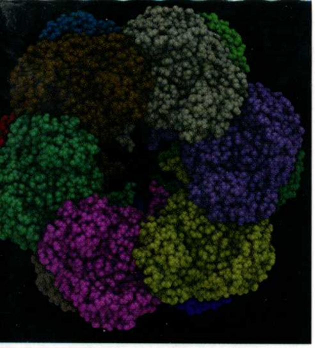
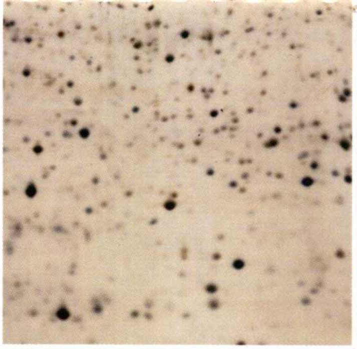
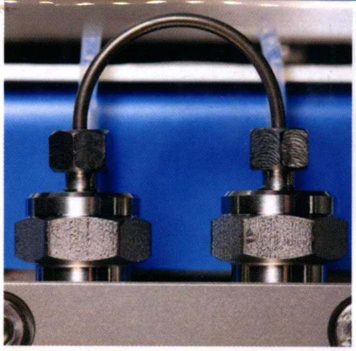






Introducing Proteomics

from concepts to sample separation, mass spectrometry and data analysis

JOSIP LOVRIĆ

WILEY-BLACKWELL

Introducing Proteomics

From concepts to sample separation, mass spectrometry and data analysis

Josip Lovrić

Faculty of Life Sciences, University of Manchester

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices:
9600 Garsington Road, Oxford, OX4 2DQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloguing-in-Publication Data
Lovrić, Josip.
Introducing proteomics : from concepts to sample separation, mass spectrometry, and data analysis / Josip Lovrić.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-03523-8 (cloth) – ISBN 978-0-470-03524-5 (pbk.)
1. Proteomics. I. Title.
[DNLM: 1. Proteomics. QU 58.5]
QP551.L875 2011
572′.6 – dc22

2010029400

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

This book is published in the following electronic formats: ePDF ISBN 978-0-470-67021-7.

Typeset in 9/12pt Times-Roman by Laserwords Private Limited, Chennai, India
Printed in Singapore by Markono Print Media Pte Ltd.

First Impression 2011

Introducing Proteomics

To my family, I hope they'll have me back . . . . . .

A narućito naknađujem ovo knjigu za moje roditelje.

Preface

The term proteomics was coined in the mid-1990s by the Australian (then post-doctoral) researcher Marc Wilkins. The term arose in response to the spirit of the day; researchers working in genetics developed genome-wide approaches and were very successful at the time. Researchers working on proteins rather than genes also felt that the time was right for a more holistic approach – rather than working on a single protein at a time, many (if not all) proteins in a single biological system should be analysed in one experiment. While surely the will was there and some good foundational work was done at the time, it still took about another five years before technologies were developed far enough, so that proteomics became a concept that could deliver some (but still not all) of the answers researchers hoped to be able to get by using it.

Historically proteomics was driven mainly by researches coming from the field of 2D gel electrophoresis. These 'bluefingers' joined forces with experts in mass spectrometry and bioinformatics. It was the combination of these fields together with the genomic revolution that created the first proteomic approaches. These were inevitably studies using 2D gel electrophoresis in combination with mass spectrometry of 'isolated' spots, often using MALDI-ToF mass spectrometry. In the beginning the development of the ionization technologies of MALDI, ESI and nano ESI were critical steps to allow the mass spectrometric analysis of biological material with reasonable sensitivity.

Together with advances in all fields concerned, it was major developments in gel free, hyphenated peptide separation technologies that allowed proteomics to prosper in more recent times. Recent developments in gel based proteomics were confined mainly to more convenient sample handling and more pre-fabricated devices and most important computer based image analysis and new protein dyes, allowing for less variable results in a shorter time with less manual input. 2D HPLC in combination with tandem mass spectrometry is a hallmark of the development of hyphenated technologies. Modern proteomics is driven by the development of ever improving software to deal with the huge amount of data generated, allowing better and more efficient data mining; new mass spectrometers allowing new imaging approaches or qualitatively better approaches through improvements in versatility, accuracy, resolution or sensitivity. Also developments in labelling reagents and affinity matrices allow more intelligent approaches, more tailored to specific questions, such as quantitative analyses and analyses of phosphorylations and other posttranslational modifications. Nano-separation methods become more routine and combination of multi-dimensional separation approaches become feasible, allowing 'deeper' views of the proteome. And if all these developments were not enough, there is a plethora of more specialized developments, like the molecular scanner (Binz *et al.*, 2004), MALDI imaging mass spectrometry for tissues, organs and whole organisms such as the mouse or rat (Caldwell and Caprioli, 2005) or Laser Capture Microdissection (Jain, 2002) which enables proteomics analysis from just a dozen of cells (Nettikadan *et al.*, 2006). In a book like this it is impossible to do justice to all these developments, and they will be mentioned as we go along, especially in Chapter 5 on strategies in proteomics. Sadly, some fields such as 3D structural analyses have had to be omitted.

Next to complete 'work floors', the mass spectrometers and separation devices (e.g. nano HPLC, free flow electrophoresis equipment) that come with the territory represent the biggest capital investment for laboratories getting involved in proteomics, ranging from some US$ 160 000 to more than a million dollars per item. In the early days of proteomics, many developments were driven by scientists rather than industries. Since 2000, proteomics has become big business, with the potential for companies to sell hundreds of mass spectrometers instead of a dozen a year to the scientific community.

Away from 'classical' approaches there have been huge developments in very diverse fields such as protein fluorescent staining, chemical peptide modification, ultra-accurate mass spectrometers, microscope assisted sample collection, improved sample treatment, isobaric peptide tagging and of course bioinformatics, to name just a few, that have opened up a whole range of new possibilities to tackle biological problems by proteomic approaches. It is this diverse group of fields that contribute towards making proteomics such a vibrant and interesting field, on the one hand, but also a field that may seem difficult to get started in, on the other hand.

This is where we aim to place this book: to give an introduction to the complete field of proteomics without delving too deeply into every single area within it, because for most of these areas there is excellent specialist literature available.

In this respect the book aims primarily to give a basic understanding of the most important technologies. At the same time it intends to allow the reader to develop an understanding of the possibilities, but also the limitations, of each of the technologies or their combinations. All this is presented with the aim of helping the reader to develop proteomic approaches that are suited to the needs of their specific research challenge.

WHO WILL BENEFIT FROM THE BOOK

This book is aimed at diverse groups of potential users. In the academic world it is written easy enough to be useful and aimed at undergraduate students to give an introduction to the field of proteomics; so many biochemical/physical principals are explained at that level.

On the other hand, this book will also be useful for postgraduate students and more senior researchers in academia and industry. While it brings an overview and an explanation of principles to postgraduate students who may be about to start to work on a proteomic project, it will also explain the possibilities and limitations of a potential proteomics approach for a principal investigator and give them an idea of the sort of financial and intellectual commitments necessary.

It will be a useful tool for experienced researchers in the field of proteomics to 'catch up' on areas that were outside their focus for a while or have developed only recently. It may also help scientists to understand the needs of a certain approach and help them with their planning; be it for starting collaborating with someone in the field of proteomics, or to help such a collaboration to be successful or for writing a new grant in this field.

While this book does not contain recipes or manuals for instruments, it will be of great benefit in helping people to get trained practically in the field, since it explains all the major principles and puts them in a wider perspective.

I hope it will also help researchers from (apparently) distant areas of research to develop new approaches and identify fields in which further research into technologies might be necessary and possible to help proteomics to become and remain one of the sharpest tools in the box of biological and medico-pharmaceutical research.

REFERENCES

Binz, P.A., Mueller, M., Hoogland, C. *et al.* (2004) The molecular scanner: concept and developments. *Curr Opin Biotech*, **15**, 17–23.

Caldwell, R.L. and Caprioli, M.R. (2005) Tissue profiling by mass spectrometry. *Mol Cell Proteomics*, **4** (4), 394–401.

Jain, K.K. (2002) Application of laser capture microdissection to proteomics. *Methods Enzymol*, **356**, 157–167.

Nettikadan, S., Radke, K., Johnson, J. *et al.* (2006) Detection and quantification of protein biomarkers from fewer than 10 cells. *Mol Cell Proteomics*, **5** (5), 895–901.

Acknowledgements

My sincere thanks go to all those individuals who helped in the making of this book, be it by advice, by letting me have their data or by helping in the production process of this book. Special thanks are due to Dr Alistair McConnell, Dr Chris Storey, Dr David Knight, Fiona Woods, Gill Whitley, Dr Guido Sauer, Haseen Khan and her amazing project team, Izzy Canning, Nicky McGirr, Dr Paul Sims and Dr Songbi Chen.

Contents

1

Introduction

1.1 WHAT ARE THE TASKS IN PROTEOMICS?

1.1.1 The proteome

In genomics, one of the main aims is to establish the composition of the genome (i.e. the location and sequence of all genes in a species), including information about commonly seen polymorphisms and mutations. Often this information is compared between different species and local populations. In functional genomics, scientists mainly aim to analyze the expression of genes, and proteomic is even regarded by some as part of functional genomics. In proteomics we aim to analyze the whole proteome in a single experiment or in a set of experiments. We will shortly look at what is meant by the word analysis. Performing any kind of proteomic analysis is quite an ambitious task, since in its most comprehensive definition the proteome consists of all proteins expressed by a certain species. The number of these proteins is related to the number of genes in an organism, but this relation is not direct and there is much more to the proteome than that. This comprehensive definition of the proteome would also account for the fact that not a single individual of a species will express all possible proteins of that species, since the proteins might exist in many different isoforms, with variations and mutations, differentiating individuals. An intriguing example are antibodies, more specifically their antigen binding regions, which exist in millions of different sequences, each created during the lifetime of individuals, without their sequence being predictable by a gene. Antibodies are also a good example of the substantial part played by external influences, which define the proteome; for example, the antibody-mixture present in our bodies is strictly dependent on which antigens we have encountered during our lives. But of course a whole host of more obvious external factors influence our proteome, but not the genome (Figure 1.1).

Furthermore, the proteome also contains all possible proteins expressed at all developmental stages of a given species; obvious examples are different proteins in the life cycle of a malaria parasite, or the succession of oxygen binding species during human development, from fetal haemoglobin to adult haemoglobin (Figure 1.2).

On top of all these considerations, there are possible modifications to the expression of a protein that are not encoded by the sequence of its gene alone; for example, proteins are translated from messenger RNAs, and these mRNAs can be spliced to form different final mRNAs. Splicing is widespread and regulated during the development of every single individual, for example during the maturation of specific cell types. Changes in differential splicing can cause and affect various diseases, such as cancer or Alzheimer's (Figure 1.3).

As if all this was not enough variability within the proteome, most proteins show some form of posttranslational modification (PTM). These modifications can be signs of ageing of the protein (e.g. deamidation or oxidation of old cellular proteins; Hipkiss, 2006) or they can be added in an enzymatically regulated fashion after the proteins are translated, and are fundamental to its function. For example, many secreted proteins in multicellular organisms are glycosylated. In the case of human hormones such as erythropoietin this allows them to be functional for longer periods of time (Sinclair and Elliott, 2004). In other cases proteins are modified only temporarily and reversibly, for example by phosphorylation or methylation. This constitutes a very important mechanism of functional regulation, for example during signal transduction, as we will see in more detail later. In summary, there are a host of

Introducing Proteomics: from Concepts to Sample Separation, Mass Spectrometry and Data Analysis, Josip Lovrić

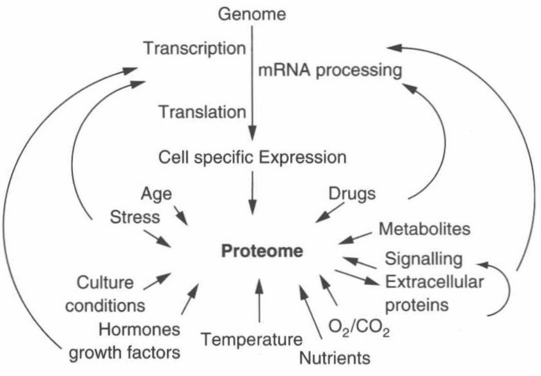


Figure 1.1 Influences on the proteome. The proteome is in a constant state of flux. External factors constantly influence the proteome either directly or via the genome.

relevant modifications to proteins that cannot be predicted by the sequence of their genes. These modifications are summarized in Figure 1.4.

Moreover, it is important to remember that the proteome is not strictly defined by the genome. While most possible protein sequences might be predicted by the genome (except antibodies, for example), their expression pattern, PTMs and protein localization are not strictly predictable from the genome. All these factors define a proteome and each protein in it. The genome is the basic foundations for the 'phenotype' of every protein, but intrinsic regulations and external influences also have a strong influence (Figure 1.5).

1.1.2 A working definition of the proteome

For all the above mentioned reasons most researchers use a more practical definition of the word 'proteome'; they use it for the proteins expressed in a given organism, tissue/organ (or most likely cell in culture), under a certain, defined condition. These 'proteomes' are then compared with another condition, for example two strains of a microorganism, or cells in culture derived from a healthy or diseased individual. This so-called differential proteomics approach has more than a description of the proteome in mind; its aim is to find out which proteins are involved in specific functions. This is of course hampered by the number of proteins present (some changes may occur as mere coincidences) and by the many parameters that influence the functionality of proteins, expression, modification, localization and interactions. While differential proteomics seems a prudent way to go, we have to keep in

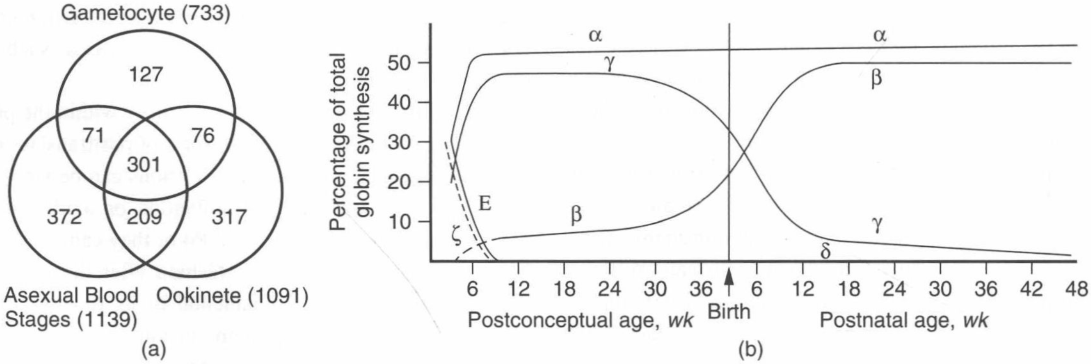


Figure 1.2 The composition of the proteome changes during ontology. (a) Plasmodium, the agent causing malaria, has a complex life cycle. Its asexual blood stage cycle lasts about 24 hours, then the sexual stages (gametocytes) develop within 30 hours and develop into the ookinetes after fertilization. A comprehensive proteomic study of these and other stages of the life cycle detected more than 5 000 proteins. The Venn diagram shows the number of total proteins identified in each specific stage in parentheses. The numbers in the Venn diagram represent the number of proteins involved in sexual development exclusive to one of the three stages shown in the picture. Over a third of the proteins in each state were found exclusively in one stage only, about 30–50% were common to all stages and about 10–20% were found in more than one of the three stages. (b) Humans express different globin species during their ontogenesis. These globin proteins come from different genes and bind the haeme group to form haemoglobins with specific characteristics essential for different stages of development. The figure shows how the relative production of different globin species changes in early human development. (a) Hall *et al.* (2007). © 2005 American Association for the Advancement of Science. (b) Modified from Wood (1976) and reproduced with permission. © 1976 Oxford University Press.

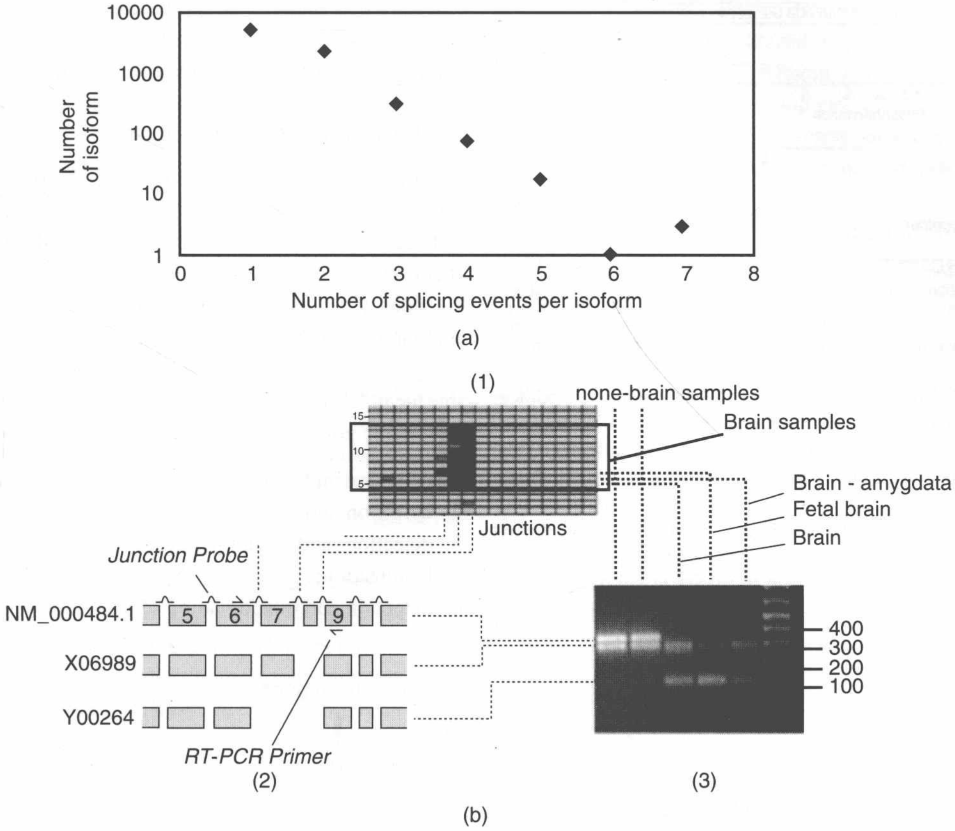
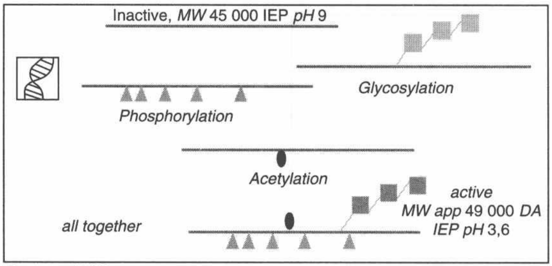


Figure 1.3 The importance of splicing. (a) The known frequency of splicing events for human proteins (Wang *et al.*, 2005). Splicing events were extracted form the SWISS-PROT database, one of the best-annotated databases for proteins. It can be assumed that there are a huge number of non-annotated splicing events. The number of proteins showing a certain number of splicing isoforms is shown. In the case of one splicing event per isoform, no alternative splicing isoform is annotated. (b) The mRNA for human β-amyloid precursor protein is spliced in brain tissues as compared to non-brain tissues. Alternative splicing of amyloid precursor protein may play a role in the development of human Alzheimer's disease. Screens for alternative splicing were performed on mRNAs microarrays (1) using splice event specific probes spanning two exons (2) and then confirmed by specific PCR reactions (3), using primers whose product length is influenced by splicing events. (a) Wang *et al.* (2005). © 2005 National Academy of Sciences, USA. (b) From Johnson *et al.*, *Science*, 2003; 302:2141–44. Reprinted with permission from the American Association for the Advancement of Science.

mind that the methods chosen for proteomic analyses will also determine the results; for example, if we use a gel-based approach, membrane proteins are almost completely excluded from the analyses. Furthermore, most analyses have a certain cut off level for the low abundant proteins. This means that proteins below (say) 10 000 copies expressed per cell are not easily measurable, because the approaches are usually not sensitive enough.

Even within this limited definition of proteomics we still face substantial tasks, as the proteome is defined not only by the physical state of the proteins in it (expression and modifications) but also by their subcellular location and their membership in protein–protein complexes of ever changing compositions. For instance, it makes a big functional difference to its activity if a transcription factor is inside or outside the nucleus and a proteomic study that fails to analyze the transcription factor's sub-cellular location will miss major changes in the activity of this transcription factor (Figure 1.7). A kinase that needs to be in a multiprotein complex to be active will be inactive when it is only bound to parts of that complex, an important difference that will be missed if we analyze only the



Common PTMs:
Phosphorylation
Glycosylation
Acetylation
Dehydroalanine (convertion from cysteine)
Disulfide bonds (oxidation)
Methylation
Sulfonation
Sumolylation
Ubiquitination
Farnesylation
Deamidation

Figure 1.4 Proteins are regulated by posttranslational modifications. Genes and splicing define the primary sequence of proteins. The primary sequence contains motives that allow different PTMs. Which of them are actually found on a protein at any given time in a specific tissue cannot be predicted. Often a combination of PTMs is necessary for active proteins. PTMs can change the 3D structure of proteins. They also change parameters such as apparent molecular weight and isoelectric point in gel-based protein separations.

presence of a protein but not the interaction partners. The same holds true for kinases that switch complexes and thereby regulate their target specificity (Kolch, 2005).

1.1.3 The tasks in proteomics

Most proteomic studies aim to correlate certain functions with the expression or modification of specific proteins; only few aim to describe complete proteomes or compare them between different species. For a functional correlation we need to analyze the most important protein features of functional relevance. We have already mentioned the analysis of proteins in proteomic studies – just what does this mean? Proteomic analyses can be summarized in terms of specific goals:

1. detection and quantification of protein level;
2. detection and quantification of protein modifications;
3. detection and quantification of sub-cellular protein localization;
4. detection and quantification of protein interactions.

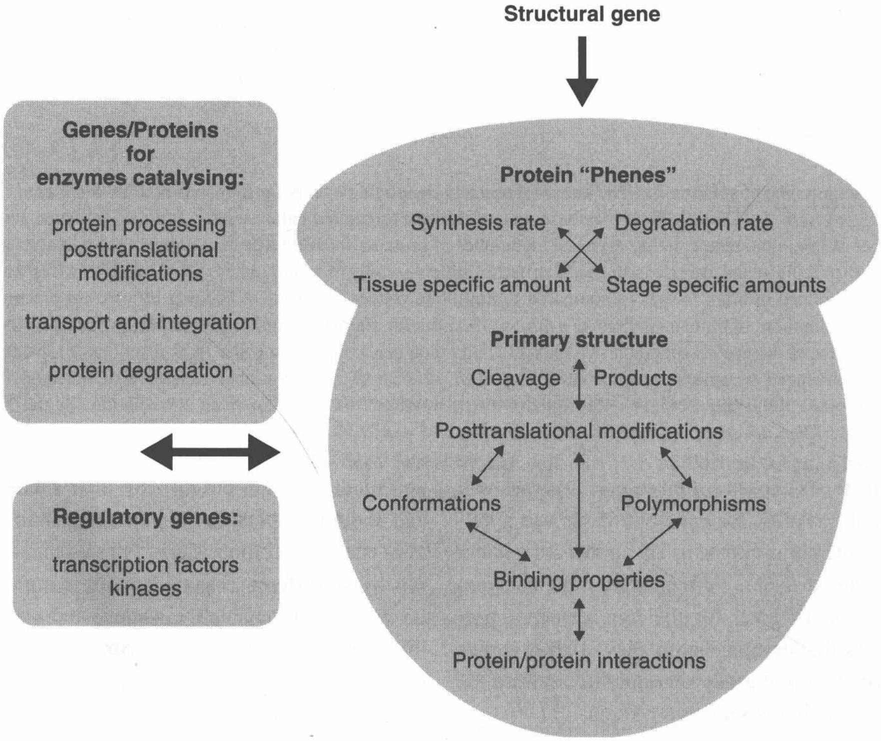


Figure 1.5 Proteins have a 'phenotype'. Similar to whole organisms, proteins can be regarded as having observable traits that are derived by genetic factors as well as influences from the surroundings they experience during their 'life'.

Historically, protein expression has been the first parameter analyzed by proteomics. While this involves a certain form of quantification (present/not present means usually at least a three- to tenfold difference in expression level), it is much harder to quantify proteins on a proteomic scale and many of the latest technological developments focus on this aspect (see Chapters 2–5). Since the abundance of proteins can vary from presumably a single protein to over a million proteins per cell, the quantifications have to cover a dynamic range of over 6 orders of magnitude in cells and up to 10 orders of magnitude in plasma (Patterson and Aebersold, 2003).

PTMs are very important for the function of proteins, and proteomics is the only approach to analyze them on a global scale. Nevertheless, the current approaches (e.g. phosphoproteomics) are by no means able to analyze all possible PTMs, and this remains a hot topic in the development of new technologies.

Before the onset of life cell imaging technology, fractionation of cells was the only method to analyze the subcellular localization of proteins. While being relatively crude and error-prone due to long manipulation times, fractionation studies are very successful in defining protein function. This holds true especially when not only organelles but also functional structures such as ribosomes (Takahashi *et al.*, 2003) or mitotic spindles can be intelligently isolated (Sauer *et al.*, 2005).

The detection of protein interactions is surely the most challenging of proteomic targets, but also a very rewarding one. In single studies the goal is often to identify all interacting partners of a single protein (see Figure 1.8), and several studies taken together can be used to identify, for instance, all interactions within a single signalling module (Bader *et al.*, 2003). Interactions on a truly proteomic scale have been analyzed only in some exceptional studies (Ho *et al.*, 2002; Krogan *et al.*, 2006) and the results are by no means complete, given the temporal and fragile nature of protein–protein interactions, the different results reached with different methods and their complexity.

Non-covalent and hence the most difficult to analyze are localization and interactions of proteins – although none of the above tasks is easily reached, considering the shear number of proteins involved, the minute amounts of sample usually available and the temporal resolution that might be required. Proteomic parameters can change from seconds or minutes (e.g. in signalling) to hours, days and even longer time periods (e.g. in degenerative diseases).

1.2 CHALLENGES IN PROTEOMICS

1.2.1 Each protein is an individual

Nucleotides are made up of four different bases each, and the structure of DNA is usually very uniform. Even if RNA forms more complex structures, we have many different buffers in which we can solubilise all known nucleotides. No such thing exists in proteomics. There is no buffer (and there probably never will be) that can solubilize all proteins of a cell or organism (Figure 1.6). Proteins are made out of 20 amino acids, which allows even a peptide that is 18 amino acids long to acquire more different sequences than there are stars in the galaxy or a hundred times more different sequences than there are grains of sand on our planet!

The average length of proteins is about 450 amino acids. The complexity that can be reached by such a

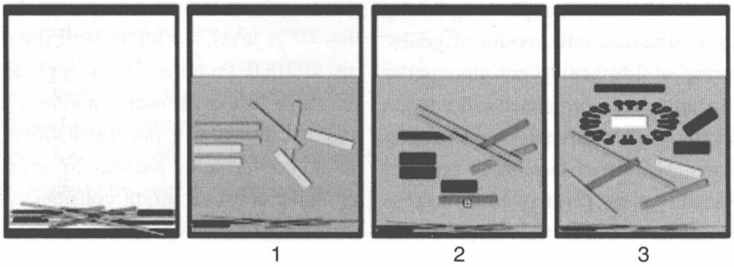


Figure 1.6 Protein solubilization. Complex mixtures of proteins (e.g. cellular lysates) can be solubilized in a variety of buffers (e.g. different ionic strength, pH). Some proteins will dissolve in one or the other buffer, but not in both, while some or most protein interactions are preserved (1/2). Adding detergents allows most proteins to be dissolved, but protein interactions are disrupted (3). Strong detergents even interfere with further manipulation or analysis of the proteins.

protein is beyond the imagination. More to the point, while almost every sequence of DNA will have fairly similar biochemical properties to any other sequence of similar length, with proteins the situation is totally different. Some proteins will bind to materials used for their extraction and so get lost in analyses, others will appear predominant in a typical mass spectrometry (MS) analysis because they contain optimal amounts and distributions of arginine and lysine. If proteins are very hydrophobic, they will not even get dissolved without the help of detergents. Some proteins show aberrant behaviour with dye; either they are stained easily or very badly. This behaviour makes absolute quantifications and even relative comparisons of protein abundances very difficult. Proteins can display highly dynamic characteristics; their abundances can change dramatically within minutes, by either rapid new synthesis or degradation. Some proteins are more susceptible to degradation by either specific ubiquitin dependent or independent proteolysis than others. These processes in turn can be triggered during cellular processes such as differentiation or apoptosis (active cell death). There are more than 360 known chemical modifications of proteins (see the 'Delta Mass' listing on the Association of Biomolecular Resource Facilities website, http://www.abrf.org). These include natural PTMs such as phosphorylation, glycosylation and acetylation, as well as artefacts such as oxidation or deamidation that might occur naturally inside cells but also as artefacts during protein preparation. There are of course also totally artificial modifications occurring exclusively during protein isolation, such as the addition of acrylic acid.

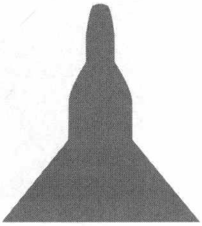
1.2.2 The numbers game

This variety explains how relatively complex organisms can manage to rely on a relative small amount of genes. The least complex forms of life are found among the viruses; in a typical example, a dozen genes will encode about 40 proteins by means of alternative RNA processing and controlled proteolysis. On top of this, these proteins are alternatively processed (e.g. by glycosylation) to regulate their function in different phases of the viral life cycle. In these relatively simple life forms the proteome is much more complex than the genome would suggest, and the more complex the life form, the more this gap widens. Bacteria have about 3 000–4 500 genes. In a typical example (if there are any 'typical' examples of these fascinating organisms!) like *Escherichia coli* there are 4 290 protein encoding genes plus about 90 only producing RNA. Splicing of mRNA is rare; PTMs are present in a variety of forms, but do occur rarely. In yeast (*Saccharomyces cerevisiae*) we detect about 6 000 genes and these are moderately modified. Splicing is a regular event, and so are differential glycosylation, phosphorylation, methylation and a host of other PTMs, resulting in a much higher number of protein isoforms than the pure addition of nuclear and mitochondrial genes would suggest. In multicellular organisms such as insects (e.g. the fruit fly, *Drosophila melanogaster*) or worms (e.g. the roundworm, *Caenorhabditis elegans*) we encounter about 13 400 and 19 000 genes, respectively. All known popular mechanisms to enlarge the number of proteins from one gene are observed. Finally, let us have a look at the highest evolved life forms, as we wish to see ourselves. Only a couple of years ago, before the completion of the human genome project phase 1, it was widely accepted that we might have about 100 000 genes. The human genome project still does not know the exact answer, but we assume between 20 000 and 40 000 genes for our species, and most scientist agree on a figure of about 25 000. We are left wondering how we manage to be so much more complex than worms with just slightly increased numbers of genes. The answer lies within the increasing complexity on the way from the genome to the proteome (see Table 1.1).

Assuming we have about 30 000 genes, a single individual will have about 200 000 differentially spliced forms of mRNA and roughly the same number of proteins, as identified by identical sequence, over the course of his or her development. Adding all found or presumed common polymorphisms (e.g. different alleles or single-nucleotide polymorphisms) we encounter on the DNA level, we might well speak of twice the number of 400 000 proteins. If we include the PTMs, numbers increase further. It seems a conservative estimate that on average about five posttranslationally modified isoforms exist per protein, leading to about 2 million different proteins that one might consider analysing in a comprehensive proteomic experiment! There are, of course, no methods at hand to do any such experiment at present!

Obviously, not all possible proteins encoded for by the genome will be expressed at all times in a given practical sample. It is safe to assume that a mammalian cell line expresses some 10 000–15 000 genes at any

Table 1.1 Numbers in proteomics. From a fixed (and in humans still only estimated) number of genes, a larger number of mRNA splice variants is generated. The number of proteins is larger than the number of mRNAs due to N-terminal processing, removal of signal peptides and proteolysis. Each protein can carry various PTMs. The most popular analysis method in proteomics performs analyses on the level of tryptic peptides (MS and MS/MS), as peptides are more informative with the instruments/strategies available. Peptides can be chemically modified by PTMs or by one or more of several hundred known chemical modifications. All figures are estimates.

Number of human genes (tentatively)	3×10^4	
Number of mRNAs	$1–2 \times 10^5$	
Number of proteins	$1–2 \times 10^5$	
Number of protein isoforms with differential PTM	2×10^6	
Number of all detectable tryptic peptide (no PTM)	$>1 \times 10^6$	
Number of all detectable tryptic peptides with natural PTM	1×10^7	
Number of all different tryptic peptides including PTMs and artificial chemical modifications	$>3 \times 10^7$	

given time, or slightly less than half the proteome of the species. Tissues consist of several cell types (plus blood cells, arteries, lymph nodes, etc.) and have a larger complexity. Thus we could encounter the products of perhaps 15 000–20 000 genes in a given tissue sample, or about half of the proteome.

Another problem in numbers arises from the dynamic range in which proteins are encountered. Proteins can be expressed from the rare one protein per cell up to several million proteins per cell (Futcher *et al.*, 1999), whereas there are usually only one or two genes per cell. And of course the Nobel prize winning invention of the polymerase chain reaction allows the amplification of one single molecule of DNA or RNA to any amount needed for repetitive analyses; there is no such thing for proteins. Researchers face the challenge of analysing a small number of proteins (one per cell?) in the presence of very abundant ones (10 million copies per cell; Ghaemmaghami *et al.*, 2003), and it is obviously difficult to quantify any measurements with results ranging over seven orders of magnitude! The most sensitive way to analyze unknown proteins is the use of mass spectrometers, which is another reason why they are so popular in proteomics. Most proteomic approaches can measure peptides down to the low femtomole level, more advances and complex approaches might reach attomole levels, and well characterized proteins can be detected down to the zeptomole level.

1.2.3 Where do proteins hang out?

Apart from other parameters, the location of each protein is most important for its function. Good examples are transcription factors, which might be in an inactive conformation in the cytoplasm and have to translocate to the nucleus to get activated (Kawamori, 2006). So to define a proteome functionally we need to know exactly where proteins are . . . very exactly indeed. A protein being inside or outside an organelle makes a difference of about 20 nm in position, for example! The spatial distribution is also regulated within short time scales; as a typical example we can think about growth factor receptors accumulating within minutes of stimulation in degrading vesicles (e.g. epidermal growth factor: Aguilar and Wendland, 2005). These different locations cannot all be addressed equally well; it is, for instance, difficult to compare protein distribution in cells with different polarity (e.g. apical and distal in epithelial cells). Proteins might be located not only outside or inside an organelle (e.g. the nucleus – Figure 1.7), but also inside its membrane(s) or in other sub-cellular structures (e.g. ribosomes, or skeletal components). Most organelles and many sub-cellular structures can be isolated to quite high purity to analyze the proteins contained in/on them. However, the higher the purity, the longer and more complicated the isolation procedure (usually involving differential centrifugation), and the more time there is for the samples to acquire artefactual changes, as the example from work in our laboratory shows: we label cells radioactively to investigate phosphorylations and a two-hour cellular fractionation procedure allows about 90% of the label to be removed (by phosphatases) when compared to a direct lysis of whole cells in high concentration urea sample buffer. Other possible artefacts include proteolysis or deglycosylation. Together, they can result in proteins dissociating from their 'correct' position. Even without