Principles and Practice

M.R. Wilkins R.D. Appel
K.L. Williams D.F. Hochstrasser (Eds.)

Proteome Research
Concepts, Technology and Application

Second Edition

With 46 Figures, 31 in Color and 16 Tables

Springer

Professor Marc R. Wilkins, PhD
School of Biotechnology and
Biomolecular Sciences
University of New South Wales
Sydney NSW 2052
Australia

Professor Keith L. Williams, PhD
BKG Group
P.O. Box 580
Avalon NSW 2107
Australia

Professor Ron D. Appel, Ph.D.
Proteome Informatics Group
Swiss Institute of Bioinformatics
Computer Science Department
University of Geneva
CMU, rue Michel-Servet 1
CH-1211 Geneva 4
Switzerland

Professor Denis F. Hochstrasser, M.D.
Department of Structural Biology
and Bioinformatics
Department of Genetic & Laboratory
Medicine
Laboratory Medicine Service
Geneva University and University Hospital
24, rue Micheli-du-Crest
CH-1211 Geneva 14
Switzerland

Library of Congress Control Number: 2007927822

ISBN-13: 978-3-540-71240-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springer.com

Editor: Dr. Sabine Schreck, Heidelberg
Desk Editor: Anette Lindqvist, Heidelberg
Cover Design: WMXDesign GmbH, Heidelberg
Production and typesetting: SPi

Printed on acid-free paper 39/3180 5 4 3 2 1 0

To Sobet, Catherine, Brynnie and Anne-Catherine
To Adrien, Naara, Jarrah, Asheley, Lucile, Virginie, Sandrine and Nelson

Preface to Second Edition

This book is the result of a long-standing friendship between two research groups – one in Sydney, Australia, and the other in Geneva, Switzerland. It was stimulated by a previous book on proteomics which we produced together in 1997. Many of the authors who contributed to the original book have also written for this new book, but with an additional 10 years of experience. What is interesting about the authors of this book is that many of them have developed new proteomic technology and techniques, commercialized this technology via different routes, established proteomics and/or bioinformatics companies, and applied proteomics to large numbers of problems of scientific, clinical and industrial importance. We believe this body of experience is unusual and unique and makes this book of relevance to proteomic researchers in all areas of academic and industrial biology and medicine.

For it to be possible to write and produce this book, we are grateful for the efforts and patience of the authors of all chapters. We also acknowledge support from Australian and Swiss universities and research institutes, the Swiss Institute of Bioinformatics and the companies Proteome Systems and Geneva Bioinformatics (GeneBio) which employed some authors during their writing. We also acknowledge support from public funding agencies, including the Australian National Health and Medical Research Council and the Swiss National Science Foundation which have supported our research in recent years.

Finally, we would like to acknowledge the efforts of proteomic researchers worldwide, whose work we draw on, discuss and occasionally critique. Proteomic research is a fast-paced, growing, yet challenging area. We hope that this book will serve to further grow the field and to encourage many new researchers (young and old) to join in this endeavour. There remains much work to be done!

April 2007

Marc Wilkins, Ron Appel,
Keith Williams, Denis Hochstrasser

Foreword to Second Edition

Ten years has elapsed since the publication of the first book on proteomics by the editors of the present book. Rather than 'proteomics', the book was entitled *Proteome research: new frontiers in functional genomics*. The idea was to establish a continuity with the Genome Analysis Project, and especially the sequencing of the human genome which was under way. However, it was already clear to some of us that a new revolution in biology was being launched: the introduction of a new paradigm permitted shifting the focus of investigation from DNA sequences to structures and functions of proteins, interacting between themselves and with other molecules, including DNA, in ways not encoded in DNA sequences. After completion of the sequencing of DNA of human and other species, the picture became even clearer. As is often the case in the history of science, the previous paradigm dominated by DNA technologies allowed for discoveries which turned this paradigm upside down. 'Proteomics' – the study of the proteome, i.e. the complete set of proteins in a cell or tissue – is one of the words being used today to name the new paradigm, together with the more general expressions 'biocomplexity' and 'systems biology'. But one should not be mistaken: proteomics is not a plain continuation of genomics. DNA sequences are being used now as an indispensable source of data regarding the first level of protein structures. However, this only marks the beginning of an entirely new story. Moreover, the same protein may have completely different functions in different tissues, even in the same cell, depending upon its localization in the cell and the state of activity of the latter. Expressed DNA sequences do not tell much about three-dimensional structures of proteins or their modifications in cellular microenvironments, nor about the dynamics of their synthesis, activation and inactivation, all of these determining their functions. Knowledge of the proteome is not limited to the pattern of expressed proteins identified from DNA sequences in DNA microarrays. This has prompted a change in the whole of biological thinking. For several decades, after the extraordinary discoveries of DNA structures and functions in the 1960s, molecular genetics and genomics were a source for *explanations*, giving answers to century-old questions regarding the nature of processes specific to living beings, such as metabolism and reproduction.

These explanations were based mainly on the metaphor of a computer program written in DNA sequences, the so-called genetic program. In spite of their being relatively simplistic, such explanations were accepted by the majority of biologists owing to their heuristic value. Protein physicochemistry, a very active field in the 1950s, was not fashionable anymore and had been almost abandoned. Among the reasons advocated were DNA technologies were easier and looked more promising. At the same time, the authors of these two books were developing two-dimensional gel electrophoresis techniques and mass spectrometry dedicated to the analysis of proteins and global protein expressions in cells and tissues. Thus, they emerged at the front line of biological research when it became clear that genomics by itself was able to provide knowledge of one-dimensional structures only, and very little knowledge of function.

This second book describes the progress made during the last 10 years. The main efforts aimed not only at improving the techniques, with the help of bioinformatics and data bases of DNA libraries, but also at tackling the more difficult problems of following protein modification and function in conditions as close as possible to their in vivo states. Progress has been made in developing reliable techniques to provide catalogues of proteins used as signatures of different normal and pathological cellular states, under well-defined conditions such as cancer versus normal cells in a given tissue. However, while this work was under way, it became clearer and clearer that post-translational modifications of proteins had to be taken into account as exhaustively as possible if protein structures were to be related to biological functions. In addition to phosphorylation, glycosylation, methylation and other covalently related modifications, more subtle intermolecular interactions are being looked for and protein–protein interaction maps are already being investigated. All these tools provide additional data which question more and more the complexity of functional regulations. How are all these interaction networks being regulated and how do they produce the observed functions? It is unlikely that a simple universal answer will be given to this question, in the form of the universal genetic code, ‘identical from bacteria to elephants’ according to the saying. Rather, local, ad hoc models will have to be designed and adapted to particular questions. Some medical applications are already being reported in diagnosis and drug development. Hopefully, they will develop into individualized medicine if not only individual genomes but also proteomes are made available in some distant future.

In any case, proteomics belongs to a world of postgenomics. This world has opened up a new era where more and more questions are raised rather than answers given owing to the formidable complexity being revealed. For example, there are different proteomes to be studied in more than 200 cell types (for humans only) expressing protein patterns differently, at different times, and in different conditions.

As George Klein put it nicely in a seminar on the cellular signaling pathways possibly disturbed in cancer: “Biologists must not only accept to live with complexity but to love complexity”. He was quoting Tony Pawson on cell signal transduction who pointed out that the complexity we see is nothing compared to the real complexity that exists.

Proteomics, as it is presented in this book, will most likely help biologists to 'love complexity', i.e. to be stimulated by the new problems and by the technical and theoretical tools being developed to approach them more and more efficiently.

February 2007 Henri Atlan
Professor Emeritus of Biophysics in Paris and Jerusalem
Director of the Human Biology Research Center at the Hadassah University
Hospital in Jerusalem and Director of Research at the Ecole des Hautes
Etudes en Sciences Sociales in Paris

Contributors

ALLARD, L.
Dpt R&D Immunoessais et Protéomique, BioMérieux S.A., Chemin de l'Orme, 69280 Marcy L'Etoile, France

APPEL, R.D.
Proteome Informatics Group, Swiss Institute of Bioinformatics, Computer Science Department, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

BAIROCH, A.
Swiss-Prot Group, Swiss Institute of Bioinformatics, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

BINZ, P.-A.
Geneva Bioinformatics (GeneBio) S.A., 25, avenue de Champel, 1206 Geneva, Switzerland, and Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

BOSCHETTI, E.
Ciphergen Biosystems Inc., Fremont, CA 94555, USA

BOUGUELERET, L.
Swiss-Prot Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

CITTERIO, A.
Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Polytechnic of Milan, Via Mancinelli 7, Milan 20131, Italy

CORTHALS, G.L.
Protein Research Group, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, P.O. Box 123, 20521 Turku, Finland

COUTÉ, Y.
Biomedical Proteomics Research Group, Department of Structural Biology and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Gavin, A.-C.
EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Gooley, A.A.
SGE Analytical Science Pty Ltd., 7 Argent Place, Ringwood, VIC 3134, Australia

Herbert, B.R.
Proteomics Technology Centre of Expertise, Faculty of Science, University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007, Australia

Hernandez, P.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Hochstrasser, D.F.
Department of Structural Biology and Bioinformatics, Department of Genetic & Laboratory Medicine, Laboratory Medicine Service, Geneva University and University Hospital, 24, rue Micheli-du-Crest , 1211 Geneva 14, Switzerland

Hoogland, C.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Lescuyer, L.
Department of Genetic & Laboratory Medicine, Laboratory Medicine Service, Geneva University Hospital, 24, rue Micheli-du-Crest , 1211 Geneva 14, Switzerland

Lisacek, F.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Packer, N.H.
CORE of Functional Proteomics and Cellular Networks, Macquarie University, Sydney, NSW 2109, Australia

Palagi, P.M.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

Righetti, P.G.
Department of Chemistry, Materials and Chemical Engineering "Giulio Natta", Polytechnic of Milan, Via Mancinelli 7, Milan 20131, Italy

ROSE, K.
Department of Structural Biology and Bioinformatics, University of Geneva,
CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland

SANCHEZ, J.-C.
Biomedical Proteomics Research Group, Department of Structural Biology
and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211
Geneva 4, Switzerland

WALTHER, D.
Proteome Informatics Group, Swiss Institute of Bioinformatics, CMU, rue
Michel-Servet 1, 1211 Geneva 4, Switzerland

WILKINS, M.R.
School of Biotechnology and Biomolecular Sciences, University of New
South Wales, Sydney, NSW 2052, Australia

WILLIAMS, K.L.
BKG Group, P.O. Box 580, Avalon, NSW 2107, Australia

ZIMMERMANN-IVOL, C.G.
Biomedical Proteomics Research Group, Department of Structural Biology
and Bioinformatics, University of Geneva, CMU, rue Michel-Servet 1, 1211
Geneva 4, Switzerland

Contents

1 Ten Years of the Proteome

Marc R. Wilkins and Ron D. Appel

Abstract

The concept of the proteome is now over 10 years old. As with all anniversaries, it is a good time to look back and reflect on what has been achieved in the area that we now call proteomics. What has been done well? What has been done not-so-well? What has been achieved, and what still eludes us? This review will briefly explore some of these questions, with respect to protein separations, mass spectrometry, and proteomic bioinformatics.

1.1 Introduction to the Proteome

The editors of this book have been carrying out research and development in proteomics for more than 20 years. They developed techniques for the analysis of proteins and global protein expression (Williams et al. 1991; Hochstrasser et al. 1988) and software algorithms and tools for the interpretation of the results obtained using such analytical tools (Appel et al. 1988; Wilkins et al. 1995). While the idea of observing the protein expression of genomes in a holistic manner rather than one protein at a time arose with the advent of 2-D gels, the concept of the proteome itself was only introduced by Marc Wilkins in 1994 at a conference in Siena, Italy[1], having coined the term earlier that year in association with his then PhD supervisor Keith Williams. The first papers that began to use the term were published shortly thereafter (Wilkins et al. 1995; Wasinger et al. 1995), and the first book on proteomics was published in 1997 (Wilkins et al. 1997). Ten years has now passed since the publication of that first book, and as with all anniversaries, it is a good time to look back and reflect a little on what has been achieved in the area we now refer to as proteomics. What has been done well? What has been done not-so-well? What has been achieved, and what still eludes us? Here we will suggest answers to these questions. At the same time, we will comment on what we have sought to achieve in this book, and provide a brief précis on its contents.

[1] First Siena conference, 2D electrophoresis: from protein maps to genomes, 5–7 September 1994.

M.R. Wilkins, R.D. Appel, K.L. Williams and D.F. Hochstrasser (Eds.)
Proteome Research: Concepts, Technology and Application

1.1.1 What's in a Word?

The words 'proteome' and 'proteomics' have been widely adopted by the biological community. In the 10 years since their introduction, their use has grown very rapidly (Fig. 1.1). In fact over 4,000 proteomics research and review articles were published in 2005. This has been fuelled by increasing numbers of journals that have arisen to serve the field, including *Proteomics*, *Proteomics-Clinical Application*, *Practical Proteomics*, *Journal of Proteome Research*, *Molecular and Cellular Proteomics*, *Proteome Science*, *Current Proteomics*, *Genomics and Proteomics*, *Briefings in Functional Genomics and Proteomics*, *Genomics Proteomics Bioinformatics* and *Expert Review of Proteomics*. In addition, proteomics research is increasingly published in a variety of other journals, so it has become established as a valuable means to obtain insight into the complexities of biological systems.

If we are simply measuring the progress of a field by its use of language, we might ask if the growth of proteomics is just a reflection of the so-called -omics revolution, or does it show a true growth in the field? The volume of work published in two other newer -omics areas, metabolomics and glycomics, is tiny by comparison, with 433 and 115 manuscripts having been published in 2005, respectively. Proteomics is clearly more widespread and established.

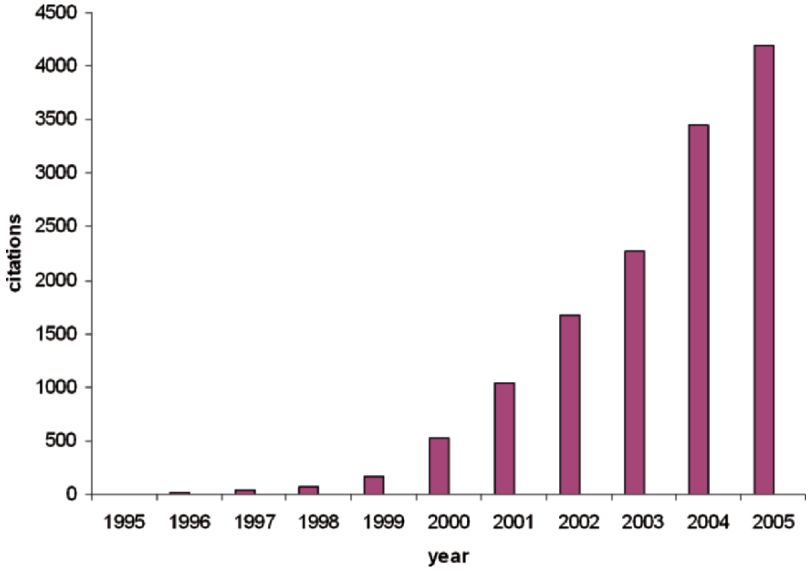

Fig. 1.1 Publications in the field of proteomics and proteome research have grown rapidly in the last 10 years. This was measured by querying the NCBI PubMed database for each year with the words 'proteome' or 'proteomics'. Note, however, that some articles may have been counted twice by this approach

1.1.2 Could Things Have Been Different?

So, would the world have been a different place had the term ‘proteome’ not been coined? Some commentators have argued that a combination of technical advances in separations technology (gel-based and chromatography-based), in mass spectrometry, and the explosion of information available from genome sequencing efforts have largely driven an increased interest in protein chemistry (Blackstock 2004).

While this is certainly true, it may be argued that the new language has brought renewed focus and legitimacy to protein chemistry that had previously been absent, largely due to the enormous shadow cast by genomics and other nucleic acid based approaches. The new language has also influenced biochemical thinking to move from a one-protein-at-a-time perspective to a more global view. Linguistically, it has been argued that thought cannot exist without language.[2] The proteome and proteomics are examples of this, as are other -omic words which were coined thereafter.[3] The new language and terminology has already helped a gamut of analytical technology to find its place in science and literature. New language in other fields will likewise legitimise emerging technology, focus thinking and also assist the funding of research in these areas.

1.2 Proteomics Is Technology-Driven

If we are to ask what has been done well in proteomics to date, one would have to pay particular attention to the development and dissemination of new technology. In a 10-year period, there have been a number of significant advances that, together, have transformed protein chemistry into the science of proteomics. Importantly, it has been a combination of conceptual breakthroughs and technical advances in separations techniques, mass spectrometry, protein chemistry and bioinformatics which have made this possible. The flood of nucleic acid sequence and genomic information, made available in sequence databases, was another essential co-requisite.

1.2.1 Protein Separations

Initially, proteomics researchers had a goal of visualising all proteins from a proteome on a single, or perhaps one acidic range and one basic range (2-D) polyacrylamide gel. This was happening in the late 1980s, and there was enormous excitement about the possibility of being able to see all

[2] Ferdinand de Saussure, Professor of Linguistics at Geneva University 1901–1913.

[3] See Chitty (2006) for an amusing list of new -omic words.

proteins in a proteome. However, it did not take long to realise that the separation and visualisation of all proteins from a proteome was not a straightforward task. In the mid-1990s, the availability of the first genome sequences and predicted proteomes allowed theoretical 2-D gels to be calculated, showing where each protein spot should be found (Urquhart et al. 1998). This revealed a bimodal distribution of proteins, with the majority of proteins having isoelectric point (pI) 4–6.5 and another group of proteins having pI 8–12. Most proteins had a mass of less than 100 kDa. The comparison of these theoretical maps with experimental 2-D gel separations immediately highlighted shortcomings with 2-D gels in that they were poor in resolving very acidic, very basic or very high mass proteins. A meta-analysis of proteins seen on 2-D gels and those predicted theoretically from genomes of *Escherichia coli*, *Saccharomyces cerevisiae* and *Bacillus subtilis* highlighted two additional issues (Wilkins et al. 1998). The first was that hydrophobic proteins were largely absent from the 2-D gels and that low-abundance proteins present at less than 1,000 copies per cell were likely to be undetectable, owing to limitations on the loading capacity and staining sensitivity of the 2-D gel process.

Since that time, a series of important technical advances have been made to help us see more proteins in the proteome. The latest advances associated with 2-D electrophoresis are discussed in Chap. 2. Broadly speaking, a number of strategies have been adopted. These include the running of narrow pI range gels to ‘zoom in’ on a particular region of the proteome, the fractionation of samples into either biologically (e.g. organelles) or physicochemically distinct fractions (e.g. membrane proteins) that can then be analysed appropriately, the enrichment or depletion of proteins of interest from a sample, along with new solubilisation and gel running techniques to assist in the analysis of the more difficult proteins. Importantly, fractionation has provided an avenue to load more of the relevant portion of samples of interest onto 2-D gels, thus assisting in the detection of lower-abundance proteins.

To completely bypass many of the challenges of working with complex mixtures of proteins, a conceptually different strategy emerged for protein analysis in proteomics. Called ‘shotgun proteomics’, probably inspired by the shotgun DNA sequencing approaches that were developed by Venter et al. (1998), it involves taking complex mixtures of proteins or indeed a whole proteome, and digesting all proteins to peptides with endoproteinases of known specificity. The resulting mixtures of peptides, which are physicochemically more homogenous than their parent proteins although greater in number, are then analysed using 2-D liquid chromatography and tandem mass spectrometry. Peptide fragment data are matched against sequence databases (Wolters et al. 2001) to determine the proteins present in a sample. Whilst this approach has limitations, notably the loss of protein isoforms (see Chap. 5), it provides an alternative to gel-based analyses for the separation and identification of large numbers of proteins from a proteome.

1.2.2 Mass Spectrometry

The last 20 years has brought astonishing advances in mass spectrometry technology. These advances have helped establish the science of proteomics. Mass spectrometers, whilst remaining expensive, now have remarkable mass accuracy and resolution, can analyse femtomolar quantities of peptides and proteins, and are increasingly automated. Two means of ionisation of proteins and peptides are in widespread use, electrospray ionisation and matrix-assisted laser desorption/ionisation, and these are teamed with a variety of mass analysers and detectors (see Chap. 3).

Mass spectrometers have all but superseded Edman degradation as the method of choice for protein identification. Two techniques, namely peptide mass fingerprinting and peptide fragmentation, can be used. Peptide mass fingerprinting has been used in a number of massive projects, for example more than 20,000 proteins were analysed as part of a large-scale analysis of yeast protein complexes (Gavin et al. 2002). However, peptide mass fingerprinting is losing favour to higher-confidence peptide fragmentation approaches that are able to fragment multiple peptides from the same protein. Nevertheless, it should be noted that mass spectrometers typically do not sequence peptides or proteins *per se*. They instead allow us to infer sequences by matching peptide fragmentation data against sequence databases. Routine *de novo* sequencing remains complex and is thus a work in progress (see Chap. 3).

In addition to protein identification, a myriad of new mass spectrometry approaches have been developed for the quantitative analysis of two or more samples. Such comparisons are of great scientific interest for the detection of biomarkers and the understanding of the multiplicity of changes that can occur when a proteome is perturbed by intrinsic or extrinsic forces. Previously, the comparison of protein expression in two or more samples was done by 2-D gel electrophoresis and computer image analysis (see Sect. 4.2). This approach has been successfully used in a large number of studies and remains widespread. The newer mass spectrometry based approaches are a significant advance and essentially use different stable isotopes to label proteins from two or more samples (Gygi et al. 1999). The samples are then mixed together and co-analysed. The high mass accuracy of the mass spectrometers allows the isotopic variants to be separated and relative quantitation to be undertaken. This concept has now been developed in a number of different ways (see Sect. 4.3) and whilst not perfect is providing a new means to undertake comparative analysis of two or more complex samples.

A final area in which mass spectrometry is now playing a major role is in the characterisation of proteins. Post-translational modifications of proteins are of increasing interest as they are key to the control and modulation of many processes inside the cell. Our recent appreciation of their roles in protein–protein interaction networks, whereby interactions between many proteins require the presence of certain post-translational modifications (Pawson and Nash 2003), is providing even greater impetus for their study.

Many sophisticated analytical strategies have been developed for the analysis of modifications (see Chap. 5) and these have now been applied, in some cases, on a proteome-wide scale. Protein phosphorylation has been a particular focus (Beausoleil et al. 2004). These analyses of modifications, whilst of large scale, remain incomplete. Yet they are giving the first glimpses of the dynamics of post-translational modifications in the proteome.

1.2.3 Making Sense of All the Data

New strategies for proteomic analysis and improvements in separation and analytical technologies have, without doubt, increased the amount and complexity of proteomic data. However, it is the combination of analytical approaches with sophisticated new bioinformatics that has allowed researchers to better generate, analyse, visualise and contextualise proteomic data and thus better understand their biological system under study.

Software for the quantitative analysis of protein expression on 2-D gels, particularly in association with new fluorescent stains, has drastically improved our capacity to find qualitative and quantitative expression differences between two gels or two populations of samples (see Chaps. 4, 6). Protein-identification software, vital to most aspects of proteomics, has incorporated statistical methods to allow identification confidence to be calculated. Bayesian and non-Bayesian statistics have been applied to the problem of protein identification by peptide mass fingerprinting (Perkins et al. 1999). For shotgun proteomics experiments, where thousands of protein identifications cannot possibly be verified by a human, searching against 'normal' and 'randomised' sequence databases is now used to estimate false-positive rates and thus overall identification confidence. The issue of protein-identification confidence has been the subject of much discussion, and proteomics journals have now released guidelines on protein identification which authors are expected to follow for their work to be published (Wilkins et al. 2006; Carr et al. 2004). In addition to improved strategies for protein identification, data-processing pipelines have been developed to automate the peak-picking and peak-matching processes for the hundreds to thousands of mass spectra that may be generated from the larger proteomics experiments. Workspaces have also been developed for the management and storage of the huge volume of data produced (Rauch et al. 2006).

Dramatic advances in the bioinformatics of post-translational modifications have also been made in recent years. Software tools for the discovery of protein modifications in mass spectrometry data are available, and are used for the analysis of peptide mass and peptide fragmentation data (see Chaps. 3, 5). Modifications such as methylation, acetylation, oxidation and phosphorylation can thus be found. The analysis of protein glycosylation, which produces enormously complicated mass spectrometry fragmentation spectra, is expected to become commonplace now that glycan structure databases and 'glycan mass fingerprinting' structure assignment tools have been developed.

The most profound advance in proteome bioinformatics has been its capacity to bridge the gap between technology and biology. Bioinformatics has been developed to allow the visualisation of cells and tissues after their direct laser scanning with mass spectrometry. This is a stunning new advance that is giving insight into the micro- and macroheterogeneity of protein expression in cells (see Chap. 6). In differential-display experiments, visualisation tools have become indispensable to highlight small changes that are undetectable when analysing each data item separately (see Chap. 6). A bioinformatics capacity to map all differentially expressed proteins onto the gene ontology also provides a 'big picture' understanding of the molecular function and biological processes that may be changed in association with a phenotype (see Chap. 7). It can reveal which changes in proteins may be functionally related. Where proteomic studies find differential expression of enzymes, the bioinformatic contextualisation of such proteins in the metabolome or 'reactome' (Reactome 2006) can reveal direct links between the proteome and metabolites in the cell. Bioinformatics is also allowing us to better understand the complexities of protein–protein interactions and interaction networks and how these change in association with disease (see Chap. 8). Figure 1.2, for example, shows the result of mapping protein function onto an interaction network. It is expected that these and other increasingly rich visualisations will assist in understanding the complexities of the proteome and the cell.

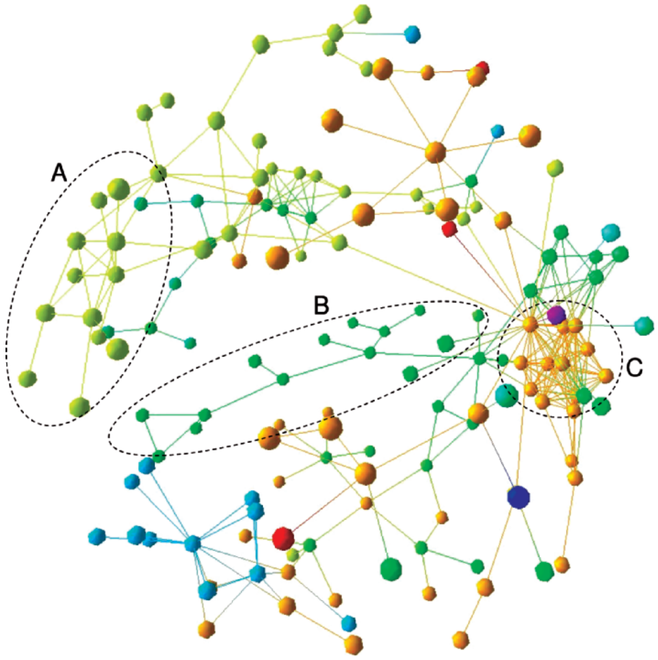

Fig. 1.2 Co-visualisation of protein–protein interaction and protein function. Some groups of directly interacting proteins have the same colour, indicating a common molecular function. Examples of molecular functions performed by such groups include *A* RNA binding (*yellow*), *B* structural molecule activity (*green*) and *C* protein binding (*orange*). (From Ho 2006)

1.3 What Has Proteomics Delivered?

A question we must always bear in mind when assessing emerging biomedical technology is what has it delivered to date or what is it likely to deliver? For proteomics, has the excitement associated with new methods translated into biological insights of scientific importance? This is a difficult question at the best of times, and since the technology has shifted the biological paradigm from a one-protein-at-a-time view to a new 'global' view, the question becomes almost impossible to answer without the benefit of the passage of time. However, it is clear that proteomics has already provided insight in a number of key areas. These may be enunciated as follows:

1. The proteome is no longer largely unknown. Substantial audits of protein expression, from gel-based studies coupled with mass spectrometry to those based on shotgun proteomics and tandem mass spectrometry, have given and will continue to give insight into which proteins are present in a particular cell or tissue. This is not to say we know the function of each protein in the proteome, but at a minimum we now have great insight into what proteins are present at any one time.
2. The 'higher order' of the proteome, obtained from large-scale studies of protein–protein interactions in the cell using proteomic techniques, is just starting to be revealed (see Chap. 8). The widespread adoption of this view will require another paradigm shift as it requires a global protein-based view of the cell and an acknowledgement that proteins do not act alone but participate in protein–protein interactions to form functional units in the cell.
3. Proteomics is providing a major new avenue for the discovery of medical biomarker proteins of diagnostic and/or prognostic significance. As proteomic technology is supremely well suited to the analysis of soluble proteins, the analysis of proteins from body fluids has been and will continue to be a fruitful endeavour. This is explored in detail in Chap. 9.
4. Proteomics is providing high-resolution data to supplement existing biomedical techniques. Toxicology, which has traditionally relied on histopathology and the evaluation of a small number of blood-associated proteins and metabolites, is using proteomics to better understand the effects and side effects of drugs (reviewed in Wilkins 2006). Immunoproteomics, the application of proteomics to the discovery of immunoreactive proteins and peptides, is starting to give stunning insight into how the body distinguishes self from non-self and what happens when this goes wrong (reviewed in Purcell and Gorman 2004).
5. Metaproteomics, a new term used to describe the shotgun proteomics analysis of mixtures of microbial species, is providing insights into microbial diversity and interactions that would otherwise be impossible to achieve. Microbial species that are difficult to culture in the laboratory can

be studied directly by a shotgun proteomics analysis of environmental samples. Whilst this currently requires a parallel metagenomic analysis (Venter et al. 2004) to allow protein identification, it is expected that this approach will become increasingly widespread.

1.4 What Still Eludes Us?

Finally, we may wish to ask which aspects of the proteome remain unexplored, and where has proteomics yet to be effectively applied? Whilst not an exhaustive answer to this question, the following points may be made:

1. The separation and detection of all proteins in the proteome remains a challenge. Low-abundance proteins are particularly elusive, owing to the large differences in concentration of proteins in many samples. Fractionation of samples can help with this, as may new 'equaliser' technology (see Chap. 2), but new approaches are still required to address this issue. Proteins that are very large, very basic and very acidic also remain problematic for 2-D gel analysis.
2. It is not possible to compare one interactome with another. The incredible complexity of the interactome, and the fact that interaction networks are built from the results of thousands of individual experiments, makes it impossible to currently compare one interaction network with another. Blue native electrophoresis, which separates large numbers of protein complexes under gentle conditions (Schagger 2001), may provide a means to address this.
3. *De novo* sequencing of proteins remains difficult. Researchers studying unusual organisms for which there is little nucleic acid sequence data cannot identify proteins of interest. They are also precluded from using shotgun proteomics techniques. De novo sequencing could address this issue; however, it remains a work in progress. Improvements in mass spectrometry and bioinformatics such as 'open-modification search' strategies (see Chap. 3) are required before this can become a robust and widespread technique.
4. We cannot monitor changes in the proteome in real time. The need to destroy cells for proteomic analysis, and a lack of alternative technology to mass spectrometry, makes it impossible to understand the myriad of changes that continuously occur in the cell. Whilst it is not clear how we may achieve such a feat, advances in high-magnification microscopy of living cells may prove to be a fertile ground for future developments.
5. Proteomics is currently semiquantitative, not quantitative. A capacity to undertake absolute rather than relative quantitation is desirable. Immunoassay techniques have been used to quantitate a large proportion of the *S. cerevisiae* proteome in copies per cell (Ghaemmaghami et al. 2003).