



Introducing **Proteomics**

From concepts to sample preparation,
mass spectrometry and data analysis

JOSEF UDENIC

JOHNS WILEY & SONS
WILEY-BLACKWELL

Introducing Proteomics

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

This edition first published 2011, © 2011 by John Wiley & Sons, Ltd.

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

The right of the author to be identified as the author of this work has been asserted in accordance with the UK Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

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

To my family, I hope they'll have me back
A naručito naknađujem ovo knjigu za moje roditelje.

Contents

Preface	ix
Acknowledgements	xi
1 Introduction	1
1.1 What Are the Tasks in Proteomics?	1
1.2 Challenges in Proteomics	5
1.3 Proteomics in Relation to Other -omics and System Biology	10
1.4 Some General Applications of Proteomics	12
1.5 Structure of the Book	18
References	18
2 Separation and Detection Technologies	21
2.1 Introduction to Experimental Strategies in Proteomics	21
2.2 Gel-Based Separation	31
2.3 Visualization and Analysis of Proteins/Peptides in Gels	40
2.4 Gel-Free Separation Technologies	54
2.5 Visualization of Proteins/Peptides from Hyphenated Methods	74
2.6 Chips in Proteomic Applications	81
References	81
3 Analysis of Peptides/Proteins by Mass Spectrometry	83
3.1 Basic Principles of Mass Spectrometry for Proteomics	83
3.2 Ionization Methods for Small Amounts of Biomolecules	101
3.3 Mass Analyzers and Mass Spectrometers	116
3.4 Concluding Remarks on Mass Analyzers for Proteomics	170
References	170
4 Analysis and Interpretation of Mass Spectrometric and Proteomic Data	173
4.1 Introduction	173
4.2 Analysis of MS Data	174
4.3 Analysis of MS/MS Data	192
4.4 Quantification of LC MS and MS/MS Data from Complex Samples	209
4.5 Bioinformatic Approaches for Mass Spectrometric Proteome Data Analysis	213
References	218

5 Strategies in Proteomics	221
5.1 Imaging Mass Spectrometry	221
5.2 Qualitative Proteomics	223
5.3 Differential and Quantitative Proteomics	234
5.4 Analysis of Posttranslational Modifications	257
5.5 Interaction Proteomics	261
5.6 Proteomics as Part of Integrated Approaches	266
References	271
Index	275

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.

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

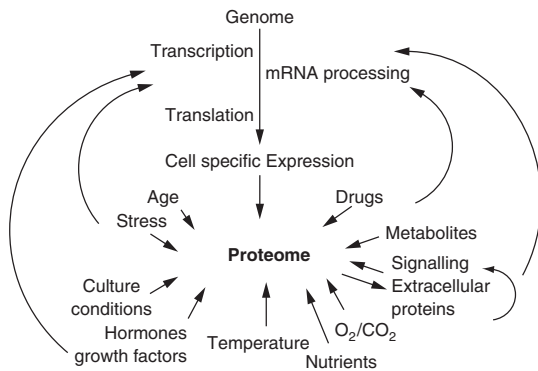


Figure 1.1 Influences on the proteome. The proteome is in a constant state of flux. External factors constantly influence the proteome either directly or via the genome.

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 micro-organism, 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

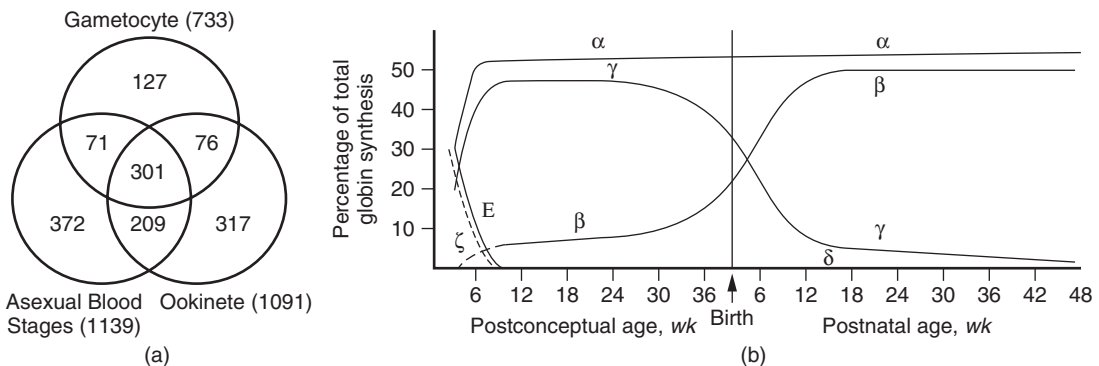


Figure 1.2 The composition of the proteome changes during ontology. (a) *Plasmodium*, the agent causing malaria, has a complex life cycle. Its asexual blood stage cycle lasts about 24 hours, then the sexual stages (gametocytes) develop within 30 hours and develop into the ookinetes after fertilization. A comprehensive proteomic study of these and other stages of the life cycle detected more than 5000 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.

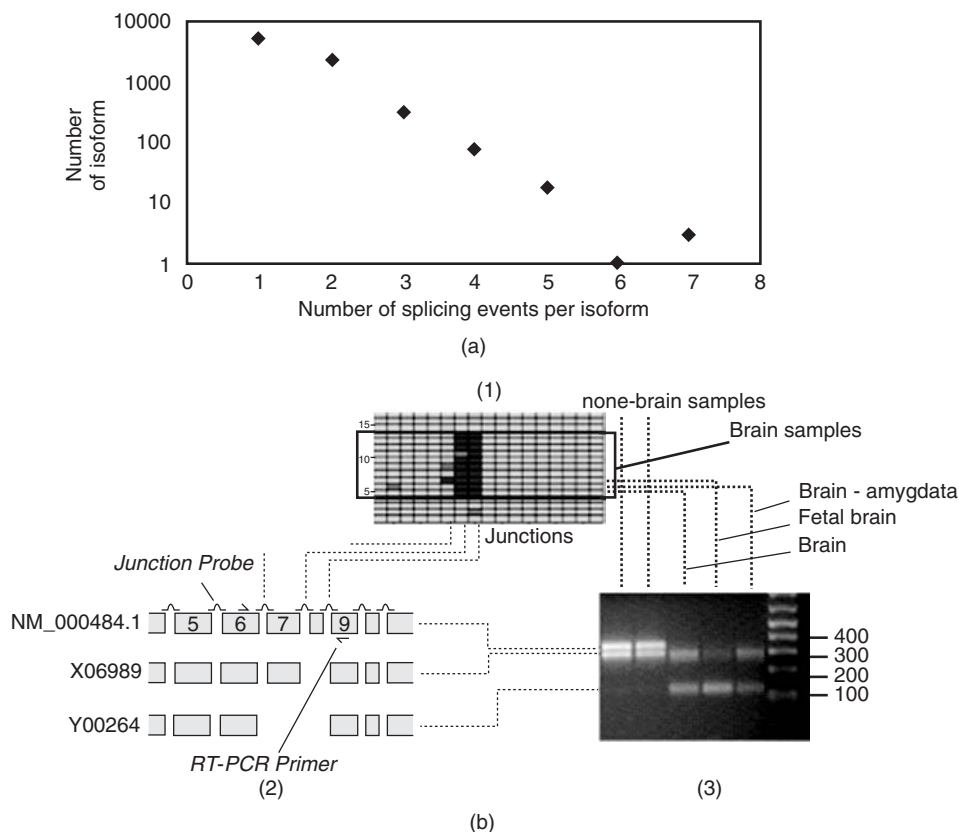


Figure 1.3 The importance of splicing. (a) The known frequency of splicing events for human proteins (Wang *et al.*, 2005). Splicing events were extracted from the SWISS-PROT database, one of the best-annotated databases for proteins. It can be assumed that there are a huge number of non-annotated splicing 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

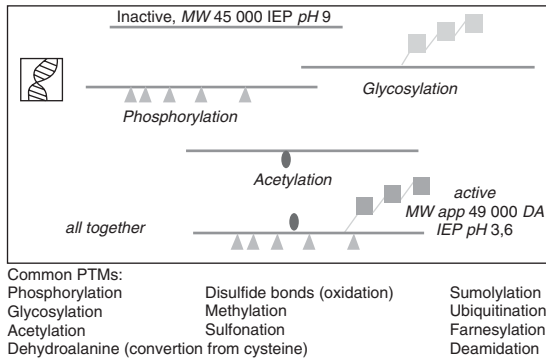


Figure 1.4 Proteins are regulated by posttranslational modifications. Genes and splicing define the primary sequence of proteins. The primary sequence contains motives that allow different PTMs. Which of them are actually found on a protein at any given time in a specific tissue cannot be predicted. 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.

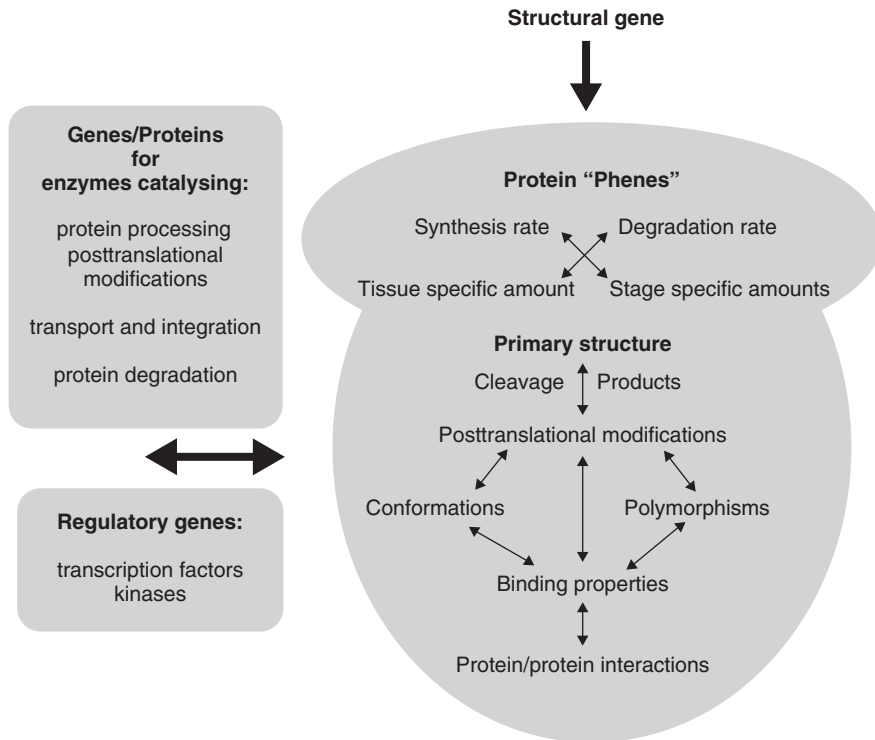


Figure 1.5 Proteins have a 'phenotype'. Similar to whole organisms, proteins can be regarded as having observable traits that are derived by genetic factors as well as influences from the surroundings they experience during their 'life'.

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 sub-cellular 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 sheer 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

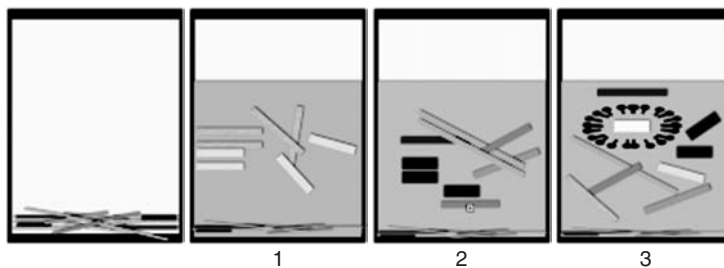


Figure 1.6 Protein solubilization. Complex mixtures of proteins (e.g. cellular lysates) can be solubilized in a variety of buffers (e.g. different ionic strength, pH). Some proteins will dissolve in one or the other buffer, but not in both, while some or most protein interactions are preserved (1/2). 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*

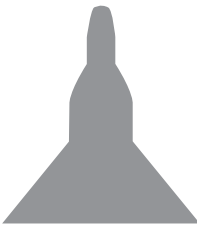
This variety explains how relatively complex organisms can manage to rely on a relative small amount of genes. The least complex forms of life are found among the viruses; in a typical example, a dozen genes will encode about 40 proteins by means of alternative RNA processing and controlled proteolysis. 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

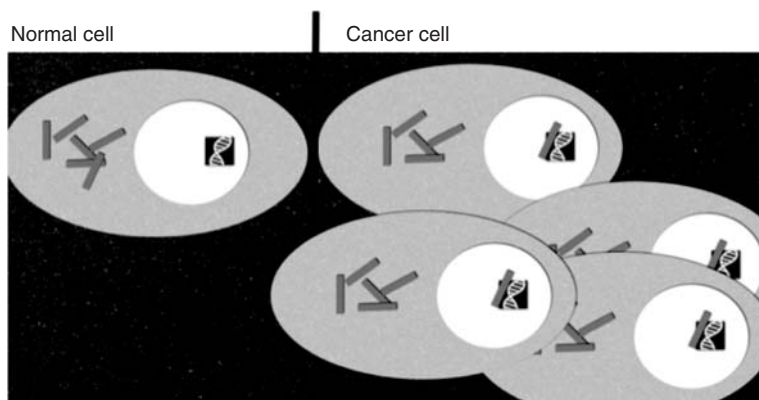


Figure 1.7 Importance of localization in proteomics. The cell in the left panel contains the same amount of red proteins as the cancer cell on the right. However, some of the proteins are in the nucleus, where they can activate transcription and cause cancer. If sub-cellular localization were not analyzed, a quantitative proteomic approach would miss this important difference.

this it is often difficult to judge if a protein is specifically associated with an organelle, or if it just ‘sticks’ non-specifically to the organelle, as a result of the cell lysis and often mediated by artificial associations with other, perhaps denatured, proteins or nucleic acids. On the other

hand, proteins, which are in the living cell (let us call this *in vivo* for our purposes) associated with certain organelles, might get lost during the isolation process. These proteins can be small proteins that ‘leak’ out through artefactual damage in the membranes or pores in the organelles; for

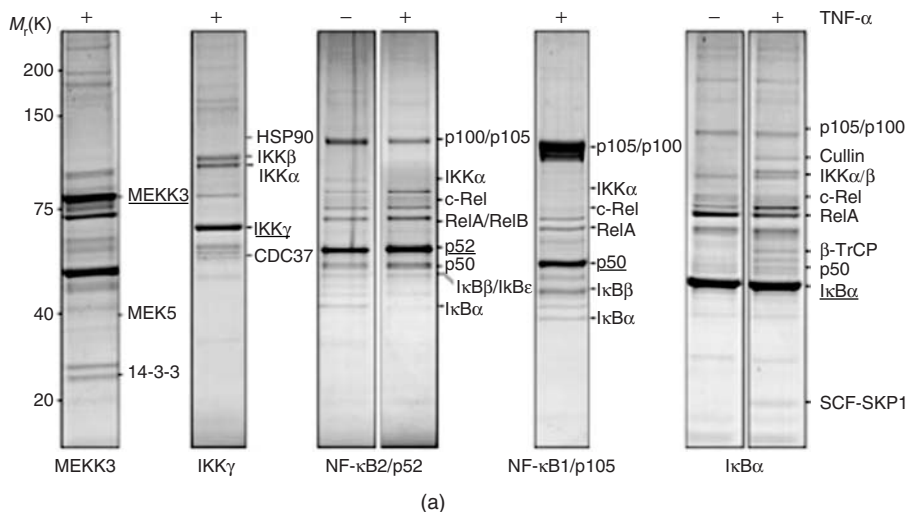


Figure 1.8 Analyzing protein interaction on proteomic levels. To analyze the complex interaction in the human TNF- α /NF- κ B signal transduction pathway selected components were tagged and affinity-purified using a tandem affinity tag approach (see Chapter 5). The affinity tagged proteins (underlined) as well as co-purifying (i.e. physically interacting) proteins were resolved by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE), and unknown protein bands were cut out from gels and identified by liquid chromatography (LC) coupled MS/MS analyses. To cover as many as possible of the interactions some components were ‘knocked down’ from the human cells used for the experiments by RNAi. Parts of the results of hundred such experiments are combined in a database and presented graphically (b). Presentations follow internationally agreed rules for easier interpretation. Even with this amount of work, not all the physical interaction of the proteins involved has actually been analyzed. Reproduced from Bouwmeester *et al.* (2004) courtesy of Nature Publishing Group. © 2004 Nature Publishing Group.

example, proteins below 45 kDa can diffuse freely in and out of the nucleus, in addition to any specific mechanism for importing, exporting or retaining them. Another species of proteins that can get lost are weakly associated proteins on the outside of organelles (as opposed to transmembrane proteins or internal proteins); they are held in place by delicate protein–protein interactions, which will be discussed in the next chapter.

Proteomic studies on sub-cellular structures have been very successful in mapping their composition and function and they have been hugely helped by the onset of gel-free proteomic methods such as free flow electrophoresis and especially multidimensional protein identification technology, known as MudPIT (also called shotgun proteomics; see Chapter 3).

1.2.4 *Proteins always hang out with their mates*

No protein can exert its function alone – there always has to be an interaction with another protein. Structural proteins are often found in huge complexes, and even if they only contain one protein their structure and composition are an important functional feature. As an example just think of tubulin in microtubules – it can be found in long microtubules, short fragments and also in combination with other proteins, often regulating its association/dissociation parameters. Enzymes are often activated and/or kept in place by their association with other proteins. They often even have to be assembled in close association with other proteins (chaperones) in order to fold into a functional form, and that is subject to intricate regulations. For example, a specific class of so-called heat shock proteins (proteins that generally stabilize correct protein folding) has to be associated with some fragile kinases in order to keep them active (e.g. Raf/HSP90: Kolch, 2000) and the regulation of this association is signalling and cell cycle dependent (Lovric *et al.*, 1994). It can be so specific that blocking the function of the heat shock protein kills the cells by inactivation the kinases. Other typical interactions are enzymes and their substrates; often even the substrate preference or specificity is regulated by protein interactions (e.g. Jun binding by extracellular signal-regulated kinase). A very good example of this is the KSR1 protein within the MAP kinase module: using the same kinases with different adaptor proteins, different substrates get phosphorylated (Casar *et al.*, 2009). It is impossible to analyze all these interaction on a proteomic scale, but several proteomic studies have added impressively to our understanding

of either the interactive partners and functions of single proteins (Figure 1.8) or whole protein complexes (e.g. ribosome or transcription complexes). However, results from interaction studies are very complex, and it can be difficult to understand their significance. Depending on the methods used, it might be difficult to understand whether, for example a protein shows a weak but specific interaction or a strong but unspecific interaction (e.g. one that does not occur in living cells) and one has to be careful comparing and combining data from different studies, because they might have been derived using different technologies.

1.3 PROTEOMICS IN RELATION TO OTHER -omics AND SYSTEM BIOLOGY

At the moment there are an ever growing number of new -omics coming into being, next to the classical genomics (Figure 1.9). The main ones are transcriptomics, phosphoproteomics, glycomics and metabolomics.

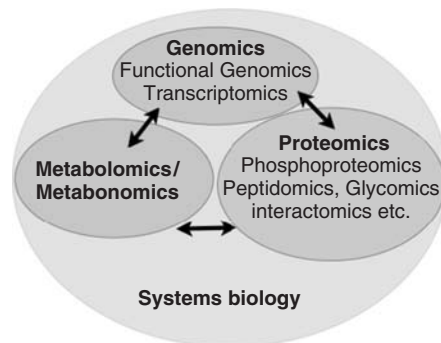


Figure 1.9 The new biology: -omics and systems. Each of the -omics tries to analyze its own sphere of components in a quantitative and qualitative manner (e.g. metabonomics), trying to understand regulatory processes. Related -omics are pharmacogenomics (the study of how genetics affects drug responses) and physiomics (physiological dynamics/functions of whole organisms). Studies in each of the -omics seem troublesome enough, but since the members of all three major -omics are interconnected and influence each other, system biology tries to reach an understanding of the quantitative and qualitative properties of a whole organism or system. An important part of systems biology is the study of how organisms respond to changes (internal or external perturbations) on every level. Mathematical models are often derived to test or expand understanding. Based on findings from the -omics, systems biology depends on rigorous quantitative information (e.g. rate constants of all enzymes, involving signalling kinases, under physiological conditions) to feed its models.

Clearly genomics is a pre-requisite for proteomics. Mass spectrometry is the analytical tool of choice in proteomics, because it is fast, cheap and accurate. However, no one really identifies a protein or a modification by MS, as is always stated; most of the time the mass spectrometer produces data that are highly likely to match the data derived by computer from genomic data. On the other hand, genomic databases can be corrected by data derived from proteomic studies (from mass spectrometers). Proteomic data can discover faults in the genomic database and deliver proof that an inferred gene (and the gene product!) really exists. Going down the information hierarchy, transcriptomics analyses the transcription of DNA into (mainly) mRNA. Transcriptomics derives most of its interest from the assumption that changes in transcript levels are reflected at the functional level, that is, at the level of proteins. Many studies have shown that this is on average not strictly true, as shown in Figure 1.10.

Usually, if an mRNA equilibrium changes, this will be reflected in some sort of change at the protein level; it has to be controlled, however, because of controls on the level of mRNA stability, splicing and translational control. Of course, just because there is more of a protein, that does not necessarily mean it is more active, so transcriptomic studies should really be backed up by proteomic evidence. Combining both technologies, it is also possible in many cases to back up proteomic data and to find the mechanism that led to the changes in protein levels, for example. There are also other reasons why combining proteomics and transcriptomics is beneficial; it is virtually impossible to measure all proteins in proteomics studies as usually the less abundant ones are missed or poorly characterized. Since transcriptomics can be very sensitive, but miss out on several regulation levels, combining technologies has the advantage of increasing coverage of the analyses.

Phosphoproteomics and glycomics are special fields in proteomics; they deserve their names (like other more specialized -omics) since it is impossible with standard proteomic technologies to achieve any reasonable coverage for phosphorylation or glycosylation of proteins. If we estimate that in a typical proteomic approach using a cultured cell line we can analyze about 30–50% of all the different protein species (covering perhaps more than 95% of the total amount of proteins), it is a reasonable estimate that we would be able to analyze maybe around a dozen or so phosphorylated proteins or peptides. Using the best current approaches we still would not be able to

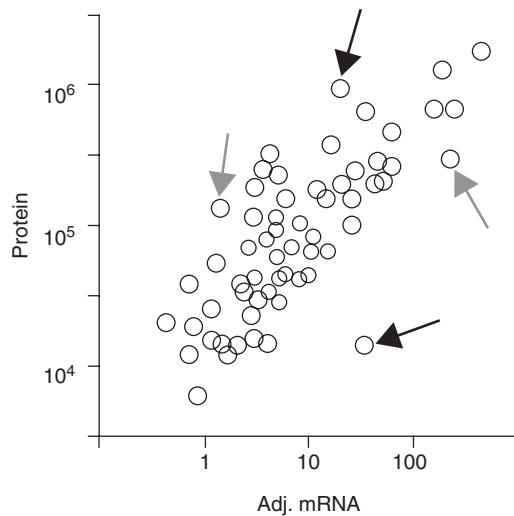


Figure 1.10 Correlation between mRNA and protein levels. Amounts of mRNA and proteins per yeast cell are compared for about 80 genes. Only relatively abundant proteins can be used for this measurement, as reliable data for absolute protein amount are more difficult to obtain for low abundance proteins. On average there are 4000 protein molecules present per mRNA molecule. The correlation coefficient is 0.76. Although this is a good trend, the variation between mRNA and protein amount is on average 10-fold. The grey (black) arrows show that for identical amounts of protein (mRNA) the mRNA level (protein level) can vary about 100-fold (Futcher *et al.*, 1999). Thus it is not possible to reliably predict the amount of protein based on mRNA analyses. Similar relationships between mRNA and protein level can be observed for the mRNA/protein relationship for one gene, when compared in between different tissues in higher organisms. It is prudent to assume that variations are even larger for rare mRNAs or proteins. Adapted from ‘A Sampling of the Yeast Proteome’ B. Futcher, G.I. Latter, P. Monardo, C.S. McLaughlin, and J.I. Garrels, *Mol Cell Biol*, 0270-7306/99/\$04.0010 Nov. 1999, p. 7357–7368, Vol. 19, No. 11 Copyright © 1999, American Society for Microbiology.

detect more than about 2000 phosphorylated peptides or proteins, and we would still not be able to analyze more than perhaps 200 in a quantitative way (i.e. which residues at which ratio are phosphorylated at any given time). If we start with an estimated 10000 different proteins expressed in a certain cell type in a typical experiment, a look at Table 1.2 shows that we would expect some 50000 different phospho-isoforms of these proteins; in other words, our coverage in detection of phospho-isoforms is 4% and far lower in quantitative analysis of phospho-proteins. Surely the analysis of PTMs is a field in which still a lot of further development is needed!

Metabolomics is very different from the -omics discussed so far. It is nearly impossible to link metabolites to single genes directly; they do not encode for metabolites, many different genes are involved in the regulation of each single metabolite, and many metabolites are derived from external sources, like other organisms. Metabolomics has been used very successfully to monitor diseases in newborns and to describe the state of microorganisms. If you look at it from a clinical perspective, screening of metabolites is a very efficient way to screen for dysfunctional genes and proteins. On average, more than 100 genes and their products influence one metabolite. In a typical study about 500 metabolites are controlled – barring redundancies, enough for a potential 50 000 proteins to be controlled! Given the complexity of metabolomics, each combination of metabolite concentrations can be derived from different scenarios on the level of regulation, so it is difficult to find out exactly which dysfunctional enzymes might be responsible for a given metabolic pathology.

This is a good time to have a look at the relatively new field of systems biology. One way to describe systems biology is to say that it is the research field that collects all information available on a system (say, a cell or organism) in order to figure out how the whole system (involving every signalling pathway, every executive pathway, every metabolite) works and is controlled. Since no regulatory circuit is entirely separated from the rest (in fact most seem intensively interconnected) we cannot look at a single pathway; we have to have a look at the whole system, hence the term ‘systems biology’.

An important aspect of systems biology is the aim to simulate a complete system in the computer (*in silico*). For this an enormous amount of data needs to be known; all the enzymes and proteins involved, all concentrations of all metabolites and regulators, all ratios of synthesis, breakdown and half-life for all components, all binding constants and distributions, to name the most important. If a system can be modelled, we can try to unbalance it. If the system reacts like the *in silico* approximation, we might just have a correct understanding of the system. For some systems impressive results have been achieved, from complete imitations of bacterial metabolism to explanations of how signalling pathways in higher organisms regulate differentiation and growth (von Kriegsheim *et al.*, 2009). Only if we can understand cells and organisms in this way will we be able to understand

and cure cancer or metabolic diseases or viral infections. Therefore, in a way, proteomics should be delivering a lot of data to systems biology so that we can understand functional relationships on a truly systemic scale (Figure 1.9).

1.4 SOME GENERAL APPLICATIONS OF PROTEOMICS

Before the term ‘proteomics’ was coined some of its typical technologies were already in use in isolation – for example, the comparison of different maize specimens for their identification and control of variability. To distinguish different variants it is enough to generate a good separation of some marker proteins; using two-dimensional gel electrophoresis, one can usually choose from about 600–2 000 protein ‘spots’ (Figure 1.11). For this kind of analysis it is not even necessary to know why the proteins migrated in different ‘spots’. The spots can arise from different proteins being expressed, or from slight sequence variations of the same (homologue) proteins or from different PTMs on proteins with the same sequence.

Proteomics can also be used for the comparison of species to analyze evolutionary relationships. Humans and chimpanzees are said to be 98.7% identical at the genomic level; when you look at a chimpanzee you would certainly feel (or hope) that the differences are somewhat larger than 1.3%. Genomic studies are very powerful for establishing evolutionary relationships between different strains, species or even higher evolutionary units such as kingdoms. However, at the genomic level the evolution of regulatory differences such as splicing or gene regulation is not very good. Using proteomics, or even organ specific proteomics, this level of evolution can be analyzed. The proteomic study of brain proteins from humans and chimpanzees showed that about 40% of the brain proteins showed either quantitative or qualitative differences (Figure 1.12). This result is a lot more in keeping with our expectations when comparing humans and chimpanzees.

The previous examples showed us the main application of proteomics, the so-called differential proteomics approach. In differential proteomics one is not interested so much in analysing every protein encountered; rather, two sets of proteins are compared, arising from similar but distinct samples. Differential proteomics involves the screening or quantitative/qualitative analysis of as many proteins as possible. However, only a part of these proteins will later be analyzed in any depth, for example to

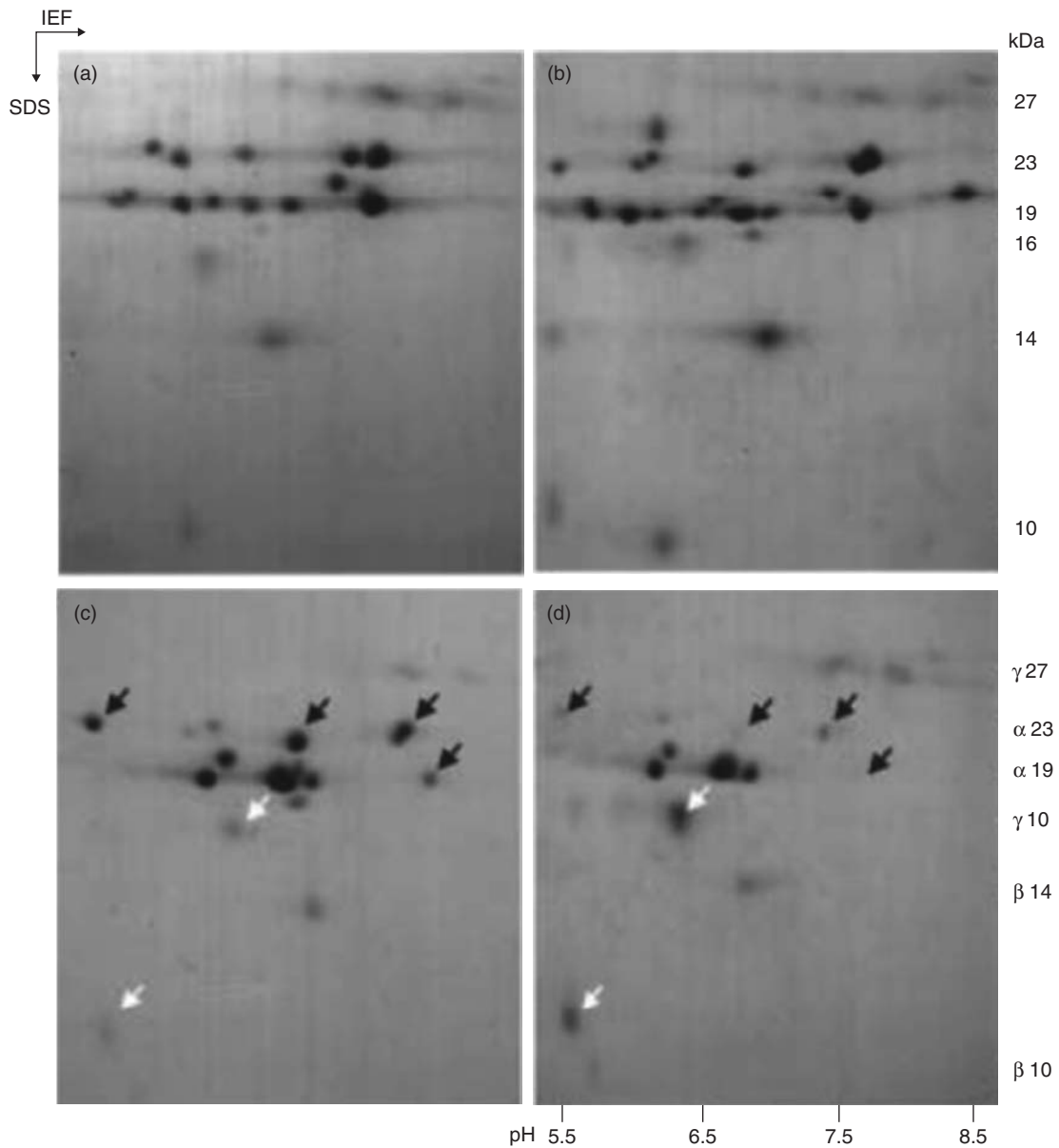
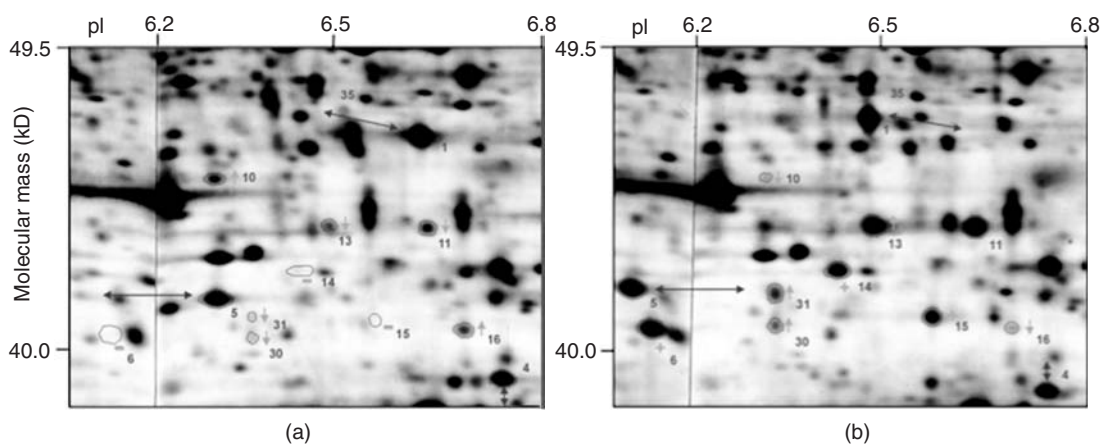


Figure 1.11 Proteomics for the analysis of genetic variability in maize. Several genetic traits influence the quality of maize corns, affecting the group of zein proteins. Zeins are the main proteins in mature seeds; their sequences are not known. A 2D gel electrophoresis of zein proteins isolated from maize powder. Proteins are separated by their isoelectric point horizontally and by their apparent molecular weight vertically. The differences in migration pattern could be based on entirely different amino acid sequences, different modifications, or (most likely) a mixture of both. The zein proteins are affected in the o2 maize line (panel d), with lower quality corns. The arrows in panel d show the zein proteins. Panels a, b and c show other inbred maize strains. The o2 mutant shows increased levels in some zein proteins (white arrows) and diminishing amounts in others (black arrows). By comparing similarities from these gels the variability of 45 inbred maize lines was analyzed in this study, to help and breed the best quality maize lines. Reproduced with permission from Consoli & Damerval (2001). Copyright © 2001 Wiley-VCH Verlag GmbH & Co. KGaA.



Comparison	Analyzed spots	Differences	
		Qualitative	Quantitative
Human-chimpanzee	538	41 (7.6%)	169 (31.4%)
<i>M. musculus</i> - <i>M. spretus</i>	8767	668 (7.6%)	656 (7.5%)

Figure 1.12 Differential proteomics for evolutionary studies. Brain proteins from humans (a) and chimpanzees (b) were separated by 2D SDS PAGE (see legend of Figure 1.11) and gel images were analyzed for qualitative changes (i.e. presence or absence of 'spots', indicated by + or -, or shift in position, indicated by double-headed arrows) or quantitative changes (i.e. more or less of the spot is present, indicated by up or down arrows). Results of repeat experiments were analyzed and are shown in the table below the images. Note that qualitative changes between humans and chimpanzees were as low as changes between two strains of mouse, while quantitative changes were about 4.5 times higher. For this type of analysis the genomes do not need to be known – it is enough to analyze spots on gel images. In (a) and (b) only 200 spots each out of 8 500 spots visible on a large scale gel are shown. From Enard *et al.*, *Science*, 2002; 296: 340–3. Reprinted with permission from AAAS.

identify the gene, analyze PTMs, establish the purity of seeds or distinguish pathological from harmless bacteria (Figure 1.13) – in other words, to identify a biological marker for a pathogen.

However, most differential proteomics studies are designed not only to detect differential proteins but also to identify them or their differential modifications by matching mass spectrometric data to predictions from databases. Typical applications for differential proteomics are the comparison of body fluids or cells or tissues from healthy and diseased states. The diseases range from hydrocephalus to cardiovascular disorders, genetic disorders, dementia (e.g. Alzheimer's) and diverse cancers. When it comes to analysing cancers, differential proteomics is also an important tool for cancer classification. The tissue of origin, the grade of de-differentiation and the level of spread throughout the body are used for classical cancer classification. This is

sometimes helped by limited genetic analysis (e.g. test for chromosomal abnormalities, like losses or translocations of big regions of the chromosomes) or the expression of certain antigens (e.g. specific proteins or glycosylation) known to be tumour or tumour-stage specific. Some cancers have been analyzed by genomic or transcriptomic analysis, and this has delivered a better understanding of their development inside the body. The same holds true for proteomic classifications; instead of having a dozen parameters as in the case of standard classifications, proteomics analyzes thousands of potential tumour markers. This also allows the occurrences of certain changes to be grouped (clustered), when changes in individual proteins are not being very helpful. These new and better classifications are important for choosing different potential treatments and predicting their outcome.

Related to differential proteomics is the field of biomarker discovery. The biggest surge in proteomic

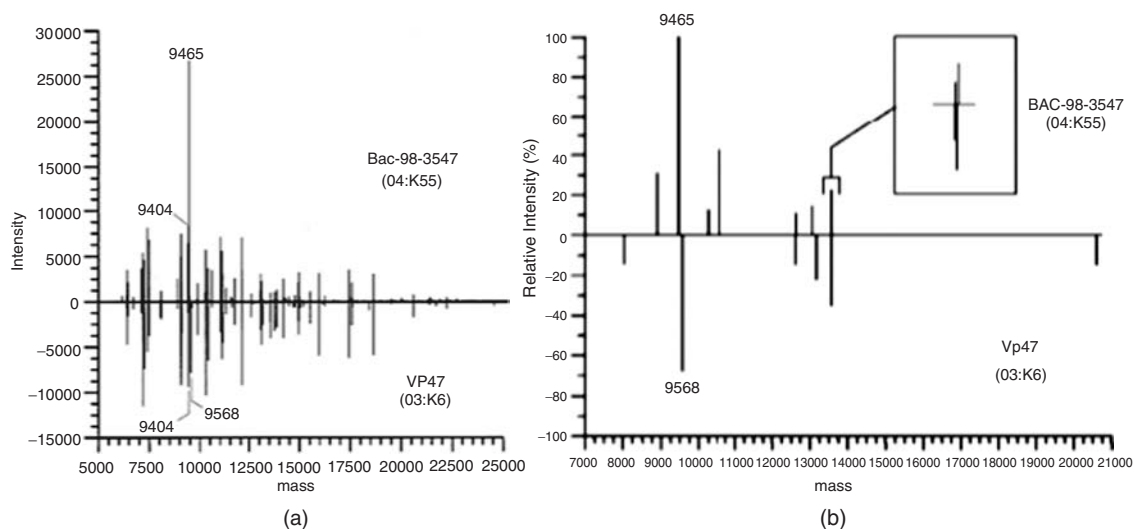


Figure 1.13 Proteomics for the rapid identification of pathogenic bacterial strains. The *Vibrio parahaemolyticus* strain O3:K6 is a dangerous contaminant in sea food, while Q4:K55 is a commonly found harmless strain of the bacterium. To distinguish them, an online LC MS approach was set up. The proteins were separated by LC and the masses of all proteins eluting from the LC were measured by MS. The intensities for both strains were joined for analysis in one graph (a) with O3:K6 derived masses plotted in the negative direction and subtracted from the Q4:K55 derived signals. Thus the graph in (a) shows the differential signal of both strains. A detailed view (b) reveals that the differences in mass are sometimes very low, only 1 Da, reflecting mutations at the protein level (PTMs are rare in bacteria). Additional LC MS and LC MS/MS analyses revealed which proteins are differential and the nature of the mutations. These differences can now be used for very fast and specific detection of this dangerous pathogen in food. Reprinted with permission from Williams *et al.*, *Journal of Clinical Microbiology*, 2004; 42:1657–1665. © 2004 American Society for Microbiology.

applications is the analysis of potential biomarkers using proteomic methods. At the time of writing, thousands of studies have been undertaken to detect biomarkers for diverse conditions, ranging from various cancers, to resistance to chemotherapy, heart conditions, kidney function and the function of the immune system (Figure 1.14). It is very significant that while most studies are very promising, few evaluated biomarkers have yet emerged. On the one hand, this seems to reflect the lack of maturity of proteomic methods, the lack of standardization and thus the problem of world-wide collaboration in studies with statistically significant sample sizes. On the other hand, this lack of immediate results reflects the complexity of the task; we do not know enough about the interactions in complex organisms to understand the huge amount of variability we encounter. It is the aim of this book to contribute to an appreciation of this complexity from the proteomic side of the analysis; the biological samples bring their own complexity as well.

Other fields to which differential proteomics is applied with great success include the study of signalling events and the elucidation of other cellular processes such as DNA replication, transcriptional control, translation, differentiation and the cell cycle. One important feature of proteomics in this setting is that it can analyze the composition of sub-cellular structures with high spatial and temporal resolution. By correlating changes in the composition of structures during biological processes, it is possible to obtain detailed knowledge of the functions of the proteins involved.

Proteomics is regularly used to analyze the reaction of organisms and cells to a changed environment, for example growth under different culture conditions and different food sources or for the analysis of stress response. The stresses analyzed can be very different in nature (e.g. temperature, nutrients, oxygen, osmotic stress, toxins), some of which are very interesting (e.g. during transplantations or more generally for the survival of operations).

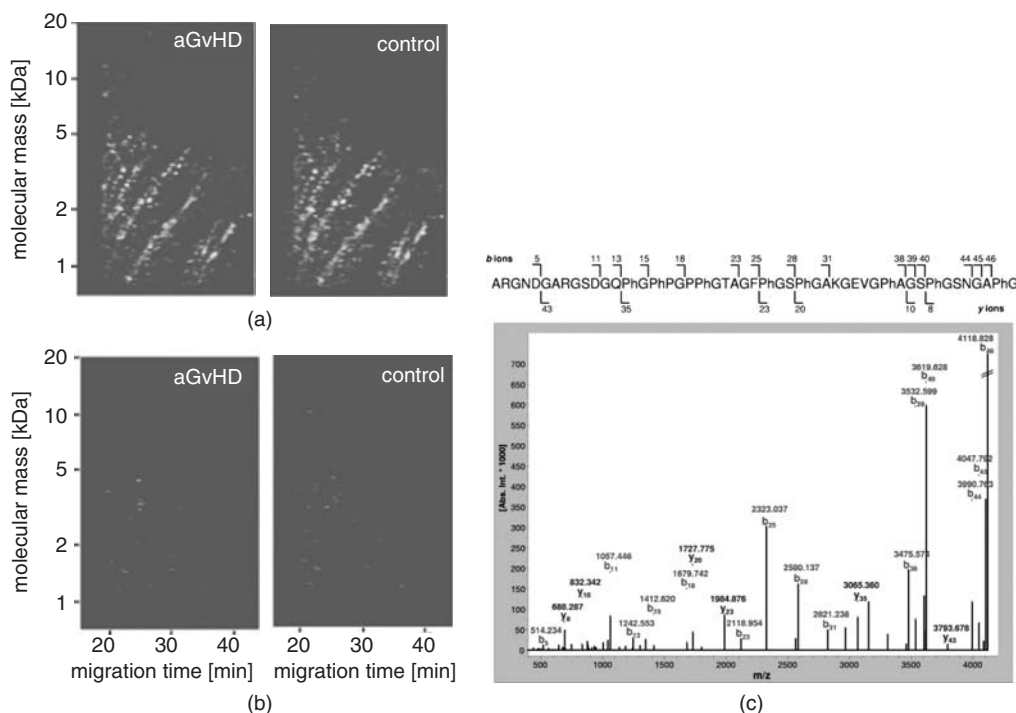


Figure 1.14 Proteomics in biomarker discovery. After treatment for some haematopoietic diseases like cancers, patients undergo stem cell transplantation. The transplanted stem cells can initiate a fatal immune reaction (acute graft versus host disease) against the new host. Biomarkers for the early stages of this disease were found by comparing peptides of serum samples from healthy and diseased individuals. Peptides were separated by capillary electrophoresis coupled with MS, for peptide analysis/identification. In (a) all peptides detected in the samples are shown as white dots, indicating their elution time and mass. In (b) only the differential peptides are shown, after extensive data analysis, and in (c) the identification of one of the diagnostic peptides from its fragmentation pattern in tandem MS is shown. Peptide fragments and their breakpoint inside the deduced peptide sequence are indicated. Reprinted with permission from Weissinger *et al.*, *Blood*, 2007; 109: 5511–19. © 2007 American Society of Hematology.

Using similar approaches, differential proteomics has also been put to good use in pharmacological studies, so a new term, ‘pharmacoproteomics’, has been coined. The main challenges here are to identify the modes of action of drugs, identify new drug targets and evaluate possible toxicities, side effects and resistances. One disease that has been tackled by different proteomic studies is diabetes. Diabetes affects some 200 million people worldwide. It is caused in 90% of cases by decreased pancreatic insulin production or resistance to insulin in the target tissues (e.g. muscle, adipose tissue and liver), where insulin normally induces increased uptake of blood glucose, leading to hyperglycaemia. Different reference maps of 2D gels have been published (e.g. from insulin producing and target tissues), with the aim of helping to understand the effects of anti-diabetic drugs and their side effects.

Differential proteomics is also very helpful in comparing different strains of microorganisms; it delivers more levels of complexity on top of genetics at which homologies and differences can be analyzed. These new complexity levels include the actual expression of similar or identical genes and their pattern of PTM. This is very helpful when it comes to deciding just how close strains of microorganisms are to one another and where the differences derive from. It has been shown that surface enhanced LASER desorption ionization (SELDI), a rapid MALDI MS based technology using an array of different absorbing surfaces for sample preparation, is a fast tool for discriminating different bacterial strains (Barzaghi *et al.*, 2006). Different strains of bacteria can also be analyzed using proteomics, for example to find markers that correlate with different pathologies, as