collaborators, metabolomics databases can be classified into four categories. (1) spectral reference databases commonly used to identify and classify metabolites. NIST, GMD, MassBank, GNPS and MetLin are some examples of this group. (2) Compounds, species and databases of the metabolite profile of specific compounds are responsible for the correlation of metabolites with target biological activities or associations of drug discovery. In this category, we can mention PubChem, Human Metabolome DataBase (HMDB), Dictionary of Natural Products and DrugBank. (3) Databases of metabolic pathways that are directly associated with regulatory pathways, including proteins and genetic interactions. The databases in this group are KEGG, WikiPathways, BioCyc, HumanCyc and Reactome. Finally, (4) LIMS (Laboratory Information Management System) metabolic databases consisting of experimental project repositories, raw data, sample details, laboratory and spectral information. SetupX, LIMS Sesame, Metabo LIMS and MeMo, Metabolomics Workbench, GNPS, XCMS online and MetaboLights are some examples of this category. In the following sections, we will provide an overview of the state-of-the-art in databases that are commonly used in metabolomics for all categories previously described. Mass-based repositories are so far the most widely used and therefore will be the main focus in this chapter.

9.4.1 Spectral Reference Databases

The spectral reference databases are the most used repositories for metabolomics. The GC-MS system was the first mass-based technique that implemented standard electronic ionization conditions (70 eV), allowing storage and comparison of mass spectra. Since then, the comparison of reproducible datasets has become the main methodology for spectroscopists in the difficult task of molecular identification. Since the 1990s, a step forward was achieved as a result of the increase in computational processing power, allowing the spectral combination between the target and standard spectra to be performed in an automated way, converting the spectral profiles into scalar vectors.[85] This has simplified the detection and annotation mode of known compounds, called dereplication, and has opened a window for large-scale analysis of thousands of metabolites in mammals, plants and microorganisms.[86–88] Currently, this is the main methodology for data analysis and interpretation of most metabolomic databases.

Throughout the evolution of metabolomics, numerous spectral databases have emerged based on MS techniques, including different ionization modes such as electronic impact (EI), ESI and laser desorption/ionization (MALDI) distributed over several chromatography settings such as liquid, gas and capillary electrophoresis.[89,90] The number of NMR spectral databases has also increased in the last decade. Traditionally, NMR metabolomics databases are based on 2D experiments because of the high degree of signal overlapping

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

in 1D NMR experiments. However, technological advances in deconvolution tools have supported the mono-dimensional spectral matching of ^{1}H and ^{13}C in NMR databases.[91]

Among the most widely used spectral databases are MassBank,[92] Metlin,[93] MMCD,[94] NIST, GOLM[95] and SDBS for MS. For NMR, we can cite BioMAgRes-Bank,[96] NMRshifdb,[97] NMR ChemSpider,[98] the ACD/Labs NMR database and the Sigma-Aldrich NMR database. While most of these databases address life sciences issues, there are many other metabolomics platforms associated with pharmaceutical research, toxicology, forensic investigations, environmental analyses, food control and industrial applications. In Table 9.2, a set of spectral references databases is summarized.

The NIST Mass Spectral Library was initially developed based on EI-MS and quickly became one of the most popular MS spectral databases. In the first versions, NIST was composed mainly of four subunits: (1) main, (2) replicate, (3) salt and (4) MS/MS libraries; in addition to the linear retention index (RI) and Kovats RI to more than 44 000 compounds. Since the NIST14 version, the library has dramatically increased and expanded the number of volatile compounds and lipids, and added new sets of small molecules and peptides. NIST14 also introduced ESI, and atmospheric pressure chemical ionization (APCI) techniques in analysers of the Q-ToF type, triple quad and ion trap. The latest version (NIST17) contains 652 475 MS/MS spectra performed by collision cells or ion trap systems of more than 15 000 different compounds including 1436 biologically relevant peptides.[89]

The NIST17 MS/MS spectral database was acquired using various analytical platforms such as Thermo Scientific, Agilent, Jeol, Waters, Bruker, Sciex, Varian, Applied Biosystems and Fisons in single or combined analysers such as Orbitrap, LTQ, Q-TOF, QqQ, Fourier transform ion cyclotron resonance (FT-ICR), and magnetic and electronic sectors in the fast atom bombardment, ESI, APCI, chemical ionization (CI), liquid secondary ion MS and EI ionization modes. The NIST17 package also includes a series of services: (i) MS Search Software (version 2.3), a free platform for searching for compounds and their corresponding spectra; (ii) the automated mass spectral deconvolution and identification system (AMDIS), an easy-to-use software interface that aims to aid in the discovery and identification of compounds by spectral matching; and (iii) Lib2NIST, a converter that allows the user to add and convert in-house MS spectra into the NIST database. The main disadvantage of NIST17 compared to other spectral databases is that it requires a commercial license.[80]

MassBank is a MS spectral data website formed by a consortium of several mass communities containing free access to more than 17 515 MS and 29 464 MS$_n$ spectral data analysed in positive and negative modes using ionization sources ESI, EI, CI and APCI and more than 15 112 metabolites. As of June 2017, 67.3% of all acquired spectra were associated with the LC-ESI system, followed by 27.1% for the EI mode, 1.4% for CI and 1.3% for APCI. Each substance is described with a set of ontological data and physicochemical properties, including chromatographic methods, retention time, precursor ions,

Table 9.2 Metabolomic databases involving mass spectra, compounds, metabolic pathways and experimental design.

Database	Content	URL (website)
Mass spectral reference databases		
AtMetExpress	LC-MS and MS/MS from *Arabidopsis thaliana*	http://prime.psc.riken.jp/lcms/AtMetExpress/
BinBase	A database system for automated metabolite annotation	http://eros.fiehnlab.ucdavis.edu:8080/binbase-compound/
Bio-MassBank	LC-MS spectra from biological samples	http://bio.massbank.jp/
FiehnLib	GC-MS spectra and RI library	http://fiehnlab.ucdavis.edu/projects/FiehnLib/index_html
GMD@CSB.DB (GOLM Database)	GC-MS spectra, retention time, experimental methods and protocols	http://gmd.mpimp-golm.mpg.de
GNPS	LC-MS and MS/MS from natural products and peptides	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp
LipidBank	LC-MS using orbitrap technology from natural lipids	http://lipidbank.jp/
LipidMaps	LC-MS using orbitrap technology from lipid species	http://www.lipidmaps.org
MaConda	LC-MS of contaminants	http://www.maconda.bham.ac.uk/search.php
MassBank	CE-MS, LC-MS, GC-MS, MALDI-MS including MS/MS spectra from biological samples	http://www.massbank.jp
MassBase	LC-MS, GC-MS and CE-MS for biological samples	http://webs2.kazusa.or.jp/massbase/index.php
METLIN	LC-MS and MS/MS (FT-MS spectra) for biological samples	https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage
MoNA	CE-MS, LC-MS, GC-MS including MS/MS spectra from biological samples	http://mona.fiehnlab.ucdavis.edu/spectra/search
MoToDB	LC-MS from tomato fruit	http://www.ab.wur.nl/moto/
MS/MS Fragmet viewer	LC-FT/ICR-MS and FT-MS/MS from natural products	http://webs2.kazusa.or.jp/msmsfragmentviewer/
MS2T	LC-ESI-QTOF/MS from *A. thaliana*, rice, wheat, and soybean	http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html

(*continued*)

Table 9.2 (*continued*)

Database	Content	URL (website)
MzCloud	ESI-MS, and APCI-MS and MS/MS spectra for organic and inorganic compounds	https://www.mzcloud.org/home
NIST	GC-MS and MS/MS libraries; GC methods/retention time database for most diverse scientific areas	http://nistmassspeclibrary.com
PMR	GC and LC-MS data for plants, eukaryotic and microorganism	http://metnetweb.gdcb.iastate.edu/PMR/
ReSpect	Q-TOF-MS and QqQ-MS spectra for phytochemicals	http://spectra.psc.riken.jp/menta.cgi/respect/search/spectrum
SDBS	EI-MS spectra from organic compounds	http://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi
SoyMet DB	GC and LC-MS from soybean	http://soymetdb.org
The MetabolomeExpress	GC-MS for metabolomics	https://www.metabolome-express.org/
Wiley 10th	EI-MS spectra from compounds	-
Compound-centric databases		
CAS	Database for chemical information gathered from the literature	http://www.cas.org/
ChemBank	Repository for small molecules including biological and medical information	http://chembank.broadinstitute.org
ChEMBL	Database manually curated containing chemical bioactive information about molecules with drug-like properties	https://www.ebi.ac.uk/chembl/
Chemical entities of biological interest (ChEBI)	Comprehensive database for organic molecules with biological interest	http://www.ebi.ac.uk/chebi/
Chemspider	Comprehensive database for chemical structure; physicochemical and biological properties	http://www.chemspider.com
DrugBank	A database that combines detailed drug (*i.e.* chemical, pharmacological and pharmaceutical) data with comprehensive drug target (*i.e.* sequence, structure, and pathway) information	https://www.drugbank.ca

Flavonoid viewer	One of the largest database for known flavonoids	http://webs2.kazusa.or.jp/mfsearcher/flavonoidviewer/
FooDB	Comprehensive resource on food constituents such as macronutrients and micronutrients including features that give foods their flavour, colour, taste, texture and aroma	http://foodb.ca
EssOilDB	Database containing more than 123 000 essential oil reports with data from 92 plant taxonomic families	http://nipgr.res.in/Essoildb/
Human Metabolome Database (HMDB)	A database from human small molecules	http://www.hmdb.ca
KNApSack	A database containing a comprehensive metabolite-species data correlation	http://kanaya.naist.jp/KNApSAcK/
LMSD	A consortium database for natural lipids and lipid species	http://www.lipidmaps.org/data/structure/
Manchester Metabolome database (MMD)	A database for endogenous and exogenous used in metabolomic studies	http://dbkgroup.org/MMD/
MetRxn	It is a collection of metabolite and reaction entities. More than 44 000 metabolites and 35 473 reactions	http://ec2-54-213-167-41.us-west-2.compute.amazonaws.com
NuBBEDB	A natural product database from Brazilian biodiversity	http://nubbe.iq.unesp.br/portal/nubbedb.html
Plant Metabolome Database (PMDB)	Annotated database for metabolites in plants	http://www.sastra.edu/scbt/pmdb/
PubChem	Comprehensive database for small molecules	http://pubchem.ncbi.nlm.nih.gov
***In vivo*/*In silico* Metabolite Database (IIMDB)**	The database includes known biochemical compounds collected from existing biochemical databases plus computationally generated human phase I and phase II metabolites of these known compounds	http://metabolomics.pharm.uconn.edu:2480
Yeast Metabolome Database (YMDB)	A manually curated database of small molecule metabolites found in or produced by *Saccharomyces cerevisae*	http://www.ymdb.ca

(*continued*)

Table 9.2 (*continued*)

Database	Content	URL (website)
Databases of metabolic pathways		
AraCyc	A metabolic pathway database from *A. thaliana*	http://arabidopsis.org/biocyc/
BioCyc	A compendium of pathway databases for several organisms	http://biocyc.org/
EcoCyc	A database that describes the genome, metabolic pathways, and regulatory network of Escherichia coli	https://ecocyc.org
HumanCyc	Database of human metabolic pathways, enzymes, metabolites, and reactions	https://humancyc.org
IPAD	A pathway analysis database	http://bioinfo.hsc.unt.edu/ipad/
iPath	A tool for biological pathway visualization	http://pathways.embl.de
KaPPA-View4	A database for analysis of omics pathway data	http://kpv.kazusa.or.jp/en/
KEGG pathway	A database that describes pathway maps for metabolism, genetic, environmental information	http://www.genome.jp/kegg/pathway.html
MapMan	A web package for analysing omics data	http://mapman.gabipd.org/
MetaCrop	A database for crops pathway analysis	http://metacrop.ipk-gatersleben.de
MetaCyc	Metabolic database that contains pathways, enzymes, metabolites, and reactions from all domains of life	http://metacyc.org/
Pathos	Facility that correlates metabolites identified by MS in the metabolic pathways	http://motif.gla.ac.uk/Pathos/
Pathvisio	A tool for visualizing biological pathways	http://www.pathvisio.org
PlantCyc	A database that contain plant metabolic pathways	http://www.plantcyc.org/
Reactome	A free, open-source, curated and peer reviewed pathway database for the visualization, interpretation and analysis of pathways	http://reactome.org
UniPathway	Manually curated resource for the representation and annotation of metabolic pathways	http://www.unipathway.org/
Wikipathway	A database of biological pathways maintained by and for the scientific community	http://www.wikipathways.org/index.php/WikiPathways

Metabolomics LIMS databases		
COordination of Standards In MetabOlomicS (COSMOS)	A database for metabolomic standards, databases, data exchange and dissemination of metabolomics experiments	http://www.cosmos-fp7.eu
MASTR-MS	A web-based tool for experimental design, sample metadata configuration, and sample data acquisition	https://muccg.github.io/mastr-ms/
Metabolomics Workbench	NIH initiative that contains an inventory and availability of high quality reference standards for data sharing and collaboration	http://www.metabolomicsworkbench.org/about/index.php
MetaDB	A web interface used to store and disseminate metadata and protocols	https://metadb.lafayette.edu
MetaboLIMS	A web-based, portable, robust, and scalable LIMS designed to meet the production and clinical needs in metabolomics research	http://www.hmdb.ca/labm/
MetaboLights	MetaboLights is a database for metabolomics experiments and derived information	http://www.ebi.ac.uk/metabolights/
QTreds	Platform to manage to manage and monitor genomic and metabolomic research providing the location of each sample, experimental protocols and reagent inventory	http://qtreds.crs4.it/home.html
Sesame LIMS	A database that contains experimental protocols, data, sample details and laboratory resources	www.sesame.wisc.edu/
SetupX	A web-based interface that provides experimental design to data reporting and raw data	http://fiehnlab.ucdavis.edu/projects/binbase-setup

high resolution MS data and links to other databases. The MassBank platform enables quick searches using text-based entries or batch-loading data. The user can compare data between the target and the standard spectra using a 3D visualization tool available on the site.[89,92]

One of the advantages of using MassBank is the ability to evaluate a target spectrum in a set of combined spectra obtained from different instrumental configurations (merged spectra) of the same metabolite. This spectral sum increases the chances of spectral matching between the target and standard spectra, reducing the intrinsic instrumental variability of each MS experiment. Another advantage is that all content is downloadable or accessed by an API and the code and supporting information is displayed in GitHub.

On the other hand, MassBank has the disadvantage that not all records are correctly curated, and some entries present data pre-treatment problems. Efforts have been made to develop standard protocols that can help load, process and annotate metabolites and spectra in MassBank.

Spectral mass banks are also found in Europe, such as European MassBank (http://massbank.eu/MassBank/), and the United States of America, such as Massbank of North America (http://mona.fiehnlab.ucdavis.edu/).

Another spectral database, Metlin, has an arsenal of endogenous and exogenous (>240 000) metabolites of plants, microorganisms and humans, as well as drug metabolites, synthetic organic compounds and peptides with three or four amino acid units. Metlin provides LC-MS profiles including all physical-chemical characteristics, structural details, and tandem experiments performed in positive and negative modes on ESI-Q-ToF (Agilent Technologies) using different collision energies (10, 20 and 40 eV).[80,99]

The Metlin web interface offers search tools and spectral matching features in single, batch, fragment, neutral loss and MS/MS modes. The user can perform queries using *m*/*z* text format data, including different types of adducts, chemical formulas and CAS numbers. Metlin also offers MS information linked to identifiers of metabolic path databases, such as KEGG identifiers. All the contents searched can be downloaded in CVS format, including mass error (in ppm), ontological characteristics, names and MS/MS spectra.[99]

Since 2014, the Scripps Research Institute has expanded its content by incorporating isotopically labelled metabolites into a new database called isoMetlin.[93] This database allows the user to search all isotopologues derived from Metlin or by isotopes of interest, such as ^{13}C and ^{15}N. IsoMetlin is designed to support metabolism and biosynthesis studies that use labelling approaches to correctly assign the structure of metabolites.

Perhaps the main common disadvantage of Metlin and IsoMetlin is that MS data is not available for download through these sites; this hampers technological advances in the field, especially by preventing data exchange between different MS databases.

Among MS spectral reference databases, Mzcloud[89] and GNPS[100] are probably the latest and represent the next generation of MS/MS spectrum information extraction tools. MzCloud provides MS/MS and MS$_n$ spectra of over 3000 authentic chemical patterns performed in orbitrap technology and displayed

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

in a web-based interface using a hierarchical architecture to store and search for spectral data called spectra of spectral trees. In this system, the MS/MS spectra obtained from different instruments, experimental conditions, sample preparations and collision energies are stored in nodes based on precursor ions. Thus, a hierarchical network is formed from the subsequent fragmentation of generated ions of the same precursor ion.[101]

If a node has more than two spectra, the interface automatically calculates the average and composite spectra. The spectral tree strategy increases the chances of annotation of compounds by reducing interference such as noise and instrumental intervariability in the process of spectral matching. Additionally, MzCloud displays chromatographic information and data sources, along with experimental parameters and compound identifiers.[101]

The main drawbacks of this database are related to the shrinking number of natural metabolites, content not available for download and a batch service not being offered. In addition, MzCloud exhibits compatibility problems with several web browsers in Mac and Linux systems.

The GNPS database, in turn, constitutes a natural product library (which also includes other metabolites and molecules) with more than 221 000 MS/MS spectra and 18 000 unique compounds obtained from collaborations with other databases such as MassBank, ReSpect and NIST, as well as contributions from metabolomics and NP communities. Unlike the other databases, GNPS represents a crowd-sourced initiative with continuous growth and re-analysis of all sets of data. GNPS establishes new data correlation patterns monthly by reprocessing all data entries (MS/MS spectra) using molecular networking and dereplication tools.[100]

9.4.2 Compound-centric Databases (Metabolic Class-, Species- and Tissue-specific)

Compound-centric databases are fundamental to linking the presence or concentration of chemical entities to upstream information, such as the discovery of biomarkers/drug development, metabolic pathways and biological functions. Also, it provides chemical-biological information that can assist in several steps that involves the structural identification. Traditionally, these databases provide a variety of information associated with physico-chemical properties, including polarity, molecular formula and monoisotopic mass, as well as biological characteristics of metabolites present in biofluids, tissues or organs obtained from experimental data or well-curated literature. These databases exert an important role in connecting the spectral databases, for instance, mass databases, to other features associated with a given metabolite. It is important to note that this category is centred on chemical structure and corresponding ontological descriptors, rather than scientific names or identifiers.[80]

Among the most comprehensive databases are CAS (Chemical Abstracts Service databases) and Dictionary of Natural Products, fee-based services that compile published literature data, as well as public databases such

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

as Chemspider, HMDB,[102] DrugBank,[103] PubChem,[104] Chemical Entities of Biological Interest (ChEBI),[105] ChEMBL,[106] KNApSAck,[107] LMSD,[108] Plant Metabolome DataBase (PMDB),[109] Yeast Metabolome Database (YMDB),[110] Manchester Metabolome Database (MMD),[90] Flavonoid viewer[111] and other important databases described in Table 9.2. A brief discussion of the most comprehensive databases is presented below.

Chemspider and PubChem are public databases and provide, respectively, >59 million and >91 million chemical structures of small molecules. Both databases support text- and structure-based queries, and all of their content is available for download in single or batch mode. Among the available information in both websites are 2D and 3D structures and conformers; biological descriptors, including metabolite class; scientific and common names as well as different identifiers; intrinsic chemical and physical properties; information on medicines, pharmacology, biochemistry and toxicology; literature references and patents; biomolecular interactions and pathways; and links to many other databases of chemical biology such as PDB, HMDB and KEGG.[80]

The database intersection has allowed the construction of universal knowledge environments aggregating spectral, compound, pathway and methodological information in a single platform. PubChem and ChemSpider have already started paving this road, including a set of web services involving all types of biological and chemical information. PubChem BioAssay tools, for instance, provide reports about biological high-throughput screening results, compounds and proteins bioactivity, structure–activity analysis and bioactivity clustering, in addition to spectral visualization tools for MS and MS/MS spectra using links from HMDB and NIST. One of the few limitations of PubChem and Chemspider, is that there is a lack of curation of publicly deposited data from contributors in many studies.[80]

HMDB (version 3.6) is the most comprehensive online human metabolite database, providing detailed information on more than 74 000 small molecules. All HMDB information is classified into three categories: (1) chemical data including physicochemical data, names, synonyms, description of metabolites and chemical structures; (2) clinical data associated with disease biomarkers; and (3) molecular biology/biochemistry data that correlate information with genes/SNP/mutation/enzymatic data and metabolic pathways.[102] Each HMDB metabolite is organized into a set of drop-down menus called "MetaboCard" containing more than 110 fields with clinical, chemical, spectral, biochemical and enzymatic data that can be downloaded in XML format and easily imported into other relational databases. Many data fields have hyperlinks to other databases, such as KEGG, PubChem, MetaCyc,[112] ChEBI, PDB,[18] UniProt[113] and GenBank,[114] as well as offering a variety of structure and pathway views.

The HMDB database allows query searches based on structure, text, sequence and chemical structure. HMDB also provides search and spectral matching services for more than 15 000 compounds distributed in 1D NMR and 2D NMR (3824 spectra), MS (111 164 spectra) and MS/MS (17 765

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

spectra) data.[102] The most recent version provides an MS/MS spectrum visualization tool that allows peak or fragment assignments, such as is found in the METLIN database. HMDB also stores and manages sample information, experimental protocols and laboratory resources using the LIMS (MetaboLIMS) system.

The LIPID MAPS Structure Database (LMSD) is a consortium database composed of an association between LIPID MAPS, LIPID Bank, LIPIDAT, Lipid Library Cyberlipids, ChEBI and other public sources, and covers biologically relevant chemical structures and annotations of lipids. As of January 2017, the LMSD is the most comprehensive lipid database accounting for more than 40 000 unique structures classified according to the design rules proposed by LIPID MAPS and named following the IUPAC and IUBMB scheme.[108]

The LMSD supports text- and structure-based search queries by combining the following data fields: LIPID MAPS ID, scientific or common name, molecular mass and formula, and main class and subclass data fields. LMSD also provides tools to facilitate the representation of different classes of lipid structures and visualization tools that support GIF, JMol, ChemDraw and MarvinView applets. Recovery results are fully annotated and are linked to external databases. They provide users with a variety of online analytics tools that include lipid rating, experimental protocols, paths and discussion forums.

9.4.3 Databases of Metabolic Pathways

Metabolic pathway databases are data repositories that are intended to describe biosynthetic pathways, cell function, organisms and ecosystems from cross-information on genes, proteins/enzymes and metabolites. These databases are freely accessible and represent the biosynthetic pathways of biological systems (*e.g.* human body, plant and microorganisms) through the molecular building blocks of genes and proteins and chemical information using interaction diagrams, reaction networks and relations (wiring diagrams).[80] Many databases also provide information on diseases and medications (health information), such as disorders of the biological system. In MS, these databases aim to organize and facilitate the understanding of how different layers of information obtained from metabolites, amino acids and nucleotide sequences can be interpreted and integrated.

Among the most used metabolic pathway databases are KEGG, BioCyc,[112] MetaCyc, HumanCyc,[115] AraCyc,[116] MetaCrop,[117] UniPathway,[118] Wikipathway,[119] Reactome,[120] ARMeC,[121] EcoCyc,[122] KOMICS,[123] and many others described in Table 9.2.

During research queries in this type of database, it is important to keep in mind that many biosynthetic pathways are incomplete and that changes usually occur due to the discovery of new intermediates and mechanisms governing the biosynthetic pathways. One of the disadvantages of these databases is that many of them are unable to annotate metabolites using MS or NMR spectra. In addition, a prior step of identification is needed to use all the available resources of those platforms. A brief discussion about the most used pathway databases is presented below.

The KEGG database was launched in 1995 by Kanehisa Laboratories and designed to promote understanding on how transcripts, proteins and molecules regulate biosynthetic pathways/metabolism, and ultimately how biological systems interact with the environment. With a set of 16 databases, KEGG is an integrated platform that provides reference knowledge for integration and interpretation of large-scale molecular datasets generated by genome sequencing and other high-throughput techniques.

The complete set of databases that make up the KEGG platform can be subdivided into four categories: (1) system information (KEGG PATHWAY, KEGG BRITE and KEGG MODULE); (2) genomic information (KEGG ORTHOLOGY, KEGG GENOME, KEGG GENES and KEGG SSDB); (3) compound information (KEGG LIGAND); and (4) health information (KEGG DRUG, KEGG DGROUP, KEGG ENVIRON).

Among the subgroups widely used for metabolomics is the KEGG PATHWAY database, which consists of a collection of manually drawn path maps on specific molecular interactions and networks of protein–protein interactions, genetic information, environmental issues, cellular processes, organizational systems and human diseases. KEGG BRITE, a platform for the storage of htext (htext) files containing functional hierarchies and binary relations of many different types of associations between biological aspects and KEGG LIGANDS, which can be divided into four subsets (KEGG COMPOUND, GLYCAN, REACTION AND ENZYME) and was designed to bridge the gap between genomic and chemical information.

As of June 2017, KEGG has more than 511 000 paths (514 path maps) in PATHWAY, more than 185 000 hierarchies in BRITE, approximately 50 000 entries related to chemical compounds, glycans, reactions and drugs, and more than 7500 reactions in LIGAND.

Another metabolic pathway database is BioCyc, a family of over 9300 pathway/genomic databases (PGDBs), including the species *Escherichia coli* (EcoCyc), *Homo sapiens* (HumanCyc), *Bacillus subtilis* (BsubCyc) and 2844 other organisms available in the MetaCyc. These PGDBs are structured in three levels according to the amount of manual update they have received. From level 1 to level 3, manual overhauls are progressively replaced by computational exploration.

BioCyc also provides a number of packages and tools for visualizing and navigating metabolic pathways, including complete metabolic maps of organisms; a browser of genome sequences; a service for data analysis, which includes statistical analysis of gene expression, proteomic or metabolomic data using viewers; a tool called smart tables, which aims to assist biologists and chemists in the analysis of genes and metabolites; a metabolic pathway browser that links specific metabolites in metabolic networks; and a comparative analysis, a tool used for comparing pathways, metabolites, transporters and regulatory networks.

BioCyc releases three new versions every year. All content is downloadable and includes SRI's Pathways Tools software that is faster and more powerful

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

than BioCyc's rendering on the site. This software is available for license for academic and commercial groups. BioCyc also offers a wide range of user guides such as the BioCyc database collection, help pages, online guided tour and downloadable webinars.

The third metabolic pathway database that we want to mention is MetaCrop, a public database that contains manually curated information of more than 60 metabolic pathways found in cultures of agronomically important monocotyledonous and dicotyledonous species. Among these crops are barley (*Hordeum vulgare*), wheat (*Triticum aestivum*), rice (*Oryza sativa*), maize (*Zea mays*), potato (*Solanum tuberosum*), rapeseed (*Brassica napus*), beetroot (*Beta vulgaris*), and two model plants, thale cress (*Arabidopsis thaliana*) and barrel medic (*Medicago truncatula*).

The MetaCrop platform supports text-based searches on metabolic routes, including localization information (species, organs, tissue, compartment and stage of development), transport processes and reaction kinetics. The MetaCrop website provides an easy-to-navigate interface allowing exploration from overview path information to single reactions through visualization tools. It is also possible to look for substances, and the corresponding pathways and associated reactions. The elements available on its website can be exported as SBML files, allowing the user to create specific metabolic models. These models can also be introduced into other software for simulation or data analysis.

9.4.4 Metabolomics Laboratory Information Management System (LIMS) Databases

The LIMS databases emerged in the late 1990s and are based on a computer system (electronic lab notebook) to manage and track laboratory information such as samples, users, protocols, experimental, instrumentation, raw data, data processing, and experimental results.[80]

Like other omics techniques, standardization is a key aspect for the development of the LIMS metabolomics databases, especially for their conversion and integration of data with other information platforms such as genomics, transcriptomics and proteomics. The LIMS databases are fundamental for MS, since the data obtained from them depend on a series of characteristics ranging from the preparation of the sample to the type of ionizers and analysers, as well as software used for the interpretation of data.

If the metabolomics LIMS databases require high levels of standardization, the dynamics of metabolomics, on the other hand, require constant changes and the introduction of new elements into its architecture. This duality has played a key role in the evolution of the metabolomics field, but has also brought major challenges in building and implementing efficient protocols and forms that use a minimal amount of information.

Conscious of the need for data standardization, several LIMS databases have been developed to support metabolomics research. These include Metabolomics Workbench,[75] COSMOS,[124] SetupX,[125] Sesame LIMS Metabolic Modeling,[80] MetaDB, MetaboLIMS, QTreds, MASTR-MS, and others databases, as shown in Table 9.2. A brief discussion of the most commonly used LIMS databases is presented below.[126]

SetupX is a LIMS metabolism platform developed at Fiehn Laboratories (UC Davis University) and designed to store and extract useful information from biological metadata and MS experiments, including methodologies, processing, and reporting. This relational database is integrated into an MS database called BinBase and provides spectral matching and metabolic annotation services as well as a complete set of sample information in a workflow layout called Biosources.[80]

All treatment of samples involving different ontogenic and metabolic stages are recorded in the object SetupX Treatments. The information is managed and stored in these objects following a standard data representation based on an ArMet structure. Any biological system and methodologies are accepted on this platform.

The Metabolomics Workbench is a public knowledge environment containing experimental data and metadata from species, analytical tools, chemical structures, tutorials, and other educational and training resources. The Metabolomics Workbench aims to integrate and analyse large volumes of heterogeneous data from a variety of metabolomics studies. These studies include more than 20 different species, covering humans and other mammals, plants, insects, invertebrates, and microorganisms. Mass and NMR spectrometry data as well as experimental protocols for a range of metabolite classes and various types of sample preparation are also present on this platform.

Collaborations with Metabolights[127] and other databases have facilitated the development of an integration platform in conjunction with the MetabolomeXchange initiative. The goal is to allow the user to compare data in different studies of metabolomics in a single platform.

The COordination of Standards in MetabOlomicS (COSMOS) is an ongoing initiative of the EU Framework Program 7 to implement strategies for storing, exchanging, comparing and reusing metabolomics data.

The available COSMOS content in the website is divided into seven work packages (WPs) that include: WP1 Management, which ensures the efficient organization and functioning of COSMOS by means of communication in forums and group mailings; WP2 Standards Development is related to exchange formats and terminological artefacts used to query metabolomics data and experimental metadata; WP3 Database Management Systems is intended to seek metadata and to create upload facilities to a centralized repository; WP4 Data Deposition develops a harmonized and compatible strategy for data deposition and annotation of metabolites considering all diversity of partners' data; WP5 Dissemination Pipelines is a tool to assist metabolomics researchers in becoming aware of new releases of datasets that may be useful for their research; WP6 Coordination With BioMedBridges and Biomedical ESFRI Infrastructures aims to foster the interaction between COSMOS content and these two projects; and WP7 Outreach and Training is a channel that employs dissemination of COSMOS standards, including scientific publications, workshops, presentations at conferences and stakeholder meetings to reach the wider metabolomics community.[124]

All content of COSMOS follows the recommendations of the metabolomics standards initiative and operate according to two criteria: (i) it uses the ISA-Tab format for experimental information (general-purpose Investigation/Study/Assay tabular format) and (ii) adapts the XML-based formats for the instrument-derived "raw" data types using the proteomics standards initiative (PSI), for example, MzML.

9.5 Databases for Drug Discovery and Natural Products

Publications of studies on modern chemistry date back to the 18th century and their number have risen dramatically since the First World War. With the enormous number of studies, the volume and highly complex nature, efforts have been made to organize them and improve their accessibility over the last 25 years.[128,129]

The Internet has helped to promote a new process of data/information transmission but it has only existed for the past two decades. If a simple Internet search does not provide an answer, new discussion forums often provide answers that previously would have taken days or even weeks of searching. In spite of this, some information can only be obtained in an indirect and relatively time-consuming way; consequently, bank architecture techniques, chemical data, consultation and visualization have constantly developed.[129]

Chemical databases have progressed from being a mere repository of the compounds synthesized or isolated from a biological source to recently becoming powerful research tools for discovering new lead compounds or used for chemotaxonomics purposes.[128,130] The new technologies enable rapid synthesis and high-throughput screening of large libraries of compounds, and have been adopted in almost all major pharmaceutical and biotech companies. New methodologies and advances in spectroscopy have produced hyphenated techniques that combine chromatographic and spectral methods to exploit the advantages of both. In all cases, the databases are becoming so huge and complex that new tools need to be developed to manage these databases. The Internet has made web tools possible, mainly to perform queries that access remote databanks in order to extract information in a simple and rapid way.[131] Queries encapsulate several ideas that can be correlated, and the structure of the compounds and their biological activities are a starting point to understanding the relation of the biological activities and structural requirements; in other words, the physicochemical properties responsible for a determined biological response.

Even though there are several types of chemical database, regarding medicinal chemistry and natural product databases, the chemical structure is essential and useful for medicinal chemistry. There are several ways to record a chemical structure in a databank. Using a string, a linear notation is

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

interesting and useful due to its small size. The SMILES (Simplified Molecular Input Line System) is widely used. Other types of codification are based on the connectivity since two-dimensional structure representation is a graph; therefore the atoms are considered as nodes and the bonds as edges. A useful file format that can encode the structure in both a 2D or 3D fashion was developed by MDL (Molecular Design Limited) called .mol files, and it is possible to combine it with a multiple structure records designated SDF (Structures Data File), which is another file format developed by Molecular Design Limited (MDL). They are text files that adhere to a strict format for representing multiple chemical structure records and associated data fields.[129,130]

9.5.1 Databases for Drug Discovery

Nowadays, there are several databases of compounds that are suitable for drug discovery purposes. Databanks with thousands and even millions of compounds that can be used in the drug design process are available online. Although compounds in available databanks are known worldwide and therefore cannot be patented, these compounds are useful for providing information on designing other compounds with desired activity.[130] Molecular databases are easily found. There are several databases available that differ not only in size but also in nature. Some of these databases are related to molecular pathways; others are large collections of crystal structures, experimental results from biological binding experiments, side effects, drug targets, and others. Databases of small molecules are widely used, and many of them are public, such as the Zinc database,[132] ChemSpider,[98] PubChem,[104] ChEMBL[106] and KEGG DRUG[133] (some of these have been previously described). These databases are useful in drug design, since they can be used to predict possible activities or even targets, or designing new compounds. The available databases are continuously increasing not only in number but also in size and complexity.[134]

The available databases for small molecules differ significantly by their purpose. Some are them are very generic; others report specific information regarding several measures of biological activity *in vivo* and *in vitro* against diverse microorganisms or even a specific target (enzyme). The databases are used to perform several computational approaches such as QSAR (quantitative structure–activity relationship), and ligand-based and structure-based virtual screening. The evolution of web tools makes it possible to perform simple queries to find a determined structure or download a batch comprising hundreds or even thousands of compounds that can be filtered.[135]

The data regarding the structure of the compounds is added in an easy way to a database, encoding each molecule in a simple file format, which must be unique for the standardization (*canonicalization*) algorithm. This file format can be converted to others using an algorithm in order to represent the structure in 2D or 3D. For several databases, it is possible to download the structures of the compounds in .mol (MDL) format. The generation of 3D representation of a structure involves algorithms that combine a specialized system algorithm with rapid methods of conformational analysis to avoid non-bonded interactions. Several software packages that can perform these tasks very quickly using free or proprietary algorithms are available, such as CORINA,[136] Open Babel,[137] RDKit,[138] ChemAxon[139] and Balloon.[140,141]

Performing a substructure search in a small molecules database is not trivial because the structures must be encoded as fingerprints. Fingerprints are bitmaps that represent the connections of a compound between one and a defined number of atoms. A hash function is used to set a determined number of bits; therefore, a bit can encode more than one structural pattern in a compound. These fingerprints can be used to find all compounds that have the same subgraph isomorphism of a specified query.[142–144]

For drug design databases, the biological activities of the compounds are very valuable data. With this information, it is possible to investigate several other analyses such as structure-activity relationships or QSAR, propose a mechanism of action (targets or pathways), new applications for known drugs, and ADMET (absorption, distribution, metabolism, excretion and toxicity) studies.[145] For these purposes, PubChem Bioassay[146] and Chembank[147] are very large and useful databases providing information on millions of screening results. Some databases are specialized to include the affinity data of ligand–protein complexes such as ChemProt,[135,148] PDBind,[149] Biding Moad[150] and AffinDB.[151] DrugBank is a suitable database for investigating several properties and the mechanism of action of known drugs and can be used for repurposing these compounds for the treatment of others diseases.[103] However, there are several data and functionalities that can also be included. For drug design, not only the biological activity and some chemical properties are necessary, but also connecting this information with the selectivity index, ADMET properties, efficacy in functional assays, and/or potential multi-target action is also necessary. Some databases are constantly expanding the data available and including more tools to extract all necessary information. ChEMBL[106] and the PubChem database[146] are huge databases that contain data to perform the first steps of drug design.

For both of these database systems, the interface is clear and user-friendly; for example, it is possible to search by drawing the structure and then searching by substructure or similarity. ChEMBL uses a Marvin JS plugin that uses JavaScript technology but does not need Java installed in the local computer; it therefore works on browsers that do not support Java plugins such as Chrome. Some other kinds of data can be used to find a ligand using Marvin JS, for instance: name, SMILES, .mol (MDL) or InChI (IUPAC International Chemical Identifier).

The resulting list provides several properties that can be used as filters. After selecting a compound, a list of calculated properties and biological activities is displayed. Selecting a specified kind of activity, a list of all values of selected activity is shown. It is possible to perform queries using target name, assay name, document identification (of ChEMBL) or cell tissues.

All lists of activities, even thousands of records, can be downloaded in xls or tab-limited format that includes important information regarding the structure, activities, properties, and bibliography source.[106]

For drug design databases, it is essential that the web tools allow the user to perform easy and fast searches regarding not only the structure, activities and/or properties of compounds, but also how to connect the data previously selected to minimize the processes necessary for screening compounds with potential activity. The visualization tools and some simple algorithms are being implemented in these web tools, including the possibility of connecting with other databases and extracting data; and the upgrade and flexibility of these data management systems are crucial for the success of cheminformatics tools for drug design databases.[106]

9.5.2 Natural Product Databases

The traditional work of a natural product researcher can be summarized as the collection of biota samples, the preparation of extracts with the objective of evaluating them in a variety of biological tests to prioritize them, based on these assays or some other criteria, obtaining compounds that can be bioactive, and/or a new structure.[152] In order to minimize time and costs, the dereplication step, a process known as the rapid characterization of known compounds, has become a strategically important area for natural product research involved in screening programs in several commercial and non-commercial databases.[87,153] These databases can be searched with minimal information, such as chemical structure and biological data from compounds; however, the stage of dereplication requires more information, such as biogeographical and taxonomic information, and the presence of this compound (new or not) in other individuals of the same species, genus, subfamily and family. This information can also help to reduce the number of failures of structural identification by dereplication.

There are large structure-based data collections, such as ChemSpider, PubChem and ChEMBL that can be used for this purpose.[86,98,105,154,155] However, the search for chemical structures in these databases is costly because several false targets among compounds of natural and synthetic origin can be generated.

For this reason, a number of specialized natural product databases were developed that are commercially or freely available and can be searched with only minimal information, for example, the Dictionary of Natural Products (DNP),[156] NAPRALERT,[157] Marinlit for marine natural products, and Antibase for microorganisms and higher fungi materials. However, none of these provide structural collections in a format that can be rapidly integrated into a software package such as ACD/Structure Elucidator.[86]

Natural product databases exhibit a huge range of structural complexity, and due to this property are expected to contribute to the ability of such databases to provide hits.[62,158] These structures are also available in regional databases, for example, NuBBEDB,[159] SANCDB,[160] TM-CM,[161] TCM-Database@

Taiwan[162] and TCMID,[163] and many of these have been used in virtual screening research. In addition, the database information described above includes 2D structures, and several databases have selected methods and tools for generating 3D structures of small organic molecules often being used in structure-based drug design.

In addition to the databases of natural products with a focus on metabolomics studies with relation, species-metabolite, KNApSAcK Family,[107] TIPdb-3D,[164] and, AsterDB are examples of databases where it is possible to search for chemical structure by species and other associated information. Nevertheless, some data is lacking for exact dereplication since information such as exact mass and geographic data have been shown to be very important for this type of study.[165–167]

It is not sufficient to simply focus on the information included in the database; a clean interface, a fast search, a user-friendly format and consistencies across the diversity of operational systems (Microsoft Windows, Mac and Linux) are also necessary. Recently, two systems, SistematX[168] and AsterDB,[169] were created as databases, and they can be used for chemosystematics studies, dereplication and botanical correlation.

SistematX and AsterDB were developed in Java programming language version 8 or higher, and also use JSP technology version 2.1 or higher, using the MySQL database version 5.5.46–0 for Linux to maintain the system data. The system uses JSP to create the pages with specific information to each molecule and dynamic page changes by clicking on certain buttons. Intermediary pages are used to recover information from the database and insert it into the JSP.

Several APIs are used in the SistematX implementation. MarvinJS version 15.7.20, from ChemAxon,[139] is the drawing API, and it is integrated into ChemAxon JChem Web Services, an external online service that transforms the drawn structure into SMILES code, then a JChem API function turns it into a binary fingerprint. This fingerprint is used to search molecules in the database through the molecules structure drawing. The drawn molecule converted to the fingerprint is used as a fragment in the search comparing it to the database's molecule fingerprints if it exists in the database as a fragment. SistematX displays information with respect to general nomenclature as a common name, SMILES, IUPAC name, InChI, InChIKey, and CAS registry number, and some properties such as oxidation number, exact mass, and relative mass.

Additionally, Google Maps, from Google Inc., is an API used to draw maps and locations, and it is used in the system to show on a world map the registered metabolite's localization of origin. The API draws the map and receives locations from the database, two variables of type are used to represent latitude and longitude. A registered molecule may have multiple locations and species attached to it. Using a JavaScript function, it graphically places the locations on the map. When registering, by clicking on the map, it sets a marker at the mouse location and adds a line in the coordinates list below the map for each marker on the map, allowing it to automatically change the position when the values of the latitude or longitude boxes are changed. The coordinates are also transformed into an approximate address, using Reverse Geocoder, a function from the Google Maps API. Therefore all combined data could be very useful for structural elucidation since it is possible to relate species, compounds, exact mass and location of the species, and consequently the compounds, that were extracted.

The SistematX homepage is shown in Figure 9.2. Once the user enters the website (http://sistematx.ufpb.br) a structure search option is observed using the MarvinJS API at the top of the screen (Figure 9.2A). Another three search options are exhibited in the interface. The initial screen of the system with the SMILES (Figure 9.2B), name compound (Figure 9.2C) and species search modes (Figure 9.2D) is also visible.

In the first option, the user can perform the search using a complete drawn structure or molecule skeleton, fragment or substructure, which is important in the cases when the user only remembers some structural characteristics of the molecule, such as functional groups, as well as when the studies require structural similarity, the molecules' groups, or compound families. In addition, it is possible to search using the SMILES code, a chemical notation system capable of representing organic compounds, even the most complex compounds, with simple grammar, a common name or IUPAC name (or part of one of these), and a species search. In this last option, it is necessary to first input the name of the genus (which presents an autocompleting

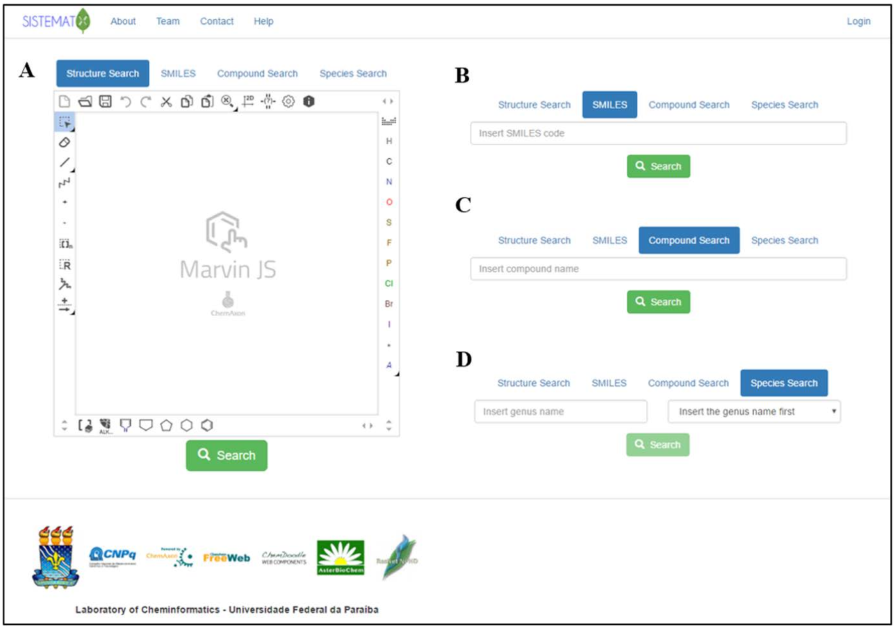

Figure 9.2 SistematX homepage with the different search options: (A) by structure; (B) by SMILES; (C) by compound name; (D) by species.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

option) and, after being selected, the system presents all of the species registered for the genus, and the user then selects a species and performs the search.

When performing a search, the mechanism generates a search results page (six results per page) with the common name; if the compound does not have a common name, IUPAC names are displayed instead. When one of the results is selected, the user has access to the data of that molecule, which is classified into six different groups.

The first group of results that appears is related to the structural representation of the searched molecule. The 2D structure is observed on the interface; on top of this appears the amplify option; by clicking it, the system displays the visualization of the molecule in 2D and 3D (ChemDoodle) and an additional option to save the 2D or 3D structure in an MDL Molfile. The second type of result exhibited by the system is associated with the compound identification. The common name, SMILES code, IUPAC name, InChI code, InChIKey cod and CAS number are all provided. Except for the common name, which is optional and is registered by the administrator, all others parameters are provided by the JChem API.

Compound data results include important characteristics for natural product chemistry. The class of metabolite of the searched molecule and its skeleton provide information about its biosynthetic pathway and aids in chemosystematic and chemotaxonomic studies. The oxidation number (NOX), which is calculated based on Hendrickson rules, is fundamental in chemotaxonomy since Gottlieb related the oxidation grade of molecules with species evolution.[170] Molecular mass is calculated using the most abundant isotope of each element (exact mass) and the average atomic mass of each element (relative mass); these data are important for users that work on the purification process and structural elucidation of molecules, due to the information this provides regarding to the purity of secondary metabolites.

In botanical data, the user can find specific information on the taxonomic rank (from family to species) of the plant from which the molecule structure was isolated and a bibliographic reference that includes journal name, volume, page and year. Because many different species can synthesize the same molecule, there is a register for each species. Meanwhile, biological data obtained in studies related to the biological activity of the searched molecule, type of activity, system, units, activity value and bibliographic references are available in this section.

Plant species have revealed clear genetic signals of local adaptation,[171] and one species can synthesize secondary metabolite or not depending on its location. Variations in the compound concentration in different sites have also been observed. Because geographical data is an important parameter in natural product research, SistematX shows geographical coordinates (latitude and longitude) for each searched molecule and an approximate location of the species from which the metabolite was isolated. In the same way, through the Google Maps API, the user can observe the species' location on the World map.

The databases of natural products available and searchable show some identical data such as chemical structure and botanical occurrence; some of them are used in virtual screening approaches that are useful for optimizing the steps of drug design. However, there are several different kinds of data available in the natural product databases that could be connected, successfully improving the use of information for various applications such as structural elucidation, metabolomics, drug design, *etc.*

9.6 Conclusions

Recent advances in MS data acquisition highlight the technique as a key feature for any metabolism-related research. Each feature of MS, from reliable data acquisition in EI-MS to high-resolution analysers such as Orbitrap, makes the technique central to many analytical issues. Most recently, the advances in user-friendly databases, as well as integrative and dynamic pipelines of data processing, have become pivotal in chemical biology studies.

The chemical biology field of knowledge is heading towards a frontier in which all metabolic processes occurring for living organisms will be trackable at any level. Such a huge challenge has been chased in recent years by many advances, including integrative omics data use. Currently, many efforts to integrate sequencing data from both genomics and transcriptomics to MS-based proteomics and metabolomics are being carried out. Omics DI, for example, intends to be used as an open source tool available for whole-omics integration.

Certainly, several data generation steps must be first standardized in order to have a network of data available to be used by each and every researcher. Next Generation Sequencing and FASTA formats are widespread for nucleic acid data. However, MS-based amino acid sequencing still stands as a challenge in standardization considering the limitations, such as lack of end-to-end sequence data, and misuse, such as the diversity of peptide genesis methods. MS-based metabolomics is also an ongoing issue. The early development of a mass spectra library in EI-MS should be used as an inspiration to shed light onto this field of research.

References

1. F. M. Cornford, *Plato's Theory of Knowledge: The Theaetetus and the Sophist*, Dover Publications, Mineola NY, 2003.
2. L. Schafer, *Zygon*, 2006, **41**, 505–532.
3. A. Goswami, *The Self-aware Universe: How Consciousness Creates the Material World*, Penguin, New York, NY, 1993.
4. R. Lanza and B. Berman, *Biocentrism: How Life and Consciousness Are the Keys to Understanding the True Nature of the Universe*, BenBella Books, Dallas TX, 2010.
5. N. Maxwell, *The Human World in the Physical Universe: Consciousness, Free Will, and Evolution*, Rowman & Littlefield, Lanham MD, 2001.

6. F. Capra, *Futurist*, 1982, **16**, 19–24.
7. F. Capra, *The Tao of Physics: An Exploration of the Parallels between Modern Physics and Eastern Mysticism*, Shambhala Publications, Boston NY, 2010.
8. J. K. Nicholson and J. C. Lindon, *Nature*, 2008, **455**, 1054–1056.
9. W. Colón, P. Chitnis, J. P. Collins, J. Hicks, T. Chan and J. S. Tornow, *Nat. Chem. Biol.*, 2008, **4**, 511–514.
10. K. Kikuchi and H. Kakeya, *Nat. Chem. Biol.*, 2006, **2**, 392–394.
11. K. L. Morrison and G. A. Weiss, *Nat. Chem. Biol.*, 2006, **2**, 3.
12. J. Griffiths, *Anal. Chem.*, 2008, **80**, 5678–5683.
13. W. B. Dunn, *Methods Enzymol.*, 2011, **500**, 15–35.
14. M. Ernst, D. B. Silva, R. R. Silva, R. Z. N. Vêncio and N. P. Lopes, *Nat. Prod. Rep.*, 2014, **31**, 784–806.
15. S. Philippi and J. Köhler, *Nat. Rev. Genet.*, 2006, **7**, 482–488.
16. G. A. Thorisson, J. Muilu and A. J. Brookes, *Nat. Rev. Genet.*, 2009, **10**, 9–18.
17. A. Bender, *Nat. Chem. Biol.*, 2010, **6**, 309.
18. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Int. Tables Crystallogr. Vol. F Crystallogr. Biol. Macromol.*, 2000, **28**, 675–684.
19. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
20. O. Fiehn, *Comp. Funct. Genomics*, 2001, **2**, 155–168.
21. R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali and M. H. Beale, *Trends Plant Sci.*, 2004, **9**, 418–425.
22. M. Y. Galperin, X. M. Fernández-Suárez and D. J. Rigden, *Nucleic Acids Res.*, 2016, **45**, D1–D11.
23. N. L. Anderson and N. G. Anderson, *Electrophoresis*, 1998, **19**, 1853–1861.
24. M. Bourgeois, F. Jacquin, V. Savois, N. Sommerer, V. Labas, C. Henry and J. Burstin, *Proteomics*, 2009, **9**, 254–271.
25. A. Shevchenko, O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie and M. Mann, *Proc. Natl. Acad. Sci.*, 1996, **93**, 14440–14445.
26. J. V. Jorrín-Novo, J. Pascual, R. Sánchez-Lucas, M. C. Romero-Rodríguez, M. J. Rodríguez-Ortega, C. Lenz and L. Valledor, *Proteomics*, 2015, **15**, 1089–1112.
27. B. T. Chait, *Science*, 2006, **314**, 65–66.
28. R. Aebersold and M. Mann, *Nature*, 2003, **422**, 198–207.
29. J. C. Williamson, A. V. G. Edwards, T. Verano-Braga, V. Schwämmle, F. Kjeldsen, O. N. Jensen and M. R. Larsen, *Proteomics*, 2016, **16**, 907–914.
30. C. Abdallah, E. Dumas-Gaudot, J. Renaut and K. Sergeant, *Int. J. Plant Genomics*, 2012, **2012**, 494572.
31. X. Han, A. Aslanian and J. R. Yates, *Curr. Opin. Chem. Biol.*, 2008, **12**, 483–490.
32. R. P. Newton, A. G. Brenton, C. J. Smith and E. Dudley, *Phytochemistry*, 2004, **65**, 1449–1485.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

33. A. Fernández and M. Lynch, *Nature*, 2011, **474**, 502–505.
34. J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar and M. Geisler, *Plant Physiol.*, 2007, **145**, 317–329.
35. K. Stingl, K. Schauer, C. Ecobichon, A. Labigne, P. Lenormand, J.-C. Rousselle, A. Namane and H. de Reuse, *Mol. Cell. Proteomics*, 2008, **7**, 2429–2441.
36. J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees and C. Orengo, *Structure*, 2010, **18**, 1233–1243.
37. K. G. Zulak, D. N. Lippert, M. A. Kuzyk, D. Domanski, T. Chou, C. H. Borchers and J. Bohlmann, *Plant J.*, 2009, **60**, 1015–1030.
38. N. J. Kruger and A. von Schaewen, *Curr. Opin. Plant Biol.*, 2003, **6**, 236–246.
39. J. Bohlmann and J. Gershenzon, *Proc. Natl. Acad. Sci.*, 2009, **106**, 10402–10403.
40. D.-K. Ro, J. Ehlting, C. I. Keeling, R. Lin, N. Mattheus and J. Bohlmann, *Arch. Biochem. Biophys.*, 2006, **448**, 104–116.
41. D. M. Martin, J. Fäldt and J. Bohlmann, *Plant Physiol.*, 2004, **135**, 1908–1927.
42. C.-H. Wen, Y.-H. Tseng and F.-H. Chu, *Holzforschung*, 2012, **66**, 183–189.
43. B. J. Townsend, A. Poole, C. J. Blake and D. J. Llewellyn, *Plant Physiol.*, 2005, **138**, 516–528.
44. T. G. Köllner, M. Held, C. Lenk, I. Hiltpold, T. C. J. Turlings, J. Gershenzon and J. Degenhardt, *Plant Cell*, 2008, **20**, 482–494.
45. N. Ikezawa, J. C. Göpfert, D. T. Nguyen, S.-U. Kim, P. E. O'Maille, O. Spring and D.-K. Ro, *J. Biol. Chem.*, 2011, **286**, 21601–21611.
46. D. E. Hall, J. A. Robert, C. I. Keeling, D. Domanski, A. L. Quesada, S. Jancsik, M. A. Kuzyk, B. Hamberger, C. H. Borchers and J. Bohlmann, *Plant J.*, 2011, **65**, 936–948.
47. T. R. Bonnett, J. A. Robert, C. Pitt, J. D. Fraser, C. I. Keeling, J. Bohlmann and D. P. W. Huber, *Insect Biochem. Mol. Biol.*, 2012, **42**, 890–901.
48. D. N. Lippert, S. G. Ralph, M. Phillips, R. White, D. Smith, D. Hardie, J. Gershenzon, K. Ritland, C. H. Borchers and J. Bohlmann, *Proteomics*, 2009, **9**, 350–367.
49. O. Fiehn, *Plant Mol. Biol.*, 2002, **48**, 155–171.
50. M. Fessenden, *Nature*, 2016, **540**, 153–155.
51. C. H. Johnson, J. Ivanisevic and G. Siuzdak, *Nat. Rev. Mol. Cell Biol.*, 2016, **17**, 451–459.
52. W. Weckwerth and K. Morgenthal, *Biotechnology*, 2005, **10**, 1551–1558.
53. E. M. DeFeo, C.-L. Wu, W. S. McDougal and L. L. Cheng, *Nat. Rev. Urol.*, 2011, **8**, 301–311.
54. J. L. Griffin, H. Atherton, J. Shockcor and L. Atzori, *Nat. Rev. Cardiol.*, 2011, **8**, 630–643.
55. K. A. Stringer, J. G. Younger, C. Mchugh, L. Yeomans, M. A. Finkel, M. A. Puskarich, A. E. Jones, J. Trexel and A. Karnovsky, *Shock*, 2015, **44**, 200–208.
56. J. Wolfender, S. Rudaz, Y. H. Choi and H. K. Kim, *Curr. Med. Chem.*, 2013, **20**, 1056–1090.

57. A. S. Edison, C. S. Clendinen, R. Ajredini, C. Beecher, F. V. Ponce and G. S. Stupp, *Integr. Comp. Biol.*, 2015, **55**, 478–485.
58. O. A. H. Jones, M. L. Maguire, J. L. Griffin, D. A. Dias, D. J. Spurgeon and C. Svendsen, *Austral Ecol.*, 2013, **38**, 713–720.
59. R. Akula and G. A. Ravishankar, *Plant Signaling Behav.*, 2011, **6**, 1720–1731.
60. J. K. Weng, *New Phytol.*, 2014, **201**, 1141–1149.
61. S. Park, Y. S. Seo and A. D. Hegeman, *J. Plant Biol.*, 2014, **57**, 137–149.
62. A. L. Harvey, R. Edrada-Ebel and R. J. Quinn, *Nat. Rev. Drug Discovery*, 2015, **14**, 111–129.
63. B. Haefner, *Drug Discovery Today*, 2003, **8**, 536–544.
64. A. Boufridi and R. J. Quinn, *J. Braz. Chem. Soc.*, 2016, **27**, 1334–1338.
65. N. D. Yuliana, A. Khatib, Y. H. Choi and R. Verpoorte, *Phytother. Res.*, 2011, **25**, 157–169.
66. A. C. Pilon, F. Carnevale Neto, R. T. Freire, P. Cardoso, R. L. Carneiro, V. Da Silva Bolzani and I. Castro-Gamboa, *J. Sep. Sci.*, 2016, **39**, 1023–1030.
67. S. Beisken, M. Eiden and R. M. Salek, *Expert Rev. Mol. Diagn.*, 2015, **15**, 97–109.
68. K. Bingol and R. Bruschweiler, *Curr. Opin. Biotechnol.*, 2017, **43**, 17–24.
69. H. G. Gika, G. A. Theodoridis, R. S. Plumb and I. D. Wilson, *J. Pharm. Biomed. Anal.*, 2014, **87**, 12–25.
70. F. Carnevale Neto, A. C. Pilon, D. M. Selegato, R. T. Freire, H. Gu, D. Raftery, N. P. Lopes and I. Castro-Gamboa, *Front. Mol. Biosci.*, 2016, **3**, 1–13.
71. P. Allard, G. Genta-Jouve and J. Wolfender, *Curr. Opin. Chem. Biol.*, 2017, **36**, 40–49.
72. Y. H. Choi and R. Verpoorte, *Phytochem. Anal.*, 2014, **25**, 289–290.
73. M. Y. Mushtaq, Y. H. Choi, R. Verpoorte and E. G. Wilson, *Phytochem. Anal.*, 2014, **25**, 291–306.
74. M. Beckmann, D. Parker, D. P. Enot, E. Duval and J. Draper, *Nat. Protoc.*, 2008, **3**, 486–504.
75. M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, S. Sumner and S. Subramaniam, *Nucleic Acids Res.*, 2016, **44**, D463–D470.
76. H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele and D. B. Kell, *Nat. Biotechnol.*, 2004, **22**, 1601–1606.
77. L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre and J. L. Griffin, *Metabolomics*, 2007, **3**, 211–221.
78. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, *AI Mag.*, 1996, **17**, 37.
79. T. Imielinski and H. Mannila, *Commun. ACM*, 1996, **39**, 58–64.
80. E. P. Go, *J. Neuroimmune Pharmacol.*, 2010, **5**, 18–30.
81. A. C. Pilon, R. L. Carneiro, F. Carnevale Neto, V. D. S. Bolzani and I. Castro-Gamboa, *Phytochem. Anal.*, 2013, **24**, 401–406.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00221

82. H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, M. Inoue, H. Toki, O. Minowa, T. Noda, J. Kikuchi, K. Auro, J. K. Nicholson, E. Holmes, J. C. Lindon, I. D. Wilson, J. K. Nicholson, J. C. Lindon, J. K. Nicholson, T. Misawa, Y. Date, J. Kikuchi, T. Asakura, K. Sakata, S. Yoshida, Y. Date, J. Kikuchi, B. J. Blaise, J. Carrola, J. Hochrein, S. Lamichhane, J. L. Ward, S. Fukuda, Y. Furusawa, D. M. Ogawa, E. Holmes, M. J. Claesson, T. A. Clayton, M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, F. Wei, K. Ito, K. Sakata, Y. Date, J. Kikuchi, T. K. Karakach, R. Knight, E. M. Lenz, M. R. Viant, J. A. Walter, I. Montoliu, F. P. Martin, S. Collino, S. Rezzi, S. Kochhar, S. Ghosh, A. Sengupta, S. Sharma, H. M. Sonawat, K. Ito, K. Sakata, Y. Date, J. Kikuchi, H. Kaiser, R. B. Cattell, J. L. Horn, R. L. Gorsuch, L. Richard, K. Zoski, S. Jurs, J. Josse, F. Husson, S. Lê, J. Josse, F. Husson, R. Suzuki, H. Shimodaira, H. C. Keun, O. Beckonert, T. Kato, Y. Date, S. Yoshida, Y. Date, M. Akama, J. Kikuchi, K. K. Cheng, I. Rubio-Aliaga, R. Dawson, S. Liu, B. Eppler, T. Patterson, D. R. Wallace, R. Dawson, B. Eppler, R. Dawson, D. Toroser, R. S. Sohal, M. al-Waiz, M. Mikov, S. C. Mitchell, R. L. Smith, C. T. Dolphin, A. Janmohamed, R. L. Smith, E. A. Shephard, I. R. Phillips, S. L. Ripp, K. Itagaki, R. M. Philpot, A. A. Elfarra, S. Fukuda, J. Kikuchi, K. Shinozaki, T. Hirayama, J. Kikuchi, T. Hirayama, Y. Sekiyama, E. Chikayama, J. Kikuchi, Y. Sekiyama, E. Chikayama, J. Kikuchi, F. Delaglio, E. Chikayama, M. Suto, T. Nishihara, K. Shinozaki, J. Kikuchi, E. Chikayama, R. Tauler, B. Kowalski, S. Fleming, R. Tauler, A. A. Smilde and B. Kowalski, *Sci. Rep.*, 2015, **5**, 15710.
83. V. Shulaev, *Briefings Bioinf.*, 2006, **7**, 128–139.
84. B. Worley and R. Powers, *Curr. Metabolomics*, 2013, **1**, 92–107.
85. P. K. Hopke, *Anal. Chim. Acta*, 2003, **500**, 365–377.
86. R. B. Williams, M. O'Neil-Johnson, A. J. Williams, P. Wheeler, R. Pol and A. Moser, *Org. Biomol. Chem.*, 2015, **13**, 9957–9962.
87. D. G. Corley and R. C. Durley, *J. Nat. Prod.*, 1994, **57**, 1484–1490.
88. J. Y. Yang, L. M. Sanchez, C. M. Rath, X. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. De Felicio, A. Fenner, W. R. Wong, R. G. Linington, L. Zhang, H. M. Debonsi, W. H. Gerwick and P. C. Dorrestein, *J. Nat. Prod.*, 2013, **76**, 1686–1699.
89. M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek and O. Yanes, *TrAC, Trends Anal. Chem.*, 2016, **78**, 23–35.
90. M. Brown, W. B. Dunn, P. Dobson, Y. Patel, C. L. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst and A. Tseng, *Analyst*, 2009, **134**, 1322–1332.
91. J. Hao, M. Liebeke, W. Astle, M. De Iorio, J. G. Bundy and T. M. D. Ebbels, *Nat. Protoc.*, 2014, **9**, 1416–1427.
92. H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.

93. K. Cho, N. Mahieu, J. Ivanisevic, W. Uritboonthai, Y. J. Chen, G. Siuzdak and G. J. Patti, *Anal. Chem.*, 2014, **86**, 9358–9361.
94. Q. Cui, I. a. Lewis, A. D. Hegeman, M. E. Anderson, J. Li, C. F. Schulte, W. M. Westler, H. R. Eghbalnia, M. R. Sussman and J. L. Markley, *Nat. Biotechnol.*, 2008, **26**, 162–164.
95. J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie and D. Steinhauser, *Bioinformatics*, 2005, **21**, 1635–1638.
96. E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao and J. L. Markley, *Nucleic Acids Res.*, 2008, **36**, 402–408.
97. C. Steinbeck, S. Krause and S. Kuhn, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1733–1739.
98. H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.
99. C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan and G. Siuzdak, *Ther. Drug Monit.*, 2005, **27**, 747–751.
100. M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, P. C. A. Boya, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.
101. MzCloud Advanced Mass Spectral Database, https://www.mzcloud.org, accessed 1 January 2017.

102. D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert, *Nucleic Acids Res.*, 2013, **41**, 801–807.
103. C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak and V. Neveu, *Nucleic Acids Res.*, 2010, **39**, D1035–D1041.
104. S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.
105. K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj and M. Ashburner, *Nucleic Acids Res.*, 2007, **36**, D344–D350.
106. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2011, **40**, D1100–D1107.
107. F. M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin and L. K. Darusman, *Plant Cell Physiol.*, 2011, **53**, e1.
108. M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill Jr, R. C. Murphy, C. R. H. Raetz and D. W. Russell, *Nucleic Acids Res.*, 2006, **35**, D527–D532.
109. M. Udayakumar, D. P. Chandar, N. Arun, J. Mathangi, K. Hemavathi and R. Seenivasagam, *Med. Chem. Res.*, 2012, **21**, 47–52.
110. T. Jewison, C. Knox, V. Neveu, Y. Djoumbou, A. C. Guo, J. Lee, P. Liu, R. Mandal, R. Krishnamurthy and I. Sinelnikov, *Nucleic Acids Res.*, 2011, gkr916.
111. O. Fiehn, D. K. Barupal and T. Kind, *J. Biol. Chem.*, 2011, **286**, 23637–23643.
112. R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse and L. A. Mueller, *Nucleic Acids Res.*, 2016, **44**, D471–D480.
113. The UniProt Consortium, *Nucleic Acids Res.*, 2014, **43**, D204–D212.
114. D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers, *Nucleic Acids Res.*, 2013, **41**, 36–42.
115. P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker and P. D. Karp, *Genome Biol.*, 2004, **6**, R2.
116. L. A. Mueller, P. Zhang and S. Y. Rhee, *Plant Physiol.*, 2003, **132**, 453–460.
117. E. Grafahrend-Belau, S. Weise, D. Koschützki, U. Scholz, B. H. Junker and F. Schreiber, *Nucleic Acids Res.*, 2007, **36**, D954–D958.
118. A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios and A. Viari, *Nucleic Acids Res.*, 2012, **40**, D761–D769.
119. T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo and A. R. Pico, *Nucleic Acids Res.*, 2011, **40**, D1301–D1307.
120. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu and L. Matthews, *Nucleic Acids Res.*, 2005, **33**, D428–D432.

121. D. P. Enot, W. Lin, M. Beckmann, D. Parker, D. P. Overy and J. Draper, *Nat. Protoc.*, 2008, **3**, 446–470.
122. I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil and P. D. Karp, *Nucleic Acids Res.*, 2005, **33**, D334–D337.
123. N. Sakurai, T. Ara, M. Enomoto, T. Motegi, Y. Morishita, A. Kurabayashi, Y. Iijima, Y. Ogata, D. Nakajima and H. Suzuki, *BioMed Res. Int.*, 2014, **2014**, 194812.
124. R. M. Salek, S. Neumann, D. Schober, J. Hummel, A. Rosato, L. Tenori, P. Turano, S. Marin, P. Conesa, K. Haug, P. R. Steve, C. Luchinat, D. Walther and C. Steinbeck, *Metabolomics*, 2015, **11**, 1587–1597.
125. M. Scholz and O. Fiehn, *Pacific Symposium on Biocomputing*, 2006, pp. 169–180.
126. L. P. de Souza, T. Naake, T. Tohge and A. R. Fernie, *GigaScience*, 2017, **6**(7), 1–20.
127. R. M. Salek, K. Haug, P. Conesa, J. Hastings, M. Williams, T. Mahendraker, E. Maguire, A. N. Gonzalez-Beltran, P. Rocca-Serra and S.-A. Sansone, *Database (Oxford)*, 2013, **2013**, bat029.
128. D. G. I. Kingston, *J. Nat. Prod.*, 2010, **74**, 496–511.
129. J. Gasteiger, *Molecules*, 2016, **21**, 151.
130. M. A. Miller, *Nat. Rev. Drug Discovery*, 2002, **1**, 220–227.
131. K. N. Patel, J. K. Patel, M. P. Patel, G. C. Rajput and H. A. Patel, *Pharm. Methods*, 2010, **1**, 2–13.
132. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.
133. M. Sitzmann, I. E. Weidlich, I. V. Filippov, C. Liao, M. L. Peach, W.-D. Ihlenfeldt, R. G. Karki, Y. V. Borodina, R. E. Cachau and M. C. Nicklaus, *J. Chem. Inf. Model.*, 2012, **52**, 739–756.
134. A. S. Reddy and S. Zhang, *Expert Rev. Clin. Pharmacol.*, 2013, **6**, 41–47.
135. J. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea and O. Taboureau, *Database*, 2016, **2016**, bav123.
136. J. Sadowski and J. Gasteiger, *The First European Conference on Computational Chemistry (ECCC 1)*, AIP Publishing, 1995, vol. 330, p. 629.
137. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
138. P. Tosco, N. Stiefl and G. Landrum, *J. Cheminf.*, 2014, **6**, 37.
139. Chemaxon, https://www.chemaxon.com/download/marvin-suite/, accessed 1 January 2017.
140. M. J. Vainio and M. S. Johnson, *J. Chem. Inf. Model.*, 2007, **47**, 2462–2474.
141. P. Sadowski and P. Baldi, *J. Chem. Inf. Model.*, 2013, **53**, 3127–3130.
142. P. Englert and P. Kovács, *J. Chem. Inf. Model.*, 2015, **55**, 941–955.
143. H. O. Villar and R. T. Koehler, *Mol. Diversity*, 2000, **5**, 13–24.
144. J. Sadowski, *Perspect. Drug Discovery Des.*, 2000, **20**, 17–28.
145. D. Lagorce, D. Douguet, M. A. Miteva and B. O. Villoutreix, *Sci. Rep.*, 2017, **7**, 46277.

146. Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B. A. Shoemaker, T. O. Suzek, J. Wang, J. Xiao, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2009, **38**, D255–D266.
147. K. P. Seiler, G. A. George, M. P. Happ, N. E. Bodycombe, H. A. Carrinski, S. Norton, S. Brudz, J. P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N. J. Tolliday, S. L. Schreiber and P. A. Clemons, *Nucleic Acids Res.*, 2007, **36**, D351–D359.
148. S. Kim Kjærulff, L. Wich, J. Kringelum, U. P. Jacobsen, I. Kouskoumvekaki, K. Audouze, O. Lund, S. Brunak, T. I. Oprea and O. Taboureau, *Nucleic Acids Res.*, 2012, **41**, D464–D469.
149. R. Wang, X. Fang, Y. Lu, C.-Y. Yang and S. Wang, *J. Med. Chem.*, 2005, **48**, 4111–4119.
150. M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin and H. A. Carlson, *Nucleic Acids Res.*, 2007, **36**, D674–D678.
151. P. Block, C. A. Sotriffer, I. Dramburg and G. Klebe, *Nucleic Acids Res.*, 2006, **34**, D522–D526.
152. J. W. Blunt and M. H. G. Munro, *Natural Products: Discourse, Diversity, and Design*, Wiley Online Library, 2014, pp. 413–431.
153. T. B. Oliveira, D. A. Chagas-Paula, A. L. Rosa, L. Gobbo-Neto, T. J. Schmidt and F. B. Da Costa, *Planta Med.*, 2013, **79**, SL26.
154. E. Bolton, Y. Wang, P. A. Thiessen and S. H. Bryant, *Annual Reports in Computational Chemistry*, Washington, DC: American Chemical Society, 2008.
155. P. J. Eugster, J. Boccard, B. Debrus, L. Bréant, J.-L. Wolfender, S. Martel and P.-A. Carrupt, *Phytochemistry*, 2014, **108**, 196–207.
156. Dictionary of Natural Products, http://dnp.chemnetbase.com/, accessed 1 January 2017.
157. J. Graham and N. Farnsworth, *Compr. Nat. Prod.*, 2010, **2**, 81–93.
158. D. H. Drewry and R. Macarron, *Curr. Opin. Chem. Biol.*, 2010, **14**, 289–298.
159. A. C. Pilon, M. Valli, A. C. Dametto, M. E. F. Pinto, R. T. Freire, I. Castro-Gamboa, A. C. Andricopulo and V. S. Bolzani, *Sci. Rep.*, 2017, DOI: 10.1038/s41598-017-07451-x.
160. R. Hatherley, D. K. Brown, T. M. Musyoka, D. L. Penkler, N. Faya, K. A. Lobb and Ö. T. Bishop, *J. Cheminf.*, 2015, **7**, 29.
161. S.-K. Kim, S. Nam, H. Jang, A. Kim and J.-J. Lee, *BMC Complementary Altern. Med.*, 2015, **15**, 218.
162. C. Y.-C. Chen, *PLoS One*, 2011, **6**, e15939.
163. R. Xue, Z. Fang, M. Zhang, Z. Yi, C. Wen and T. Shi, *Nucleic Acids Res.*, 2012, **41**, D1089–D1095.
164. C.-W. Tung, Y.-C. Lin, H.-S. Chang, C.-C. Wang, I.-S. Chen, J.-L. Jheng and J.-H. Li, *Database*, 2014, **2014**, bau055.
165. B. L. Sampaio, R. Edrada-Ebel and F. B. Da Costa, *Sci. Rep.*, 2016, **6**, 29265.
166. T. J. Schmidt, S. Rzeppa, M. Kaiser and R. Brun, *Phytochem. Lett.*, 2012, **5**, 632–638.

167. E. Gaquerel, C. Kuhl and S. Neumann, *Metabolomics*, 2013, **9**, 904–918.
168. SistematX, http://www.sistematx.ufpb.br, accessed 1 June 2017.
169. AsterDB, http://www.asterbiochem.org/asterdb, accessed 1 June 2017.
170. O. R. Gottlieb, *Phytochemistry*, 1989, **28**, 2545–2558.
171. T. Züst, C. Heichinger, U. Grossniklaus, R. Harrington, D. J. Kliebenstein and L. A. Turnbull, *Science*, 2012, **338**, 116–119.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

CHAPTER 10

Perspectives for the Future

ANELIZE BAUERMEISTER*[a], LARISSA A. ROLIM[a,b], RICARDO SILVA[c], PAUL J. GATES[d] AND NORBERTO PEPORINE LOPES[a]

[a]Núcleo de Pesquisa em Produtos Naturais e Sintéticos (NPPNS), Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil; [b]Central de Análise de Fármacos, Medicamentos e Alimentos (CAFMA), Universidade Federal do Vale do São Francisco, Petrolina, Pernambuco, Brazil; [c]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA; [d]School of Chemistry, University of Bristol, Clifton, Bristol, BS8 1TS, UK
*E-mail: ane_qui@hotmail.com

10.1 Introduction

Chemical biology has rapidly emerged in the last twenty years, mainly due to the development of many techniques such as confocal microscopy, genetic engineering, mass spectrometry (MS) and robotic screening procedures, along with many others. MS has a crucial role to play in a range of areas within chemical biology. These have been discussed previously in this book. This technique has been used in studies of animal tissue sections, natural products, enzyme technologies, biological fluids, behavioral ecology and others, allowing the identification of a range of molecules that could be signatures or markers of diseases or, in fact, potential targets for new drug

Chemical Biology No. 4
Mass Spectrometry in Chemical Biology: Evolving Applications
Edited by Norberto Peporine Lopes and Ricardo Roberto da Silva

Published by the Royal Society of Chemistry, www.rsc.org

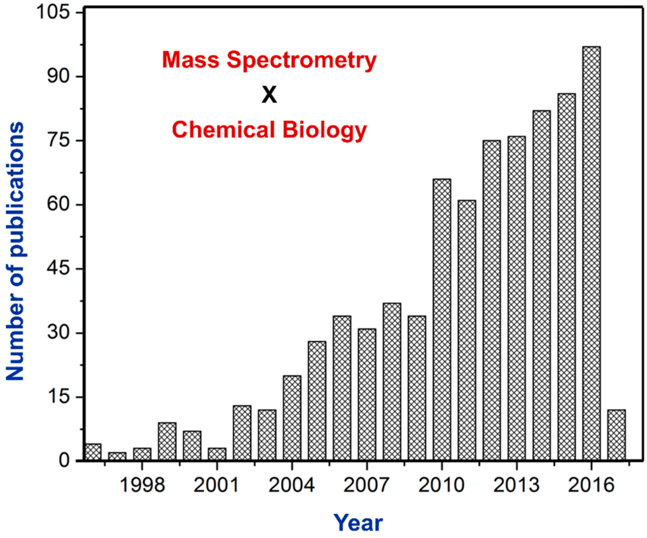

Figure 10.1 Graph showing number of publications *vs.* year (1996–2017) obtained from the Web of Science (March 2017) by crossing referencing the terms “mass spectrometry” and “chemical biology”.

development. The increasing interest in MS applications in chemical biology is clearly demonstrated in Figure 10.1.

In this book, we have presented some potential applications of MS in chemical biology, discussing advances in sample preparation procedures, and also in MS ionization sources and mass analyzers, alongside the increase in spatial resolution and the development of bioinformatics tools for data treatment. All of these advances have led to exponential growth in the field of chemical biology, resulting in the investigation of biological processes through the study of chemical compounds and their interactions.

In addition to the many successes of MS-based approaches in chemical biology, many challenges still remain for the future. For instance, MS has been widely applied to evaluate the mode of action of drugs and their interactions with proteins; however, new strategies are still needed to perform these investigations more closely to their natural or physiological environment. Many techniques and miniaturized devices have been developed and improved upon to allow more specific investigation, especially at the molecular level, in biological systems. The broad range of applications of MS to chemical biology, highlighted throughout this book and this brief introduction, illustrate the challenges for future advances in the field. For this final chapter, we have selected some of the potential and emerging fields of chemical biology in which MS has demonstrated a crucial role, including imaging, ion mobility, microfluidics, single-cell analysis, synthetic ecology and the role of real-time MS in surgical procedures. This chapter will discuss how their improvement could impact in many fields of biology, such as health and environment sciences.

10.2 MS Imaging (MSI)

MSI is an excellent example of an emerging approach that is increasingly arousing the interest of many researchers around the world (see Chapter 6 for a brief discussion). This technique allows two- or three-dimensional visualization of the spatial distribution of different types of molecules, from small organic compounds to large proteins in a biological sample, which has already seen an important impact in the field of chemical biology. MSI was developed as a tool for identifying biomarkers (proteins and/or peptides) *in situ* in biological tissues, with high sensitivity and specificity. It has been applied to scan samples of organs,[1] tissues,[2] gels,[3] or other surfaces where imaging is required. This broad range of applicability highlights the importance of this technique for many fields of science.

The MSI concept was introduced in 1962 by Castaing and Slodzian applying MS to secondary ions (SIMS).[4] However, MSI was largely applied only after the mid-90s with soft ionization development;[5–7] more specifically in 1997, when Caprioli and co-workers[8] started to apply matrix-assisted laser desorption/ionization (MALDI) to imaging biomolecules, such as peptides and proteins, in biological samples. MALDI and desorption electrospray ionization (DESI) are the most-used and well-established ionization techniques currently used for MSI. Laser desorption/ionization MS[9] and time-of-flight secondary ion MS (TOF-SIMS)[10] are other techniques that also have been applied to medical imaging.

Spatial resolution is a key point to obtain the most useful images that are closest to reality—the higher the spatial resolution, the more specific the resulting investigation becomes. Some of the most recently developed MALDI sources have laser beam diameters allowing up to 5 μm of spatial resolution to be recorded; this is smaller than a eukaryotic cell (approximately 10–50 μm). However, we believe that this would need to be reduced to femtometer scale to answer the demands for increasingly specific analyses. Another important point that needs to be considered for MALDI-MSI is the method of deposition of the matrix onto the sample surface. The matrix needs to be delivered as fine droplets onto the surface in order to form a thin homogeneous layer of crystals. This procedure can be done manually; however, automated matrix deposition methods can lead to the formation of a very homogeneous matrix crystal layer.

MSI has presented a significant breakthrough to various areas of science, especially for providing spatial information on the location of chemical compounds. The technique has allowed for the chemical and molecular dialogues of the interactions of microorganism to be studied, contributing to an increased understanding of microbial mechanisms for survival in highly competitive environments such as soils and the rhizosphere.[11] Furthermore, this valuable tool has contributed greatly to studies of the biological function of small molecules in complex systems, which lead to a better understanding of ecological behavior.[12,13] These examples show the importance of MSI in a range of scientific fields.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

Although MALDI is the preferred ionization source for imaging samples, this technique needs many (potentially complex) sample preparation steps, which could make DESI a more attractive ionization source for imaging in some cases. This technique has grown considerably in recent years, mainly due to it being an ambient ionization source, making it very easy to operate with limited sample preparation. The applications of DESI-MS will be discussed in more detail later in this chapter.

10.3 Ion Mobility

The development of the ion mobility spectrometry (IMS) technique began when Bradbury made the first measurement of ion mobility in the gas-phase in 1931.[14] Initially called plasma chromatography or ion chromatography, IMS is a technique for the separation of ions in the gas-phase. The technique works by subjecting ions to collisions with a countercurrent inert gas under the influence of an electric field. The ions are separated based on their molecular weight, charge and collision cross section (which depends on the size, geometry and spatial configuration of the ions). The separation in IMS occurs rapidly, in the order of milliseconds, and coupled to MS (IMS-MS) constitutes an emerging analytical tool for analytical chemical biology. Initially, IMS was only able to analyze volatile compounds. More recently, the development of soft ionization techniques, such as electrospray ionization (ESI), has extended the range of compound types that the IMS technique can be applied to, including non-volatile, thermally unstable and high molecular weight molecules, such as amino acids, peptides, proteins, DNA, polysaccharides and various drug complexes.

Usually, MS techniques are limited to the analysis of the primary protein structure. The coupling of MS with IMS allows the study of more complex structures. The fundamental concepts of ion mobility are governed by eqn (10.1).[15]

$$K = \frac{3}{16}\frac{z}{p}\left(\frac{2\pi k_{\mathrm{B}}T}{\mu}\right)^{1/2}\frac{1}{\Omega_{\mathrm{D}}} \tag{10.1}$$

where z = ion charge; P = gas pressure; μ = reduced mass; k_{B} = Boltzman's constant; T = gas temperature; Ω = collision cross section (CCS).

IMS-MS is able to provide valuable information on the secondary, tertiary and even quaternary structure of proteins.[16] Data obtained from IMS-MS, when supplemented with other classical biophysical methods, has increased the knowledge obtainable from the study of biomolecules and their related complexes. As can be observed in the equation, there are many factors that influence the ion mobility process. The mobility has a dependence on electric field (EF),[17] and for this reason there are a range of commercial instruments available that exploit this in different ways, such as DTIMS (gradient EF), FAIMS (oscillating EF) or TWIMS (adding radio frequency). Better resolution of analytes can be achieved by changing the drift gas properties, which must

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

also be considered, such as the gas itself (helium, nitrogen, argon, carbon dioxide, a mixture, or even chiral modifiers as S-(+)-2-butanol, to improve the separation of chiral compounds), and also the molecular weight, radius and polarizability of the gas,[18–20] as well as the pressure and temperature of the drift chamber.

Another very important factor is the physical properties of the analyte ion, for example molecular weight, charge and CCS. Ions with a higher CCS will collide with a larger number of drift gas molecules, and therefore will remain for a longer time within the drift chamber.[21] Several studies have explored the use of the CCS obtained by IMS-MS, for structural characterization of (for example) heteromeric protein assemblages, showing good correlation with their respective native structures. Protein-binding interactions have also been successfully investigated.[22–24]

10.4 Microfluidics

The requirement for increasingly sensitive tools and methodologies for molecular amplification has led to the development of miniaturized bioanalytical platforms. Microfluidic methods involving biomolecular microanalysis have been attracting increasing interest in the chemical biology field, including metabolomic profiling, environmental monitoring and studies of drug interactions. The small dimension of microfluidics ensures some key advantages such as high speed, low sample and reagent consumption, and the possibility for automation and integration with other techniques for high throughput analysis. Microfluidic technologies have also demonstrated promising applications in single-cell analysis owing to their dimensions being consistent with that of cells. The sensitivity of microfluidics analysis allows the dynamic monitoring of slight cellular and also extracellular secretions.[25]

The growing interest in microfluidic technologies has led to the development of miniaturized systems as well as their integration with other techniques. Microfluidic chips are multi-function miniaturized devices used for sample preparation and detection.[26] The first microfluidic chip was reported in 1979;[27] a miniature silicon wafer based on gas chromatography. Since then, the technology for microfluidic devices has been advancing and other materials have been developed, including micro-electromechanical systems, followed by silicon, inorganic and glass materials.

However, the physical and chemical properties of the new materials still present some limitations for biological applications. With the goal of overcoming this challenge, compatible biological materials, such as polymers (plastics and elastomers) and hydrogels, with facile surface modification, have been developed and are gradually replacing the more traditional materials used.[25] It is important to keep in mind that advances in the field of nanomaterials have made a great contribution to the development of more specific and specialized chips, which could overcome some of the challenges for chemical biology analysis.[28]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

A good example of the development of more efficient microfluidic chips is the achievement of Hattori and co-workers in 2016.[29] They developed a microfluidic shredding chip with high-pressure resistance to extract the RNA from the harder microtissues of skeletal muscle samples. For this purpose, the chip had to be manufactured with a hard material, thus polydimethylsiloxane was used with SU-8 photoresist (epoxy resin), on a flat glass plate. To make the chip permeable, they used a micropillar array combined with physical forces and chemical reagents. The working principle consists of a microchannel with low resistance, which allows the high flow of a suspension of microtissue. The collision between the cells causes the dissolution of the cellular membrane, mainly due to the presence of micropillars and the buffer, leading to the leakage of cell nuclei. The efficiency of this methodology could make it applicable to achieving extraction of any biological molecules from other hard tissues.

Many researchers have endeavored to improve the surface hydrophilicity of microfluidic devices with chemical modifications, allowing protein adsorption, for example. Such approaches could lead to efficient cell isolation and the subsequent unleashing of a whole new set of studies, including single-cell analysis. Several papers have recently been published showing the successful application of microfluidic chips for single-cell analysis (see discussion in the next section).[30–32]

Coupling microfluidic devices with MS can improve and expand the analytical performance for biological applications. Nevertheless, it is a major challenge for microfluidic analysis. It is difficult to get stable and effective interfaces between the chips and the instrumentation.[33] Nevertheless, since the first coupling of microfluidic chips to MS nearly two decades ago this is changing. The development of different chip materials and MS technologies has made this coupling easier. The successful online coupling of microfluidics to MS has enabled the analysis of valuable and complex biological samples. Moreover, MS offers an endless number of approaches for detection beyond electrochemical and spectroscopic techniques.

Many ionization sources have already been coupled with microfluidics chip, such as ESI,[34] DESI,[35] MALDI[36] and paper spray.[37] All of them have their own advantages and disadvantages, and the choice should be made according to the properties of the sample to be analyzed and the type of result one wants to achieve. ESI is the most common ionization technique considered for coupling microfluidic devices to MS due to its simplicity of interfacing. The progress in technology for microfabrication has also led to the development of multifunctional microfluidic chips for interfacing ESI to MS instrumentation. These advances have enabled the application to high-throughput and automated analysis.[26]

An example of a microfluidic chip performing flow separation by a micro-solid phase extraction (SPE) channel for the separation of complex samples by ESI-MS is shown Figure 10.2a. The chip receptors labeled “auxiliary” in the figure allow for the change of pH and/or the addition of matrices, enabling application to high-throughout and automated analysis.[26,38] Similar to

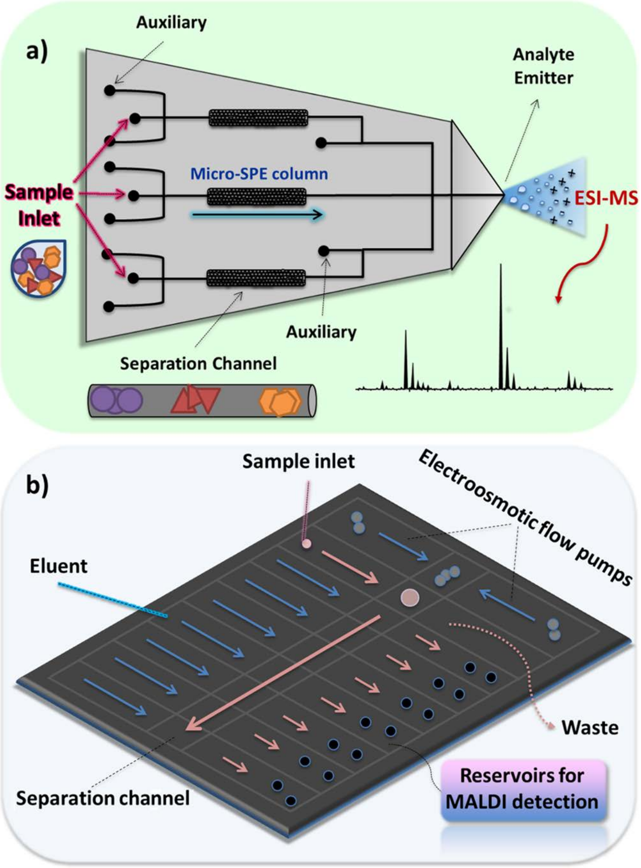

Figure 10.2 Schematic figures of (a) a microfluidic chip for separation of complex samples and automated ESI-MS detection, and (b) a microfluidic liquid chromatography (LC) device for interfacing to MALDI-MS (based on Lazar and Kabulski[36]).

the coupling to ESI, microfluidic devices have been developed for MALDI (see Figure 10.2b). This demonstrates an orthogonal extraction device with reverse phase separation for MALDI imaging. The device exemplifies and demonstrates the principle of using microfluidic chips as a tool to separate and analyze a small amount of sample with a small amount of eluting fluid using a modified surface to allow a multi-step flow system.[36]

A very good example of a multifunction chip for application to ESI-MS analysis is the on-chip digestion approach developed by Wang and co-workers in 2010[39] for cytochrome C analysis. The researchers improved the sensitivity by integrating trypsin digestion, SPE concentration, separation by electrophoresis and MS analysis. The on-chip digestion uses immobilized trypsin, which totally consumes the protein in 3 min, substantially faster than the traditional digestion methods, which typically take 2 h or more.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

Additionally, the integrated system demonstrated higher sequence coverage and potential for automation and high-throughput protein digestion.

The technological advancements of microfluidic analysis have allowed the development of many pioneering works described recently. In 2016, Wang and co-workers[40] observed significant changes in *N*-glycan profiling in leukemia cells treated with acridone, a potential antitumor compound. Due to the low amount of sample used, it was applied to a porous graphitized carbon microfluidic chip ESI-MS platform. This approach provided the separation of the glycans with high sensitivity. The authors also suggested that oligosaccharyltransferase sub-units were a potential biomarker for monitoring toxicity and antitumor activity.

Microfluidic chips have showed a great number of advantages as separation devices with high sensitivity and the possibility of integration to MS and automation. This is especially the case when applied in bioassays. Owing to the microchannel scale of the devices and the increasing advances in microvalve techniques, it is an excellent approach to investigating cellular dynamic events, monitoring key signaling biomolecules. A miniaturized version of capillary electrophoresis, also called microchip electrophoresis, has been shown to be a promising technique for separation for microfluidic analysis. As a result of so many advantages this technique appears ideal for monitoring neurotransmitters in neuronal cells. The dynamic changes in neurochemical release was investigated by Ly and co-workers in 2016,[41] using a chip-based ESI-MS approach. The microchip electrophoresis was manufactured to simultaneously analyze the neurotransmitters: dopamine, serotonine, aspartic acid and glutamic acid. The authors observed that all of the neurotransmitters were stimulated in the presence of KCl or ethanol. The dynamic release observed was distinct, which led the authors to suggest that dopamine and serotonin are packaged into different vesicle pools.

In 2009, Sen *et al.* presented the microfabrication and testing of a microfluidic nebulization chip for DESI-MS for proteomic analysis.[35] The microfluidic chip was fabricated using cyclic olefin copolymer substrates. The nebulizer chip was used to perform DESI-MS analyses of peptides (BSA and bradykinin) and reserpine on the surface of nanoporous alumina. The DESI-MS performance of the microfluidic nebulizer chip had a higher analytical quality (lower limits of detection) when compared to the results obtained using a conventional DESI nebulizer.[35]

In 2013, Lazar and Kabulski published another example of the use of microchips in association with MALDI-MS.[36] They developed a device composed of a matrix of functional elements capable of performing chromatographic separations with the integration of microchip-MS. Essentially, the device provides a MALDI-MS snapshot of the contents of the channel separation present on the chip. They presented the detection of proteins with the potential as biomarkers in MCF10A breast epithelial cells with detection limits in the low fmol range. In addition, the design of the new LC-MALDI-MS chip attracts the promotion of a new concept for performing sample separations within the limited timeframe that accompanies the dead volume of a separation channel.[36]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

In 2014, Zhang and co-workers[37] presented a paper spray method for analyzing microdroplets produced in a gravity driven microchip. Use of paper as a device gives many advantages such as easy adaptation, low-cost and easily disposable sample cartridges. Paper spray ionization MS was then performed to analyze the droplet content. This interface was assembled and controlled manually, presenting a simple and easy method of implementation. Additionally, paper spray ionization MS is promising in the direct analysis of actual samples of micro-biological/chemical reactions due to its tolerance to complex matrices. As a proof of concept, the hydrolysis of acetylcholine drops was performed to demonstrate the validation of the method for the direct analysis of micro-chemical/biological reactions. The work demonstrated that the combination of a microdroplet chip with paper spray ionization MS is a useful platform for monitoring and analyzing such reactions.

The growing interest in microfluidics has led to rapid advances in the associated technologies resulting in analysis techniques with high sensitivity and specificity, with even greater miniaturization. Microfluidic chips coupled with MS have clearly demonstrated a crucial role in studies of cellular behavior, making it possible to perform real-time monitoring of cellular reactions and events. When considered all together, these advances in technology, including the development of more compatible biomaterials, the specific chip approaches, and the coupling to MS to enhance sensitivity, we can conclude that microfluidics coupled with MS has the potential to greatly improve the field of analytical chemical biology.

Therefore, it is clear to see that microfluidics, when coupled with MS, offers a range of innovative technological opportunities to obtain new knowledge and understanding of biological systems. The power of these methods can only be fully exploited with a synergistic effort for accurate chip analysis. In addition to the many advances and advantages, as discussed above, there is still room to improve this technology. The technique is highly dependent on the matrix, and there are few options commercially available. Most of the microfluidic chips cited in this chapter were manufactured according to the sample characteristics. Therefore, the greatest challenge to the progress of this technique relies on the development of either a universal matrix or the supply of many commercially available ones. Unfortunately this could bring with it issues with regards to chip-to-chip reliability and reproducibility, which will have to be addressed by the commercial suppliers to maintain confidence with the users.

10.5 Single-cell Metabolomics

Cells, the smallest unit of all living organisms, were discovered by Robert Hooke in 1665. Since then, researchers have been working on trying to comprehend the complexity of cell systems, correlating their functions with different factors such as size, shape and composition. Despite being observed for the first time in 1869 by Frederich Miescher, the role of deoxyribonucleic acid (DNA) in genetic inheritance was not discovered until much later.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

In 1943, Oswald Avery suggested that DNA is accountable for specific and apparently inheritable transformations in bacteria. It was only in 1953, that James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin proposed the structure of DNA from X-ray crystallography, which in-turn was due to improvements in this technique. This important work led to Watson, Crick and Wilkins being awarded the 1962 Nobel Prize for Medicine for their work on understanding the structure of DNA and its significance in information transfer between living organisms.[42] However, other molecules inside cells, such as proteins, lipids, carbohydrates, and small molecules, all have specific functions that are no less important than the role of DNA, and in many cases, their exact role is still unclear.

The metabolome, as defined and discussed in Chapter 3, is the sum of all the small organic molecules (<1500 Da) produced by a cell, tissue, organ or an organism in a given time and space. It represents the products of all metabolic process, including all enzymatic reactions and the outcome of gene expression.[43] Advances in MS approaches and data analysis have revolutionized the field of metabolomics. State-of-the-art metabolomics studies can provide a functional overview of how environmental factors (such as medication and/or nutrition) could affect an organism.

To evaluate molecules in cells, it is common to study multi-cellular systems such as tissues or cell cultures, which will provide an average value as a response. However, such measurements of an average fail to provide conclusive evidence for molecular behavior inside an individual cell. The study of biological molecules in a single cell is a challenge for chemical biology. A recent study has revealed the heterogeneity of single-cell populations. It showed that cells are even different to each other on the same culture plate. Phenotype development will differ even in cloned cells due to their sensitivity to environmental influences such as culture media, light exposure, extracellular osmotic stress, and other factors.[44] As a result, you could question: “what is the advantage or benefit of single-cell studies, as opposed to studies of the average?”.

Many research groups are using single-cell approaches to help comprehend how cancer cells evolve, progress, metastasize and respond to treatment—for example a single tumor cell can lead to the death of an entire body (about 37 billion cells).[45] Metastasis is the processes of the migration of cells from the initial invading tumor to other distant sites. This process is responsible for about 90% of cancer mortality; however, its mechanism remains unknown.[46] Single-cell analysis, however, enables the identification and classification of different cell populations as normal, tumorous or metastasis cells.[47]

Neurons are another good example. They are a type of cell that connect to each other using synapses, allowing a rapid transmission of chemical or electrical signals. There is an extensive body of literature describing the heterogeneity present in neuronal cells at the genomic level.[48–50] Neurons are very sensitive, and are responsible for many functions; a single chemical or electric signal received or given out by a single neuron could change the entire life of a human being. Thus, increasing our understanding of this type of cell and how they work could help neuroscientists explain neurological issues, such as the small daily problems afflicting many people, or more serious brain diseases. Several neuropsychiatric diseases have been attributed to changes in the genomic content of single cells, such as schizophrenia and Alzheimer's disease.

The ability to study the interactions between pathogens and host cells is another advantage of single-cell analysis. There is evidence that pathogens benefit (for example through increased resistance) from the processes that also lead to the host cell's death,[51] and that the vacuoles may be an important organelle involved in this process.[52] The determination of how the growth of pathogens relate to the viability of the host cells will help in understanding the lifestyle of pathogens and how cellular events lead to host cell death during a disease.[53] The knowledge of the sequence of cytological events leading to cell death will guide the development of more effective drugs and therapies or to control pathogenic attack.

Studies published by Zenobi and co-workers have developed different approaches to make single-cell analysis possible. His group developed a microarray for MS (MAMS), a type of chip used for single-cell studies.[54,55] Cells can be easily distributed in parallel and automated in the MAMS device by pulling a droplet of cell suspension onto the surface of the microarray with a glass slide or a piece of plastic. The design and chemical composition of MAMS allows this deposition with hydrophilic reservoirs with an omniphobic surface and pores of 50–100 μm. Several compositions have been tested in the manufacture of MAMS. Urban *et al.*[54] successfully used a metal support with indium tin oxide glass slides coated with polysilaze. Figure 10.3

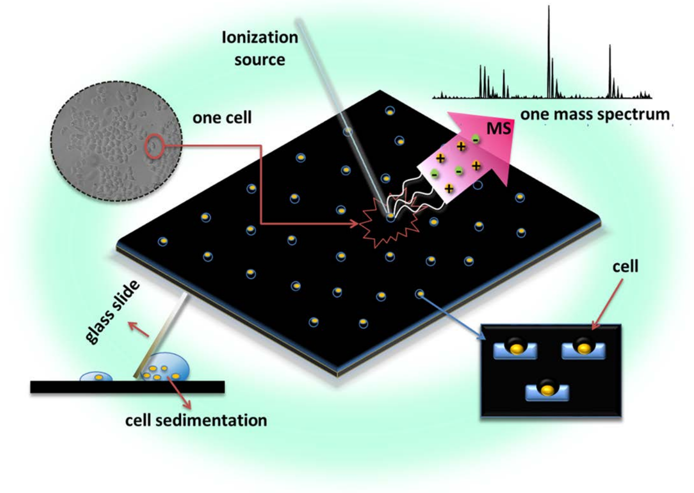

Figure 10.3 Schematic representation of MAMS and a simplified view of cell deposition on a checkerboard pattern.

shows a schematic diagram of a sample deposited in reservoirs after application of a matrix (typically 9-aminoacridine) for MALDI-TOF analysis, resulting in a single spectrum per cell. MAMS was initially developed for MS analysis, but it can also be used in fluorescence analysis or Ramam spectroscopy.

Using their invention for innovative research, Urban *et al.* monitored the metabolism of *Saccharomyces cerevisiae* at the single-cell level, and reported a measure of cellular energy charge (ATP/ADP ratio), along with a correlation between some metabolites from the glycolytic pathway.[30] In other work with single-cell approaches, the group distinguished two *Chlamydomonas reinhardtii* strains with different phenotypes using MS based on the presence or absence of chlorophyll. These reports demonstrate the type of scientifically valuable findings from single-cell metabolomic studies that would not have been possible with population-level analysis.

We could present an infinite list of examples, advantages and questions about single-cell analysis. “What is the metabolic profile of each type of cell?” It is incredible to think of the possibilities that comprehension of all the metabolic processes inside a unique cell might bring, especially the understanding of “why do cells acquire different phenotypes during embryogenesis?”, “how do cells communicate with each other?” or even “how do microbial biosynthetic pathways react to threat?”. Not to mention the possibilities of understanding the whole process of aging; that would be wonderful! Nevertheless, to make all this possible, it is necessary to develop more specific techniques and software that will enable the increasingly detailed study of single cells.

Increasingly, specific techniques have been employed to observe and analyze cellular function. For example, fluorescence microscopy is considered to represent great progress in cellular structure studies. More specifically, confocal fluorescence microscopy has become a key tool that allows the observation of molecular dynamics in living cells.[53] In addition, the numerous applications, biological molecules must be fluorescently stained. As a result, cells need to be treated with detergents in order to allow the dye to permeate the cell membrane, which can lead to the creation of artifacts.[56] Other techniques, such as optical spectroscopy[57] and capillary electrophoresis,[58] have also contributed to advances in single-cell analysis. As discussed previously, there are many well-established techniques for the measurement of DNA, RNA and proteins, however, most of them have failed to detect small molecules. MS is a technique well known for allowing the study of small organic molecules without isolation in complex systems, such as biological matrices.[59] The uses of MS in single-cell metabolomics remains relatively unexplored,[60,61] however, owing to its high sensitivity for the detection of many different metabolites and its flexibility enabling coupling to other techniques, many MS-based approaches have been developed and adapted. Capillary electrophoresis ESI-MS and laser desorption–ionization MS are amongst other MS-based techniques that have demonstrated some success in this field.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

MALDI is the technique most often applied to single-cell metabolomic experiments, mainly owing to recent improvements in the spatial resolution of the laser beam, to less than the size of a single eukaryotic cell. If you consider that single-cell analysis needs highly sensitive methodologies, lower spatial resolution would result in loss of essential information. ESI is an alternative technique that can be used in single-cell analysis. In a study in 2015, Onjiko *et al.*[62] used capillary electrophoresis ESI-MS to demonstrate that the metabolome can affect the cell phenotype in embryonic cells of the South African clawed frog (*Xenopus laevis*). In 2014, Gong and co-workers[63] related the significant metabolite diversity in epidermal cells from an *Allium cepa* bulb using a tungsten probe inserted into live cells to enrich metabolites and then desorbed/ionized for ESI-MS detection. They noted that the outer epidermal cells had more lipids while the inner epidermal cells presented more fructans.

SIMS is an MS technique widely applied for single-cell proteomic analysis, mainly for superficial studies because of its high spatial resolution and sensitivity. Another advantage is that SIMS does not need a matrix, thus avoiding issues with sample preparation and matrix use of MALDI-MS. Advances in ion probe technologies have also led to an expansion in the range of analytes detectable, making it possible to detect and localize metabolites in a two- or three-dimensional cellular sample.[64] In 2016, Huang and co-workers[65] developed facile single-cell patterning integrated with SIMS, which was able to distinguish drug-induced phenotypic alterations at the single-cell level.

MS of a single-cell would be extremely difficult to interpret, due to the many contaminants and high noise level as well as the presence of isomers. In addition, the differences between the single-cells would lead to spectra with considerable variations in peaks. In an attempt to alleviate these problems, Krismer *et al.*[32] proposed that tandem MS (MS/MS) could be used to directly identify the biological molecules in a single-cell. The authors presented the first MS/MS approach applied to the analysis of single cells of *Chlamydomonas reinhardtii*. This study demonstrated the successful assignment of the majority of peaks detected in the spectra.

Despite all the advantages of single-cell metabolomics, this technique still has some challenges to overcome. Some of them have already been discussed, along with some ideas for their potential solution. Another challenge that needs to be kept in mind is the isolation of the cells without damaging them, and in a way that allows for the metabolomic analysis to be performed. One way to isolate cells is by manually using a microscope; however, it is difficult and time consuming. There are some high throughput methodologies that allow cell isolation, as discussed earlier. MAMS (the MALDI-based chip technology) was developed specifically for this purpose, and has been successfully applied.[30–32] However, the most commonly used technique is mass cytometry. Flow cytometry is a technique used to analyze the properties of single cells in a fluid using antibodies to label cellular components of interest. Mass cytometry is an adaptation of flow cytometry coupled to MS, to allow for direct single-cell analysis.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

The development of new platforms and methodologies for the analysis of single cells increases the prospects for understanding the interactions of biomolecular structures and how such information is exploited in the prevention and treatment of diseases. The extraction of intracellular molecules is an extremely important experimental step, since the reliability of the results also depends on the efficiency of the extraction process as well as the analytical methodology. The strategy used for the extraction of biometabolites depends on the cell type being analyzed, and the approaches can be divided into invasive and non-invasive.[66]

Non-invasive methods are able to maintain the cellular structure intact. Raman,[67] infrared and nuclear magnetic resonance[68] are some examples of techniques that could be applied in the monitoring of biomolecules inside the cell without destroying it. On the other hand, invasive methods are those that destroy the cellular structure for extraction of the metabolites. Among these methods, we can mention that lysis of the cellular membrane with the use of chemical agents, the uses of a needle for withdrawing the microfluidic, or the insertion of microchips into the cell for biomolecules adsorption.

MS approaches applied to single-cell metabolomics have attracted even greater interest in recent years. It is a field that has the potential to remodel analytical chemical biology and contribute to the development of pharmaceutical and medicinal companies. Although there are still many challenges to be overcome, there is continuous research innovation in the area and many proposals for its advancement. Single-cell analysis will allow the gaining of knowledge of the true chemical signature of a single cell, and of the heterogeneity of the cellular populations. In the near future, MS, when applied to single-cell metabolomics, will almost certainly allow for correlations between genomics and proteomics to be made with metabolomics. It is a highly promising research area, which is destined to enrich, direct and inform the field of chemical biology.

10.6 Mass Spectrometry in Surgery

The diagnosis of a large number of diseases is mainly due to their manifestation in tissues, and by detection of morphological changes in different tissue types. In tumor-removal surgery, the surgeon needs to know the surgical dissection margins. Being able to spare the normal (non-tumorous) marginal tissue may have significant benefits to the patient's life, particularly in the case of brain surgery where any functional damage could lead to epilepsy or memory loss, amongst other potentially life changing problems. Traditional histopathological techniques could be applied during the surgical procedure; however, due to a wide error range, the surgeons usually resort to postoperative histopathological examination, and so get the results only after the surgery is completed. This setting highlights the urgency of developing new technologies and methodologies for the identification of diseased tissues more quickly and efficiently with the aim of improving the treatment of cancer and other diseases.[69]

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

The application of MS-based methods for tissue identification in surgery is surely amongst the greatest achievements for analytical sciences, and chemical biological in particular. The development of MS approaches has allowed the direct analysis of living human tissue. The investigation of the molecular profile and morphology of tissues is of great importance to generating knowledge about the fundamental nature of a biological system. Furthermore, the in-depth molecular profile that comprises biological tissues could lead to the identification of specific targets for normal and unhealthy tissues. Cells from different tissues have specific chemical profiles, as discussed earlier in this chapter. Although this approach is focused on improvements in the medical field, it has also stimulated further research on the comprehension of the distribution of metabolites and their function in human tissues, bringing with it important information for the chemical biology field as a whole.[70]

MS has been demonstrated to have high specificity and sensitivity and is a powerful tool for the characterization of the biomolecular profile of biological tissues. Although this major breakthrough has been recently presented, there is still great deal of work needed to overcome difficulties in implementation of this technique as a standard tool in clinics. MS was used in operating rooms in the 1980s, but with another objective entirely—namely to evaluate faulty techniques and equipment, and for monitoring the gases emitted by anesthetized patients.[71] The application of MS in surgery, as an analytical tool to guide treatments and therapies, has only become possible due to the advancement on MS itself. This is mainly due to the development of compact ambient ionization sources, such as DESI, which allowed for the direct analysis of samples outside of the vacuum system. The ionization mechanism of DESI is similar to ESI, however, the charged solvent droplets are sprayed directly onto the surface of the tissue sample to be analyzed. The impact of the focused solvent desorbs ions from the surface into the MS instrument for analysis in the usual way.[72]

Since it was first described by Cook's group in 2004,[72] DESI has rapidly become a powerful technique widely applied to evaluate biological samples. DESI-MS offers some advantages over traditional techniques, such as the ability to analyze a wide range of molecules under ambient conditions with minimal sample treatment or preparation. Additionally, DESI is a relatively soft technique that allows for the analyte ions to be measured intact without thermal degradation or fragmentation, which in turn greatly aids identification and characterization. This ease of molecular assignment has made DESI a suitable and powerful tool for rapid *in situ* analysis.[73–75] Moreover, imaging by DESI-MS has become an interesting and promising approach in histopathology evaluation to support surgical decision making, owing to its ability to provide the spatial distribution of molecules in a biological tissue surface, and may be achieved at a pixel size of 20–100 μm.[69]

A good example of imaging by DESI-MS is its use in brain cancer surgery, published recently by the Cook group.[73] In order to apply this tool intraoperatively for tumor margin evaluation, the authors demonstrated the molecular analysis of a human brain tumor during surgery. The analyses were performed under ambient conditions in the negative ionization mode, monitoring the profile of lipids and other small molecules of glioma samples. The approach used yielded diagnostic results in just a few minutes. This paper was the first published study to use MS within an operating room for the profiling of the metabolome of tumorous tissue, to inform and to improve cancer therapies.

MALDI is another technique widely applied to imaging biological tissue surfaces. For many years this technique was employed for the identification and characterization of proteins and peptides, but more recently it has seen increasingly application in the analysis of small molecules (as discussed earlier in this chapter). Due to the wide mass range that MALDI-MSI can analyze, it is able to provide a more thorough investigation of biological tissue.[76] The disadvantage of MALDI-MSI is that this technique requires the matrix to be deposited over the tissue sample and the time for crystal formation to occur. Recently, the development of glass slides pre-coated with matrix has made MALDI-MSI more applicable for performing rapid clinical analysis.[77] Despite this, MALDI is not compatible with the operating room workflow due to its requirements for sample preparation. However, the technique is still considered a promising approach to supporting decision making by surgeons in the future.

For the effective and realistic use of MALDI-MSI in real time during surgery, a number of drawbacks need to be overcome. Several published studies demonstrate the technique's potential and efficiency in profiling the metabolome of biological tissues. In 2014, Mirnezami and co-workers[78] applied MALDI-MSI to compare the chemical profile of fresh frozen sections of colorectal cancer (CRC) against healthy tissue. They reported a characteristic phospholipidic signature for the CRC tissue as well as observing biochemical differences between the CRC microenvironments.

Several MS methods have been demonstrated to support surgical therapies. Another approach is to apply mechanical disintegration of the tissue followed by ultrasonic aspiration to disintegrate the cellular components. This technique avoids excessive tissue damage, which is an extremely important factor, especially in the case of brain surgery. In contrast to the discussed off-line approaches, Schäfer *et al.*[79] developed an online coupling technique by introducing the effluent of a cavitron ultrasonic surgical aspiration device directly into the Venturi easy ambient sonic-spray ionization source. This allowed for *in situ* analysis in quasi real-time. The authors reported the detection of predominantly complex lipid constituents and also multiply charged peptides in the following tissues: meningeomas, astrocytomas, and metastatic and healthy brain.

The increasing development of MS technologies, in particular regarding ambient ionization techniques, has revolutionized the science, allowing for chemical and molecular information to be extracted from living human tissue samples. These comprehensive approaches have broadly aroused the interest of many researchers around the world in attempts to define disease biomarkers directly by MSI application. Obviously, these techniques are still

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

in their infancy and it is not possible to make major demands from them yet. However, what is presented here is a discussion of the hugely significant advances and achievements, and obvious potential of the application of real-time MS-based methods in operating rooms. These achievements present an opportunity for learning the mechanism of development of numerous diseases. Now chemists and biologists need to be creative in using this new knowledge and to applying these technological advances to enable new discoveries that will result in a better understanding of the metabolome and thus the function of molecules in living biological systems.

10.7 Synthetic Ecology

This term was introduced in 2007 by Dunham,[80] and first mentioning in a study by Shou *et al.*[81] Synthetic ecology uses the concepts developed in synthetic biology, but aims them at the study of ecology. This new area of study is not only to evaluate the building blocks of a population structure, but to mix different populations of cells in order to construct a simple ecological system of obligatory cooperation, and to use it as a robust model for establishing limits and parameters for such systems.

The study of ecology is a way of understanding how microbiota can influence human life. The emergence of synthetic ecology has provided an attractive alternative for the study of species against several parameters of the system in which they are included, enabling the discovery of the main ecological characteristics of microbial communities, for the design of synthetic consortia, and for biotechnology and biomedical applications. There are several examples of promising small synthetic ecosystems that demonstrate the feasibility of this approach, such as those described by Shou *et al.*[81] and Grosskopf and Soyer.[82] The most exciting opportunities in synthetic ecology are to enable a more systematic approach to the study of the complexity of microbial systems with a hierarchy of nested networks, metabolic and external factors.[82]

A good example of this approach is the use of DESI-MS for the analysis of the surface of the red macroalgae *Callophycus serratus*. This analysis sought to understand the signaling sites of the organisms to attract allies and defend against enemies. In 2009, Lane *et al.*[83] evaluated the algae surface seeking an understanding of the relationship to microbes. The reported defenses revealed that bromophycolides (diterpene-benzoate macrolides) were found exclusively in association with distinct surface patches at adequate concentrations for fungal inhibition (a marine fungal pathogen *Lidra thalassiae*). DESI-MS also indicated the presence of bromophycolides within the internal alimentary tissue. This is one of the earliest examples of product imaging on biological surfaces, suggesting the importance of secondary metabolites in localized ecological interactions and illustrating the potential of DESI-MS in understanding chemically mediated biological processes.

In 2012, Watrous and co-workers[84] demonstrated a methodology for profiling live microbial colonies by nano-DESI-MS without any further sample preparation. This technique provides a highly sensitive tool for the study of bacterial interactions directly from living microbial communities. This work revolutionized synthetic ecology and allows for the analysis of a wide variety of metabolites from living systems along with correlation with the phenotype.

In 2014, Shou *et al.*[81] suggested that their methodology could be useful for studying "cheaters", *i.e.* cells that eat nutrients without contributing themselves. Cheaters pose a special challenge to the stability of cooperative systems and have arisen with interesting consequences in experimental evolutions of *Pseudomonas fluorescens*. In 2015, Song and co-workers[11] demonstrated that the composition of a microbial community can be significantly affected by protozoan predation. The authors observed communication between *Pseudomonas fluorescens* and the protist *Naegleria americana* at the molecular and chemical level by combining genome transcriptome analyses, MSI and live colony nano-DESI-MS. This study provides new insights at the interface of *Pseudomonas*–protozoa interactions relating the metabolic response to bacterial survival in competitive environments.

10.8 Final Considerations

Chemists and biologists must think more about the whole systems, not just single biomolecules. Overall, scientists are used to evaluating the function of the components of cells, such as DNA and proteins, as isolated molecules, in an environment completely different from how they occur natively. Biomolecules function with infinite interactions inside a biological system. We urgently need the development of new approaches to study living systems in real-time. The authors feel, and have tried to demonstrate in this chapter, that MS has already played, is playing and can play an important role in this area.

The field of chemical biology uses the power of chemistry to achieve a deep knowledge and understanding of biological functions, applying it to pharmacology, biotechnology and medical sciences with the aim of improving the quality of life and generally advancing scientific knowledge. In the future, it is possible that single-cell analysis will play a significant role in increasing the understanding of all human cells and tissues, and even their behavior in an ever deeper and more detailed way. The application of MS in surgery will bring essential molecular information to inform surgeons during real-time interventions and even help diagnose disease or correct a misdiagnosis. The application and development of new MS-based technologies, methodologies and software, along with parallel advances in computing and bioinformatics, will most likely make possible time and space resolved real-time analysis in the future. This, in turn, will lead to a revolution in the field of chemical biology, from surgical to ecological applications.

Acknowledgements

The authors are grateful to Universty of São Paulo, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for financial support.

References

1. A. Fulop, D. A. Sammour, K. Erich, J. von Gerichten, P. van Hoogevest, R. Sandhoff and C. Hopf, Molecular imaging of brain localization of liposomes in mice using MALDI mass spectrometry, *Sci. Rep.*, 2016, **6**, 33791.
2. S. Giordano, M. Zucchetti, A. Decio, M. Cesca, I. F. Nerini, M. Maiezza, M. Ferrari, S. A. Licandro, R. Frapolli, R. Giavazzi, D. Maurizio, E. Davoli and L. Morosi, Heterogeneity of paclitaxel distribution in different tumor models assessed by MALDI mass spectrometry imaging, *Sci. Rep.*, 2016, **6**, 39284.
3. N. M. Stasulli and E. A. Shank, Profiling the metabolic signals involved in chemical communication between microbes using imaging mass spectrometry, *FEMS Microbiol. Rev.*, 2016, **40**(6), 807–813.
4. R. Castaing and G. Slodzian, Optique corpusculaire - Premiers essais de microanalyse par emission ionique secondaire, *C. R. Hebd. Seances Acad. Sci.*, 1962, **1**, 395–410.
5. E. Handberg, K. Chingin, N. N. Wang, X. M. Dai and H. W. Chen, Mass spectrometry imaging for visualizing organic analytes in food, *Mass Spectrom. Rev.*, 2015, **34**(6), 641–658.
6. A. Bodzon-Kulakowska and P. Suder, Imaging mass spectrometry: Instrumentation, applications, and combination with other visualization techniques, *Mass Spectrom. Rev.*, 2016, **35**(1), 147–169.
7. R. Longuespee, R. Casadonte, M. Kriegsmann, C. Pottier, G. P. de Muller, P. Delvenne, J. Kriegsmann and E. De Pauw, MALDI mass spectrometry imaging: A cutting-edge tool for fundamental and clinical histopathology, *Proteomics: Clin. Appl.*, 2016, **10**(7), 701–719.
8. R. M. Caprioli, T. B. Farmer and J. Gile, Molecular imaging of biological samples: Localization of peptides and proteins using MALDI-TOF MS, *Anal. Chem.*, 1997, **69**(23), 4751–4760.
9. P. Le Pogam, B. Legouin, A. Geairon, H. Rogniaux, F. Lohezic-Le Devehat, W. Obermayer, J. Boustie and A. C. Le Lamer, Spatial mapping of lichen specialized metabolites using LDI-MSI: chemical ecology issues for Ophioparma ventosa, *Sci. Rep.*, 2016, **6**, 37807.
10. T. B. Angerer, Y. Magnusson, G. Landberg and J. S. Fletcher, Lipid heterogeneity resulting from fatty acid processing in the human breast cancer microenvironment identified by GCIB-ToF-SIMS imaging, *Anal. Chem.*, 2016, **88**(23), 11946–11954.
11. C. Song, M. Mazzola, X. Cheng, J. Oetjen, T. Alexandrov, P. Dorrestein, J. Watrous, M. van der Voort and J. M. Raaijmakers, Molecular and chemical dialogues in bacteria-protozoa interactions, *Sci. Rep.*, 2015, **5**, 12837.
12. D. B. Silva, I. C. Turatti, D. R. Gouveia, M. Ernst, S. P. Teixeira and N. P. Lopes, Mass spectrometry of flavonoid vicenin-2, based sunlight barriers in Lychnophora species, *Sci. Rep.*, 2014, **4**, 4309.
13. T. M. Nunes, S. Mateus, A. P. Favaris, M. F. Amaral, L. G. von Zuben, G. C. Clososki, J. M. Bento, B. P. Oldroyd, R. Silva, R. Zucchi, D. B. Silva and N. P. Lopes, Queen signals in a stingless bee: suppression of worker ovary activation and spatial distribution of active compounds, *Sci. Rep.*, 2014, **4**, 7449.

14. N. E. Bradbury, The mobility of aged ions in air in relation to the nature of gaseous ions, *Phys. Rev.*, 1931, **37**(10), 1311–1319.
15. E. A. Mason and E. W. McDaniel, Transport properties of ions in gases, *NASA STI/Recon Technical Report A*, 1988, p. 89.
16. Y. R. Wen, F. Sobott and B. Devreese, ATP and autophosphorylation driven conformational changes of HipA kinase revealed by ion mobility and crosslinking mass spectrometry, *Anal. Bioanal. Chem.*, 2016, **408**(21), 5925–5933.
17. E. V. Krylov and E. G. Nazarov, Electric field dependence of the ion mobility, *Int. J. Mass Spectrom.*, 2009, **285**(3), 149–156.
18. G. R. Asbury and H. H. Hill, Using different drift gases to change separation factors (alpha) in ion mobility spectrometry, *Anal. Chem.*, 2000, **72**(3), 580–584.
19. L. M. Matz, H. H. Hill, L. W. Beegle and I. Kanik, Investigation of drift gas selectivity in high resolution ion mobility spectrometry with mass spectrometry detection, *J. Am. Soc. Mass Spectrom.*, 2002, **13**(4), 300–307.
20. M. D. Howdle, C. Eckers, A. M. F. Laures and C. S. Creaser, The effect of drift gas on the separation of active pharmaceutical ingredients and impurities by ion mobility-mass spectrometry, *Int. J. Mass Spectrom.*, 2010, **298**(1–3), 72–77.
21. J. M. Dilger, S. J. Valentine, M. S. Glover, M. A. Ewing and D. E. Clemmer, A database of alkali metal-containing peptide cross sections: Influence of metals on size parameters for specific amino acids, *Int. J. Mass Spectrom.*, 2012, **330**, 35–45.
22. Z. Hall, A. Politis and C. V. Robinson, Structural Modeling of Heteromeric Protein Complexes from Disassembly Pathways and Ion Mobility-Mass Spectrometry, *Structure*, 2012, **20**(9), 1596–1609.
23. A. Politis, A. Y. Park, Z. Hall, B. T. Ruotolo and C. V. Robinson, Integrative modelling coupled with ion mobility mass spectrometry reveals structural features of the clamp loader in complex with single-stranded DNA binding protein, *J. Mol. Biol.*, 2013, **425**(23), 4790–4801.
24. M. Jovanovic and J. Peter-Katalinic, Preliminary mass spectrometry characterization studies of galectin-3 samples, prior to carbohydrate-binding studies using affinity mass spectrometry, *Rapid Commun. Mass Spectrom.*, 2017, **31**(1), 129–136.
25. R. Z. Ning, F. Wang and L. Lin, Biomaterial-based microfluidics for cell culture and analysis, *TrAC, Trends Anal. Chem.*, 2016, **80**, 255–265.
26. X. J. Feng, B. F. Liu, J. J. Li and X. Liu, Advances in coupling microfluidic chips to mass spectrometry, *Mass Spectrom. Rev.*, 2015, **34**(5), 535–557.
27. S. C. Terry, J. H. Jerman and J. B. Angell, Gas-chromatographic air analyzer fabricated on a silicon-wafer, *IEEE Trans. Electron Devices*, 1979, **26**(12), 1880–1886.
28. V. Biju, Chemical modifications and bioconjugate reactions of nanomaterials for sensing, imaging, drug delivery and therapy, *Chem. Soc. Rev.*, 2014, **43**(3), 744–764.

29. K. Hattori, H. Wada, Y. Makanae, S. Fujita and S. Konishi, RNA extraction from microtissues of skeletal muscle by a microfluidic shredding chip, *IEEJ Trans. Electr. Electron. Eng.*, 2016, **11**, S123–S129.
30. A. J. Ibanez, S. R. Fagerer, A. M. Schmidt, P. L. Urban, K. Jefimovs, P. Geiger, R. Dechant, M. Heinemann and R. Zenobi, Mass spectrometry-based metabolomics of single yeast cells, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**(22), 8790–8794.
31. J. Krismer, J. Sobek, R. F. Steinhoff, S. R. Fagerer, M. Pabst and R. Zenobi, Screening of *Chlamydomonas reinhardtii* populations with aingle-cell resolution by using a high-throughput microscale sample preparation for Matrix-Assisted Laser Desorption Ionization Mass Spectrometry, *Appl. Environ. Microbiol.*, 2015, **81**(16), 5546–5551.
32. J. Krismer, R. F. Steinhoff and R. Zenobi, Single-cell MALDI tandem mass spectrometry: unambiguous assignment of small biomolecules from single *Chlamydomonas reinhardtii* cells, *Chimia*, 2016, **70**(4), 236–239.
33. D. Gao, H. X. Liu, Y. Y. Jiang and J. M. Lin, Recent advances in microfluidics combined with mass spectrometry: technologies and applications, *Lab Chip*, 2013, **13**(17), 3309–3322.
34. J. S. Mellors, W. A. Black, A. G. Chambers, J. A. Starkey, N. A. Lacher and J. M. Ramsey, Hybrid capillary/microfluidic system for comprehensive online liquid chromatography-capillary electrophoresis-electrospray ionization-mass spectrometry, *Anal. Chem.*, 2013, **85**(8), 4100–4106.
35. A. K. Sen, J. Darabi and D. R. Knapp, Design, fabrication and test of a microfluidic nebulizer chip for desorption electrospray ionization mass spectrometry, *Sens. Actuators, B*, 2009, **137**(2), 789–796.
36. I. M. Lazar and J. L. Kabulski, Microfluidic LC device with orthogonal sample extraction for on-chip MALDI-MS detection, *Lab Chip*, 2013, **13**(11), 2055–2065.
37. Y. Zhang, H. Li, Y. Ma and J. M. Lin, Paper spray mass spectrometry-based method for analysis of droplets in a gravity-driven microfluidic chip, *Analyst*, 2014, **139**(5), 1023–1029.
38. X. Wang, L. Yi, N. Mukhitov, A. M. Schrell, R. Dhumpa and M. G. Roper, Microfluidics-to-mass spectrometry: A review of coupling methods and applications,*J. Chromatogr. A*, 2015, **1382**, 98–116.
39. C. Wang, A. B. Jemere and D. J. Harrison, Multifunctional protein processing chip with integrated digestion, solid-phase extraction, separation and electrospray, *Electrophoresis*, 2010, **31**(22), 3703–3710.
40. Y. N. Wang, D. Park, A. G. Galermo, D. Gao, H. X. Liu and C. B. Lebrilla, Changes in cellular glycosylation of leukemia cells upon treatment with acridone derivatives yield insight into drug action, *Proteomics*, 2016, **16**(23), 2977–2988.
41. X. Li, H. Hu, S. Zhao and Y. M. Liu, Microfluidic platform with in-chip electrophoresis coupled to mass spectrometry for monitoring neurochemical release from nerve cells, *Anal. Chem.*, 2016, **88**(10), 5338–5344.
42. H. M. Berman, Crystal studies of B-DNA: the answers and the questions, *Biopolymers*, 1997, **44**(1), 23–44.

43. G. A. Prosser, G. Larrouy-Maumus and L. P. de Carvalho, Metabolomic strategies for the identification of new enzyme functions and metabolic pathways, *EMBO Rep.*, 2014, **15**(6), 657–669.
44. A. H. Wong, I. I. Gottesman and A. Petronis, Phenotypic differences in genetically identical organisms: The epigenetic perspective, *Hum. Mol. Genet.*, 2005, **14**(1), R11–R18.
45. E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli and S. Canaider, An estimation of the number of cells in the human body, *Ann. Hum. Biol.*, 2013, **40**(6), 463–471.
46. P. Mehlen and A. Puisieux, Metastasis: a question of life or death, *Nat. Rev. Cancer*, 2006, **6**(6), 449–458.
47. E. Kidess and S. S. Jeffrey, Circulating tumor cells *versus* tumor-derived cell-free DNA: rivals or partners in cancer care in the era of single-cell analysis? *Genome Med.*, 2013, **5**(8), 70.
48. J. F. Poulin, B. Tasic, J. Hjerling-Leffler, J. M. Trimarchi and R. Awatramani, Disentangling neural cell diversity using single-cell transcriptomics, *Nat. Neurosci.*, 2016, **19**(9), 1131–1141.
49. I. Y. Iourov, S. G. Vorsanova and Y. B. Yurov, Single cell genomics of the brain: focus on neuronal diversity and neuropsychiatric diseases, *Curr. Genomics*, 2012, **13**(6), 477–488.
50. H. Wichterle, D. Gifford and E. Mazzoni, Neuroscience. Mapping neuronal diversity one cell at a time, *Science*, 2013, **341**(6147), 726–727.
51. M. B. Dickman and R. Fluhr, Centrality of host cell death in plant-microbe interactions, *Annu. Rev. Phytopathol.*, 2013, **51**, 543–570.
52. S. Mochizuki, E. Minami and Y. Nishizawa, Live-cell imaging of rice cytological changes reveals the importance of host vacuole maintenance for biotrophic invasion by blast fungus, Magnaporthe oryzae, *MicrobiologyOpen*, 2015, **4**(6), 952–966.
53. K. Jones, D. W. Kim, J. S. Park and C. H. Khang, Live-cell fluorescence imaging to investigate the dynamics of plant cell death during infection by the rice blast fungus Magnaporthe oryzae, *BMC Plant Biol.*, 2016, **16**, 69.
54. P. L. Urban, K. Jefimovs, A. Amantonico, S. R. Fagerer, T. Schmid, S. Madler, J. Puigmarti-Luis, N. Goedecke and R. Zenobi, High-density micro-arrays for mass spectrometry, *Lab Chip*, 2010, **10**(23), 3206–3209.
55. P. L. Urban, A. M. Schmidt, S. R. Fagerer, A. Amantonico, A. Ibanez, K. Jefimovs, M. Heinemann and R. Zenobi, Carbon-13 labelling strategy for studying the ATP metabolism in individual yeast cells by micro-arrays for mass spectrometry, *Mol. BioSyst.*, 2011, **7**(10), 2837–2840.
56. S. M. Wilson and A. Bacic, Preparation of plant cells for transmission electron microscopy to optimize immunogold labeling of carbohydrate and protein epitopes, *Nat. Protoc.*, 2012, **7**(9), 1716–1727.
57. W. E. Huang, M. Li, R. M. Jarvis, R. Goodacre and S. A. Banwart, Shining light on the microbial world the application of Raman microspectroscopy, *Adv. Appl. Microbiol.*, 2010, **70**, 153–186.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00264

58. P. Nemes, S. S. Rubakhin, J. T. Aerts and J. V. Sweedler, Qualitative and quantitative metabolomic investigation of single neurons by capillary electrophoresis electrospray ionization mass spectrometry, *Nat. Protoc.*, 2013, **8**(4), 783–799.
59. M. Ernst, D. B. Silva, R. Silva, M. Monge, J. Semir, R. Z. Vencio and N. P. Lopes, A metabolomic protocol for plant systematics by matrix-assisted laser-desorption/ionization time-of flight mass spectrometry, *Anal. Chim. Acta*, 2015, **859**, 46–58.
60. R. Zenobi, Single-cell metabolomics: analytical and biological perspectives, *Science*, 2013, **342**(6163), 1243259.
61. T. J. Comi, T. D. Do, S. S. Rubakhin and J. V. Sweedler, Categorizing cells on the basis of their chemical profiles: Progress in single-cell mass spectrometry, *J. Am. Chem. Soc.*, 2017, **139**(11), 3920–3929.
62. R. M. Onjiko, S. A. Moody and P. Nemes, Single-cell mass spectrometry reveals small molecules that affect cell fates in the 16-cell embryo, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**(21), 6545–6550.
63. X. Gong, Y. Zhao, S. Cai, S. Fu, C. Yang, S. Zhang and X. Zhang, Single cell analysis with probe ESI-mass spectrometry: detection of metabolites at cellular and subcellular levels, *Anal. Chem.*, 2014, **86**(8), 3809–3816.
64. E. J. Lanni, S. S. Rubakhin and J. V. Sweedler, Mass spectrometry imaging and profiling of single cells, *J. Proteomics*, 2012, **75**(16), 5036–5051.
65. L. Huang, Y. Chen, L. T. Weng, M. Leung, X. X. Xing, Z. Y. Fan and H. K. Wu, Fast single-cell patterning for study of drug-induced phenotypic alterations of HeLa cells using time-of-flight secondary ion mass spectrometry, *Anal. Chem.*, 2016, **88**(24), 12196–12203.
66. S. S. Rubakhin, E. V. Romanova, P. Nemes and J. V. Sweedler, Profiling metabolites and peptides in single cells, *Nat. Methods*, 2011, **8**(suppl. 4), S20–S29.
67. C. W. Freudiger, W. Min, B. G. Saar, S. Lu, G. R. Holtom, C. He, J. C. Tsai, J. X. Kang and X. S. Xie, Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy, *Science*, 2008, **322**(5909), 1857–1861.
68. A. Motta, D. Paris and D. Melck, Monitoring real-time metabolism of living cells by fast two-dimensional NMR spectroscopy, *Anal. Chem.*, 2010, **82**(6), 2405–2411.
69. E. R. St John, M. Rossi, P. Pruski, A. Darzi and Z. Takats, Intraoperative tissue identification by mass spectrometric technologies, *TrAC, Trends Anal. Chem.*, 2016, **85**, 2–9.
70. L. S. Eberlin, M. Gabay, A. C. Fan, A. M. Gouw, R. J. Tibshirani, D. W. Felsher and R. N. Zare, Alteration of the lipid profile in lymphomas induced by MYC overexpression, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**(29), 10450–10455.
71. G. M. Ozanne, W. G. Young, W. J. Mazzei and J. W. Severinghaus, Multipatient anesthetic mass-spectrometry - Rapid analysis of data stored in long catheters, *Anesthesiology*, 1981, **55**(1), 62–70.
72. Z. Takats, J. M. Wiseman, B. Gologan and R. G. Cooks, Mass spectrometry sampling under ambient conditions with desorption electrospray ionization, *Science*, 2004, **306**(5695), 471–473.

73. S. Santagata, L. S. Eberlin, I. Norton, D. Calligaris, D. R. Feldman, J. L. Ide, X. Liu, J. S. Wiley, M. L. Vestal, S. H. Ramkissoon, D. A. Orringer, K. K. Gill, I. F. Dunn, D. Dias-Santagata, K. L. Ligon, F. A. Jolesz, A. J. Golby, R. G. Cooks and N. Y. Agar, Intraoperative mass spectrometry mapping of an onco-metabolite to guide brain tumor surgery, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**(30), 11121–11126.
74. M. Shariatgorji, N. Strittmatter, A. Nilsson, P. Kallback, A. Alvarsson, X. Q. Zhang, T. Vallianatou, P. Svenningsson, R. J. A. Goodwin and P. E. Andren, Simultaneous imaging of multiple neurotransmitters and neuro-active substances in the brain by desorption electrospray ionization mass spectrometry, *NeuroImage*, 2016, **136**, 129–138.
75. D. Miura, Y. Fujimura and H. Wariishi, *In situ* metabolomic mass spectrometry imaging: Recent advances and difficulties, *J. Proteomics*, 2012, **75**(16), 5052–5060.
76. M. Stoeckli, P. Chaurand, D. E. Hallahan and R. M. Caprioli, Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues, *Nat. Med.*, 2001, **7**(4), 493–496.
77. J. Yang and R. M. Caprioli, Matrix precoated targets for direct lipid analysis and imaging of tissue, *Anal. Chem.*, 2013, **85**(5), 2907–2912.
78. R. Mirnezami, K. Spagou, P. A. Vorkas, M. R. Lewis, J. Kinross, E. Want, H. Shion, R. D. Goldin, A. Darzi, Z. Takats, E. Holmes, O. Cloarec and J. K. Nicholson, Chemical mapping of the colorectal cancer microenvironment *via* MALDI imaging mass spectrometry (MALDI-MSI) reveals novel cancer-associated field effects, *Mol. Oncol.*, 2014, **8**(1), 39–49.
79. K. C. Schäfer, J. Balog, T. Szaniszlo, D. Szalay, G. Mezey, J. Denes, L. Bognar, M. Oertel and Z. Takats, Real time analysis of brain tissue by direct combination of ultrasonic surgical aspiration and sonic spray mass spectrometry, *Anal. Chem.*, 2011, **83**(20), 7729–7735.
80. M. J. Dunham, Synthetic ecology: a model system for cooperation, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(6), 1741–1742.
81. W. Shou, S. Ram and J. M. Vilar, Synthetic cooperation in engineered yeast populations, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**(6), 1877–1882.
82. T. Grosskopf and O. S. Soyer, Synthetic microbial communities, *Curr. Opin. Microbiol.*, 2014, **18**, 72–77.
83. A. L. Lane, L. Nyadong, A. S. Galhena, T. L. Shearer, E. P. Stout, R. M. Parry, M. Kwasnik, M. D. Wang, M. E. Hay, F. M. Fernandez and J. Kubanek, Desorption electrospray ionization mass spectrometry reveals surface-mediated antifungal chemical defense of a tropical seaweed, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**(18), 7314–7319.
84. J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, Mass spectral molecular networking of living microbial colonies, *Proc. Natl. Acad. Sci.*, 2012, **109**(26), E1743–E1752.

Published on 09 November 2017 on http://pubs.rsc.org | doi:10.1039/9781788010399-00288

Subject Index