Translational Bioinformatics Series: 3
Series Editor: Xiangdong Wang, MD, PhD, Prof

Xiangdong Wang *Editor*

Bioinformatics of Human Proteomics

Springer

Translational Bioinformatics

Series Editor

Xiangdong Wang, MD, PhD
Professor of Clinical Bioinformatics, Lund University, Sweden
Professor of Medicine, Fudan University, China

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Translational Bioinformatics

Series Editor: Xiangdong Wang, MD, PhD, Professor of Clinical Bioinformatics,
Lund University, Sweden;
Professor of Medicine, Fudan University, China

Recently Published and Forthcoming Volumes

Applied Computational Genomics
Editor: Yin Yao Shugart
Volume 1

Pediatric Biomedical Informatics
Editor: John Hutton
Volume 2

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

For further volumes:
http://www.springer.com/series/11057

Translational Bioinformatics

Volume 3

Bioinformatics of Human Proteomics

Editor: Xiangdong Wang

Editor
Xiangdong Wang
Medicine, Biomedical Research Center
Fudan University Zhongshan Hospital
Shanghai, China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
ISBN 978-94-007-5810-0 ISBN 978-94-007-5811-7 (eBook)
DOI 10.1007/978-94-007-5811-7
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013930144

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Translational bioinformatics is to bridge clinical application and methodology development and translate the analysis and visualization of complex medical datasets to clinical informatics. The volume *Bioinformatics of Human Proteomics* as one of the serial books entitled *Translational Bioinformatics* more focuses on the application of bioinformatics in human sample-based proteomic studies. The present volume aims to introduce new concepts and methodologies of human proteomics-based bioinformatics and present a number of outstanding studies. We hope clinicians and clinical researchers will find the volume helpful in medical practice, the selection of appropriate software to analyze the protein microarray data for medical decision-making, the development of disease-specific biomarkers, and in drug target identification and clinical validation. With increasing numbers of clinical studies on human tissue-based proteomics, translational bioinformatics integrates omics technology, metabolic and signaling pathways, biomarker discovery and development, computational biology, high-throughput image analysis, human tissue bank, mathematical medicine and biology, protein expression and profiling and systems biology together. Bioinformatics of human proteomics is a critical tool to discover and develop protein-based diagnostics and therapies for diseases.

The serial book *Translational Bioinformatics* is an effort to match disease complexity with patient information, clinical data, standard laboratory evaluations, imaging data and omic data obtained from molecular profiling experiments, in order to improve medical care, patient prognosis, and human health. We created the *Journal of Clinical Bioinformatics* (www.jclinbioinformatics.com) to elucidate how biological and medical information can be applied to the development of personalized healthcare, medication, and therapies, and translate bioinformatics and computational methods into clinical and medical applications as well as the advancement of our understanding of the molecular and cellular mechanisms of diseases. To further accelerate the translational process to clinical practice, new journals *Clinical and Translational Medicine* (www.clintransmed.com) and *Translational Respiratory Medicine* (www.transrespmed.com) were expected to foster a predictive, preventive, personalized, and practical approach toward precision medicine, and exchange

ideas between basic and clinical scientists on molecular and cellular mechanisms of disease and potential therapies, leading to improved patient prognosis.

The first part of the volume starts with the chapter "Clinical Bioinformatics in Human Proteomics Research" to describe the concept and importance of bioinformatics in human proteomics combining clinical informatics, bioinformatics, medical informatics, information technology, mathematics, and addresses clinically relevant challenges in early diagnosis, efficient therapies, and predictive prognosis of patients with disease. The chapter "Clinical Bioinformatics in Human Proteomics Research" explains how to analyze protein-protein interactions, understand the regulation of signal pathways, select different approaches discovering the interaction, and characterize protein complexes. The chapter entitled "Protein Function Microarrays: Design, Use and Bioinformatic Analysis in Cancer Biomarker Discovery and Quantization" describes available methodologies relevant to human proteomics and provides a simple approach to the design and fabrication of cancer antigen arrays suitable for cancer biomarker discovery through serological analysis of cancer patients, including raw data extraction, pre-processing, and analysis. The chapter "Proteomics and Cancer Research" discusses disease-specific biomarkers to detect early-stage cancer, predict prognosis, determine therapy efficacy, identify novel drug targets, and ultimately develop personalized medicine. The chapter "Towards Development of Novel Peptide-Based Cancer Therapeutics: Computational Design and Experimental Evaluation" demonstrates peptides as a novel class of drugs for cancer therapy and discusses three novel bioactive peptide analogues designed using the Resonant Recognition Model.

The second part of the volume focuses more on "Advances in Proteomic Methods" to introduce new technologies for sample processing, protein identification, quantification, structure, and function for the further improvement of accuracy, precision, and reproducibility. The chapter "Clinical and Biomedical Mass Spectrometry: New Frontiers in Drug Developments and Diagnosis" states the importance of protein biomarkers in drug discovery and development, high throughput multiplexed biomarker assay in clinical health care and targeted medicine, and new biomarkers in the early identification of disease and disease progression. The chapter "Disease Biomarkers: Modeling MR Spectroscopy and Clinical Applications" introduces the reference technique for evaluating the metabolism of different tissues in vivo, with special application to brain and prostate lesion characterization and tumor follow-up. Such technology with spectroscopic imaging, higher spatial resolution, lower acquisition times, and the automation of the spectra processing analysis can be also applied in many degenerative and oncologic diseases. The chapter "Processing of Mass Spectrometry Data in Clinical Applications" provides specific bioinformatic tools to assist researchers in the management of large-scale proteomic data and identify biomarkers for clinical practice with a specific focus on the identification of proteotypic peptides and the classification of proteomic data. The chapter "Bioinformatics Approach for Finding Target Protein in Infectious Disease" elaborates the systems biology approaches for identification of novel drug targets for various infectious diseases and highlights some in silico experiments and effective systems biology strategies.

The third part of the volume calls special attention from clinicians and researchers on biomarkers, network biomarkers, or dynamic network biomarkers and emphasizes the importance of dynamic interactions between proteins. The chapter "Identification of Network Biomarkers for Cancer Diagnosis" explains the reasons that the complexity and heterogeneity of carcinogenesis should be explored and validated from individual marker discovery to a systems-oriented paradigm. Network-based biomarker discovery can be one of new strategies to monitor the efficacy and efficiency of cancer intervention. The chapter "Software Development for Quantitative Proteomics Using Stable Isotope Labeling" demonstrates a technique to identify and validate protein-based biomarkers by table isotope labeling coupled with liquid chromatography and high resolution tandem mass spectrometry. UNiquant as a quantification program is introduced to analyze quantitative proteomics data. Another alternative for "Clinical Translation of Protein Biomarkers Integrated with Bioinformatics" can be achieved by maturely quantitative proteomics methods such as stable isotope labeling by amino acids in cell culture, isobaric peptide tags for relative and absolute quantification, and label-free. Targeted quantitative measurements of selected proteins can be applied for the further validation in a large population of patients. The chapter "Proteomic Approaches for Urine Biomarker Discovery in Bladder Cancer" offers a practical example to analyze urinary protein patterns of bladder cancer with the proteomic approach. The chapter "Antibody Microarray and Multiplexing" highlights statistical methods for data normalization and analysis of antibody-based microarray and the implementation of each technique to the technology and suitability on basis of sample types and experiment designs. Last but not least, the chapter "Proteomics in Anesthesia and Intensive Care Medicine" presents a vision of clinical proteomic application from the clinician.

Bioinformatics in human proteomics becomes more and more important in identification and development of disease-specific biomarkers, to diagnose various phases of diseases, monitor severities of diseases and responses to therapies, and predict prognoses and responses of patients to therapy. The application of bioinformatics in clinical proteomics benefits disease-associated specificity, sensitivity, traceability, stability, repeatability, and reliability. Integration of clinical proteomics with bioinformatics can bridge identification and validation of gene or protein-based biomarkers, network biomarkers, dynamic network biomarkers with human diseases, patient phenotypes, and clinical applications. The book intends to accelerate the discovery and development of human disease-specific biomarkers for the early diagnosis, monitoring, and evaluation of diseases and predictions of responses to therapy.

Xiangdong Wang

Contents

Contributors

A. Alberich-Bayarri, M.D. Department of Radiology, Hospital Quirón Valencia, Valencia, Spain

Nicola Armstrong Cancer Research Program, Garvan Institute of Medical Research, Sydney, Australia

School of Mathematics and Statistics and Prince of Wales Clinical School, University of New South Wales, Sydney, NSW, Australia

Larissa Belov School of Molecular Bioscience, The University of Sydney, Sydney, Australia

Jonathan Blackburn, Ph.D. Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Pietro Brunetti Proteomics and Metabolomics Laboratory, Institute for Biomedical Technologies–National Research Council, Segrate, Milan, Italy

Jiajia Chen, Ph.D. Center for Systems Biology, Soochow University, Suzhou, Jiangsu, China

Luonan Chen, Ph.D. Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences Chinese Academy of Sciences, Shanghai, China

William Chi-Shing Cho Department of Clinical Oncology, Queen Elizabeth Hospital, Kowloon, Hong Kong

Richard I. Christopherson School of Molecular Bioscience, The University of Sydney, Sydney, Australia

Shi-Jian Ding, Ph.D. Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

Mass Spectrometry and Proteomics Core Facility, University of Nebraska Medical Center, Omaha, NE, USA

Chaoqin Du Proteomics Division, BGI-Shenzhen, Yantian District, Shenzhen, China

Jessica Duarte Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

David Erlinge, Ph.D. Department of Cardiology, Lund University, Lund, Sweden

Indira Ghosh, Ph.D. School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

Xin Huang Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

T. Istivan Department of Biotechnology and Environmental Biology, School of Applied Sciences, RMIT University, Bundoora West, VIC, Australia

Health Innovations Research Institute, RMIT University, Melbourne, VIC, Australia

Hemant Ritturaj Kushwaha, Ph.D. Synthetic Biology and Biofuel group, International center for Genetic Engineering and Biotechnology (ICGEB), New Delhi, India

Xianyin Lai, Ph.D. Department of Cellular and Integrative Physiology, Biotechnology Research and Training Center, Indiana University School of Medicine, Indianapolis, IN, USA

Haihao Li Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Shitao Li, Ph.D. Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA

György Marko-Varga, Ph.D. Division of Clinical Protein Science and Imaging, Department of Measurement Technology and Industrial Electrical Engineering, Lund University, Lund, Sweden

1st Department of Surgery, Tokyo Medical University, Tokyo, Japan

Luis Martí-Bonmatí, Ph.D., M.D. Department of Radiology, Hospital Quirón Valencia, Valencia, Spain

Department of Imaging Medicine, University of Valencia, Valencia, Spain

Pier Luigi Mauri Proteomics and Metabolomics Laboratory, Institute for Biomedical Technologies –National Research Council, Segrate, Milan, Italy

Nicola Mulder Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Elena Pirogova, Ph.D. School of Electrical and Computer Engineering, RMIT University, Melbourne, VIC, Australia

Health Innovations Research Institute, RMIT University, Melbourne, VIC, Australia

Melinda Rezeli, Ph.D. Division of Clinical Protein Science and Imaging, Department of Measurement Technology and Industrial Electrical Engineering, Lund University, Lund, Sweden

Jean-Michel Serufuri Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

Bairong Shen, Ph.D. Center for Systems Biology, Soochow University, Suzhou, Jiangsu, China

Dario Di Silvestre Proteomics and Metabolomics Laboratory, Institute for Biomedical Technologies –National Research Council, Segrate, Milan, Italy

Yu-Chang Tyan, Ph.D. Department of Medical Imaging and Radiological Sciences/Center of Excellence for Environmental Medicine/National Sun Yat-Sen University-Kaohsiung Medical University Joint Research Center, Kaohsiung Medical University, Kaohsiung, Taiwan

Ákos Végvári, Ph.D. Division of Clinical Protein Science and Imaging, Department of Measurement Technology and Industrial Electrical Engineering, Lund University, Lund, Sweden

Elena Lopez Villar, Ph.D. Department of Oncohematology of Children, Hospital Universitario Niño Jesús, Madrid, Spain

Xiangdong Wang, M.D., Ph.D. Biomedical Research Center, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Department of Pulmonary, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Clinical Bioinformatics, Clinical Science, Lund University, Lund, Sweden

Duojiao Wu, M.D., Ph.D. Biomedical Research Center, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Xu Yang, M.D., Ph.D. Department of Human Complex Disease Research of Research and Cooperation Division of BGI-Shenzhen, Yantian District, Shenzhen, China

Ming-Hui Yang Department of Chemical and Materials Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan

Juanjuan Zhou Department of Human Complex Disease Research of Research and Cooperation Division of BGI-Shenzhen, Yantian District, Shenzhen, China

Jerry Zhou School of Molecular Bioscience, The University of Sydney, Sydney, Australia

Chapter 1
Clinical Bioinformatics in Human Proteomics Research

Duojiao Wu, Haihao Li, and Xiangdong Wang

Abstract Proteome analysis has rapidly developed in the post-genome era and is now widely accepted as a complementary technology to genetic profiling. The improvement in the technology of both two-dimensional electrophoresis (2-DE) analysis as well as quantitative iTRAQ has made proteomics a valuable and powerful tool to study human diseases. Clinical bioinformatics, emerging science combining clinical informatics, bioinformatics, medical informatics, information technology, mathematics, and omics science together, can be considered to be one of critical elements addressing clinical relevant challenges in early diagnosis, efficient therapies, and predictive prognosis of patients with disease. A combination of proteome

D. Wu, M.D., Ph.D.
Biomedical Research Center, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China
e-mail: wuduojiao@126.com

H. Li
Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

X. Wang, M.D., Ph.D. (✉)
Biomedical Research Center, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Shanghai Key Lab of Organ Transplantation, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Department of Pulmonary, Zhongshan Hospital, Shanghai School of Medicine, Fudan University, Shanghai, China

Clinical Bioinformatics, Clinical Science, Lund University, Lund, Sweden
e-mail: xiangdong.wang@clintransmed.org

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_1,

analysis with clinical bioinformatics has been developed as a promising experimental approach for the identification of diagnostic and prognostic markers and so on, suggesting that proteome-based analysis is a promising tool for the identification of prognostic and diagnostic markers as well as for novel therapeutic targets which could be used for the treatment of diseases. The integration of proteome-based approaches with data from genomic or genetic profiling will lead to a better understanding of different diseases, which will then contribute to the direct translation of the research findings into clinical practice.

Keywords Clinical bioinformatics • Proteomics • Biomarker • Network

1.1 Introduction

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. There are some ways of modeling a biological system, such as sequence analysis, literature analysis, and also analysis of gene expression, regulation, protein expression, mutations, and modeling biological systems, high-throughput image analysis, structural bioinformatics approaches. Modeling biological systems is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, and visualization and communication tools with the goal of computer modeling of biological systems. It involves the use of computer simulations of biological systems, like cellular subsystems, to both analyze and visualize the complex connections of these cellular processes (Bonneau 2008).

High-throughput image analysis is used to accelerate or fully automate the processing, quantification, and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images by improving accuracy, objectivity, or speed (Nicholson et al. 2007).

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence. It is one of the most important goals pursued by bioinformatics and theoretical chemistry and is highly important in the territory of medicine and biotechnology. Basing on the notion of homology, we can determine which parts of a protein are important in structure formation and interaction with other proteins. The technique called homology modeling may currently remain the only way to predict protein structures reliably (Zhang and Skolnick 2005).

Another important aspect is molecular interaction. Understanding protein–protein interactions is important for the investigation of intracellular signaling pathways, modeling of protein complex structures, and gaining insights into various biochemical processes (Pazos and Valencia 2001). There are plenty of successful stories. Biomarker is a typical successful application. Biomarker is defined

by FDA as "A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." According to the definition of proteome, we must have the idea that the structure and function of each protein and the complexities of protein–protein interactions are important and critical for our clinical application.

1.2 Clinical Bioinformatics

Clinical bioinformatics provides biological and medical information to allow for individualized healthcare. In clinical bioinformatics, the proteomic data only have meaning if they are integrated with clinical data (Matharoo-Ball et al. 2007). In pharmacogenomics, clinical bioinformatics includes elaborate studies of bioinformatics tools and various facets of proteomics related to drug target identification and clinical validation. Using clinical bioinformatics, researchers apply computational and high-throughput experimental techniques to cancer research and systems biology. Meanwhile, researchers of bioinformatics and medical information have incorporated clinical bioinformatics to improve healthcare, using biological and medical information. Using the high volume of biological information from clinical bioinformatics will contribute to changes in practice standards in the healthcare system. Clinical bioinformatics provides benefits of improving healthcare, disease prevention, and health maintenance as we move toward the era of personalized medicine (Chang 2005).

Proteome data should be furthermore correlated with clinical informatics, including patient complaints, history, therapies, clinical symptoms and signs, physician's examinations, biochemical analyses, imaging profiles, pathologies, and other measurements, as explained in Fig. 1.1. Clinical bioinformatics, emerging science combining clinical informatics, bioinformatics, medical informatics, information technology, mathematics, and omics science together (Wang and Liotta 2011), can be considered to be one of critical elements addressing clinical relevant challenges in early diagnosis, efficient therapies, and predictive prognosis of patients with disease.

In our previous paper (Chen et al. 2012), we developed a new protocol of specific biomarker evaluation by integrating proteomic profiles of inflammatory mediators with clinical informatics in patients with chronic obstructive pulmonary disease (COPD) and acute exacerbations (AECOPD), understanding better their function and signal networks. Forty chemokines were measured using a chemokine multiplex antibody array. Clinical informatics was achieved by a Digital Evaluation Score System (DESS) for assessing severity of patients. Some chemokines such as BTC, IL-9, IL-18Bpa, and CCL22 had significant correlation with DESS scores. Our preliminary study suggested that integration of proteomics with clinical informatics can be a new way to validate and optimize disease-special biomarkers, as shown in Fig. 1.2. The DESS scoring system is provided in Table 1.1.

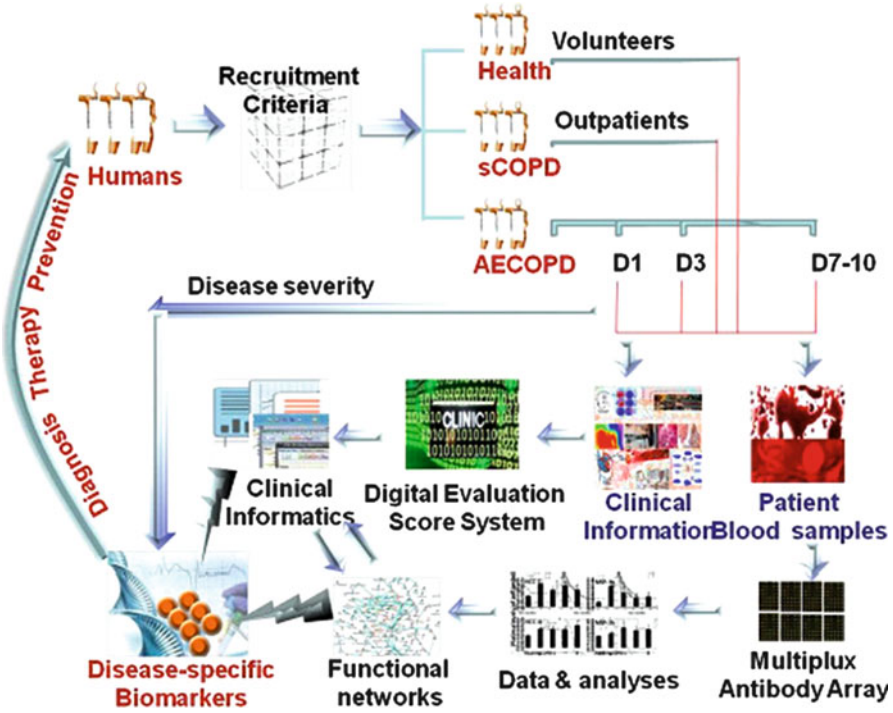


Fig. 1.1 Workflow used for integrating proteomics with clinical informatics to validate and optimize *AECOPD*-special biomarkers (Seng et al. 2009)

Systems clinical medicine is a new and important concept which integrates systems biology, omics science, or bioinformatics with clinical measurements and provides a three-dimensional imaging of protein–protein interactions to demonstrate the location and time of altered proteins, interactions, or regulations in the network. Dynamic network biomarkers show not only higher or lower expression of genes or proteins but also time-dependent stronger or weaker interactions between genes and proteins. Systems clinical medicine more focuses on the bifurcation of gene or protein interactions and the early change of network biomarkers in diseases. Dynamic network biomarkers have the advantage of demonstrating pathophysiological changes at different stages and periods. One of the examples in the field of systems clinical medicine was that the disease specificity of dynamic network biomarkers can be validated by the integration with clinical informatics by which the clinical descriptive information on complaints, sign, symptoms, biochemical analyses, imaging, and therapies was translated into the digital data (Chang 2005). Comparing dynamic alterations of network biomarkers with clinical informatics may allow us to discover disease-specific, stage-specific, severity-specific, or therapy-sensitive biomarkers.

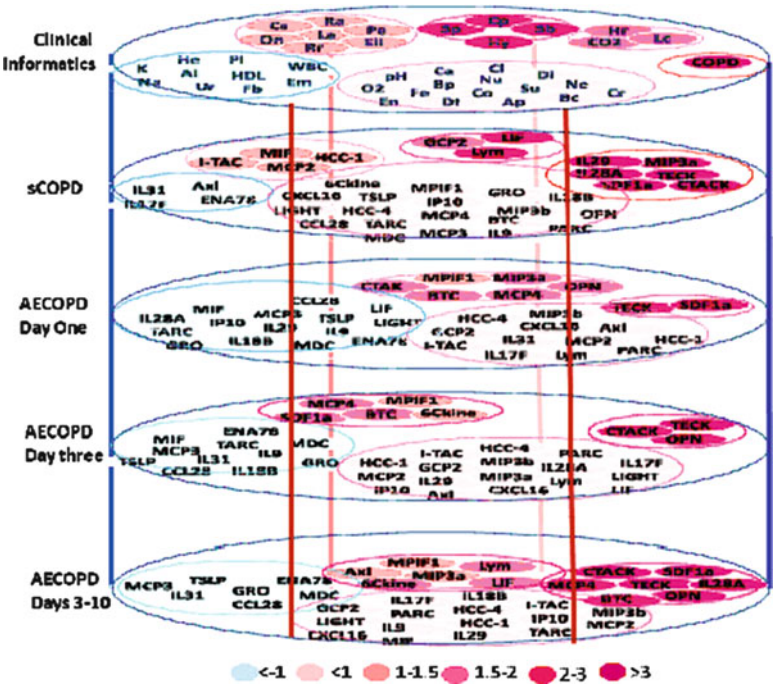


Fig. 1.2 A new protocol of specific biomarker evaluation by combination of proteomic profiles of inflammatory mediators with clinical informatics in patients with acute exacerbations of chronic obstructive pulmonary disease (*AECOPD*). Clinical informatics was achieved by a Digital Evaluation Score System (*DESS*) for assessing severity of patients. Chemokines such as BTC, IL-9, IL-18Bpa, and CCL22 had significant correlation with DESS scores (Seng et al. 2009)

1.3 Methodology of Proteome Research

To study a particular protein, specific antibodies to specific modification can be used to determine the set of proteins that have undergone the modification of interest. In order to develop personalized drugs that are more effective for individual, researchers use these techniques such as ELISA, mass spectrometry, high-pressure liquid chromatography, fluorescence two-dimensional differential gel electrophoresis (2-D DIGE), secretomics, and matrix-assisted laser desorption/ionization (MALDI). Enzyme-linked immunosorbent assay (ELISA) is a popular format of a "wet-lab" type analytic biochemistry assay that uses one subtype of heterogeneous, solid-phase enzyme immunoassay (EIA) to detect the presence of a substance, usually an antigen, in a liquid sample or wet sample. The ELISA has been used as a diagnostic tool in medicine and plant pathology as well as a quality-control check in various industries. Attempting to detect (and quantify) the presence of the antigen in the sample proceeds as follows. Antigens from the sample are attached to a

Table 1.1 Variables and point values used for new score system (history and signs)

	Points			
Variables	0	1	2	4
History				
Cough severity	No	≤1 week	1–2 weeks	≥2 weeks
Sputum	No	White and small amount	White, relatively larger amount	Yellow
Chest pain	No	Under severe activity	Under daily activity	At rest
Shortness of breath	No	Only under severe activity	In daily activity	At rest
Limitation of activity	No	Mild	Marked	Severe
Orthopnea at night	No			Yes
Edema of lower limbs	No			Yes
Chill	No			Yes
Fever (°C)	No	37.3–38	38.1–39	≥39
Duration of fever	No	≤1 week	1–2 weeks	≥2 weeks
Appetite	Good	Semiliquid diet	Liquid diet	Absolute diet
Hemoptysis	No	A little	Median	Large
Stool and urine				
Consciousness	Conscious	Hypersomnia	Confusion	Coma
Hypertension	No	≤5 years	5–10 years	≥10 years
Diabetes mellitus	No	≤5 years	5–10 years	≥10 years
Chronic obstructive pulmonary disease (COPD)	No	≤10 years	10–20 years	≥20 years
Temperature (°C)	<37.3	37.3–38	38.1–39	≥39.1
Heart rate (beat/min)	60–100			>100 or <60 or with any kind of arrhythmia
Respiratory rate (/min)	16–18	19–20 or 12–15	21–24 or 8–11	>24 or <8

Blood pressure (mmHg)	<140/90	Diastolic 140–159 or systolic 90–99	Diastolic 160–179 or systolic 100–109	Diastolic ≥180 or systolic ≥110
Nutrition	Good	Median		Poor or overweight
Enlargement of lymph nodes	No			Yes
Three depression signs	No			Yes
Barrel chest	No			Yes
Chest palpitation	Negative			Positive
Chest percussion	Negative			Positive
Rales	No	Single side <1/3 area	Single side 1/3–1/2 or bilateral <1/3	Single side >1/2 or bilateral >1/3
Heart examination	Negative			Positive
Abdominal examination	Negative			Positive
Laboratory tests				
Hemoglobin (g/L)	Male 120–160 Female 110–150	90 lower limit of normal	60–90	<60
WBC ($\times 10^9$/L)	4月10日	1.5–4	<1.5	>10 or <1.5
Neutrophil percentage (%)	50–70			>70 or <50
Platelet ($\times 10^9$/L)	100–300			>300 or <100
Albumin (g/L)	35–55		28–35	<28
ALT (U/L)	<30			>30
AST (U/L)	<50			>50
ALP (U/L)	Within normal range			Beyond
Gamma-GT	Within normal range			Beyond
Bilirubin (μmol/L)	<34.2	34.2–171	171–342	>342
Urea (mmol/L)	2.5–7.1	7.1–9	9月20日	>20

(continued)

Table 1.1 (continued)

	Points			
Variables	0	1	2	4
Creatinine (μmol/L)	40–120	120–150	150–200	>200
Cholesterol (mmol/L)	3.1–5.9	5.9–7	7月8日	>8
Triglyceride (mmol/L)	0.6–2.0	2.0–3.0	3.0–4.0	>4.0
HDL (mmol/L)	1.03–2.07	0.91–1.03		≤0.91
LDL (mmol/L)	≤3.12	3.12–3.16	3.16–3.64	>3.64
Na (mmol/L)	135–145	146–155 or 125–134	156–165 or 115–124	>165 or <115
K (mmol/L)	3.5–5.5	3–3.4	2.5–2.9	>5.5 or <2.5
Cl (mmol/L)	95–105			<95 or >105
Ca (mmol/L)	2.25–2.58			>2.58 or <2.25
P (mmol/L)	0.97–1.61			>1.61 or <0.97
Glycosylated hemoglobin, HbA1c (%)	4月6日	6月8日	8月9日	>9
pH	7.35–7.45			>7.45 or <7.35
PaO2 (mmHg)	≥90	60–90	40–60	<40
PaCO2 (mmHg)	35–45	45–50		>50
SaO2 (%)	≥90	80–90	60–80	<60
Increased numbers of tumor marker	0	1月3日	4月6日	>6
C-reactive protein, CRP (mg/L)	≤10	10月30日	30–90	>90
Prothrombin time prolonged (s)	0	Within 4 s	4月6日	>6
Fasting blood glucose (mmol/L)	<5.8	5.8–7		>7
Lung imaging				
Lung consolidation	No	Single side <1/3 area	Single side 1/3–1/2	Single side >1/2 or bilateral
Enlargement of lymph nodes	No			Yes
Pleural effusion	No	Single side <1/3 area	Single side 1/3–1/2	Single side >1/2 or bilateral
Emphysema	No			Yes

surface. Then a further specific antibody is applied over the surface so that it can bind to the antigen. This antibody is linked to an enzyme, and, in the final step, a substance containing the enzyme's substrate is added. The subsequent reaction produces a detectable signal, most commonly a color change in the substrate (Lequin 2005).

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio of charged particles (Sparkman 2000). It is used for determining masses of particles, for determining the elemental composition of a sample or molecule, and for elucidating the chemical structures of molecules, such as peptides and other chemical compounds. MS works by ionizing chemical compounds to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios. The technique has both qualitative and quantitative uses. MS is now in very common use in analytical laboratories that study physical, chemical, or biological properties of a great variety of compounds (Thelen and Miernyk 2012). High-pressure liquid chromatography (HPLC) is a chromatographic technique used to separate a mixture of compounds in analytical chemistry and biochemistry with the purpose of identifying, quantifying, and purifying the individual components of the mixture (Xiang et al. 2006).

Two-dimensional difference gel electrophoresis is a technique of gel electrophoresis used to see changes in protein abundance. The proteins from the different sample types are run on the same gel they can be directly compared. To do this with traditional 2-D electrophoresis requires large numbers of time-consuming repeats. 2-D DIGE is an emerging technique for comparative proteomics, which improves the reproducibility and reliability of differential protein expression analysis between samples (Alban et al. 2003). Secretomics is a subset of proteomics in which all of the secreted proteins of a cell, tissue, or organism are analyzed (Hathout 2007). Secreted proteins are involved in a variety of physiological processes, including cell signaling; matrix remodeling, integral to invasion; and metastasis of malignant cells (Pavlou and Diamandis 2010). Secretomics has thus been especially important in the discovery of biomarkers for cancer.

Matrix-assisted laser desorption/ionization is a soft ionization technique used in mass spectrometry, allowing the analysis of biomolecules and large organic molecules, which tend to be fragile and fragment when ionized by more conventional ionization methods. It is similar in character to electrospray ionization both in relative softness and the ions produced (Seng et al. 2009). Systems biology application is widely translated in clinical medicine, such as toward prognosis and therapy response prediction in cancer patients (Logan et al. 2010; Saratsis et al. 2012). Systems clinical medicine could provide insights into molecular mechanisms of disease behavior, helping to develop sensitive prognostic models, identifying novel therapeutic targets, and providing the framework for the development of molecularly based therapies, and, eventually, will help in developing individualized therapy to improve outcomes, with reduced toxicity.

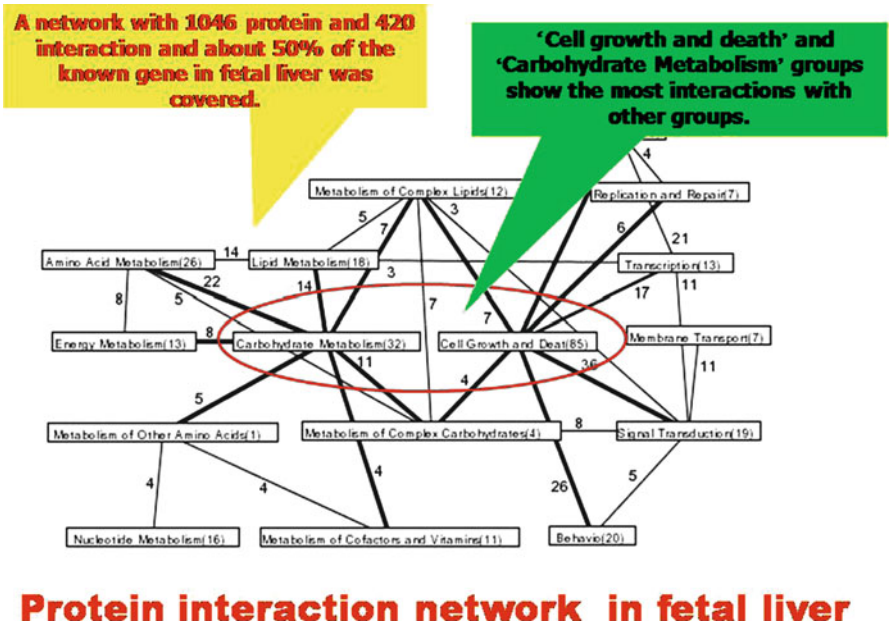


Fig. 1.3 Understanding protein–protein interactions is important for the investigation of intracellular signaling pathways, modeling of protein complex structures, and gaining insights into various biochemical processes

1.4 Application of Proteome Research in Clinical Treatment

Although a systems-level approach to drug design is still very much in its infancy, there are many leading authorities who recognize the potential of systems biology in clinical therapeutics (Zhou and Gallo 2011). The current state of systems pharmacology allows us to formulate a set of questions that could drive future research in the field. Combining pathway and network analyses, pharmacokinetic and pharmacodynamic models, and a knowledge of polymorphisms in the genome will enable the development of predictive models of therapeutic efficacy. The long-term goal of such research is to develop poly-pharmacology for complex diseases and predict therapeutic efficacy and adverse event risk for individuals prior to commencement of therapy (Zhao and Iyengar 2012). Scientists try to set up a methodology of incorporating drug pharmacology information into drug therapeutic response modeling using a computational systems biology approach. A dynamic molecule-to-phenotype model of the human system would allow in drug screening and, thus, more efficient and ethical use of time and resources. It is also important to understand the interaction between proteins rather than only focus on the expression of proteins in diseases, as indicated in Fig. 1.3.

On the other hand, it could provide valuable suggestions to adjust individual drug dosing regimens to improve therapeutic effects considering some agents (immunosuppressive agents, etc.) have wide interindividual pharmacokinetic

variability and a narrow therapeutic index (Shi et al. 2011). A dynamic hybrid systems model is proposed to study drug antitumor effect from the perspective of tumor growth dynamics, specifically the dosing and schedule of the periodic drug intake, and a drug's pharmacokinetics and pharmacodynamics information are linked together in the proposed model using a state–space approach (Li et al. 2012). It is proved analytically that there exists an optimal drug dosage and interval administration point and demonstrated through simulation study.

Systems clinical medicine is also of interest in the field of targeted therapeutics (Dietel and Schafer 2008; Williams et al. 2006). Systems biology can benefit the biomedical sciences by providing a more complete understanding of human disease, enhancing the development of targeted therapeutics (Parisi et al. 2011; Long et al. 2011). In this context, the aim of systems biology is to identify and model key system components (e.g., genes and proteins) that regulate a phenotypic behavior of interest. By measuring these key molecules, the model can be tailored to represent the biology of an individual patient. The personalized model could be used to predict the response to the available treatments and indicate the best therapy on an individual basis (Nicholson 2006). It is particularly relevant in heterogeneous conditions such as neuroblastoma (Teitz et al. 2011); a systems model of neuroblastoma offers benefits in terms of clinical management and therapeutic design for this disease and could also assist developments in related areas of biology and pathology. Also, multiple myeloma (MM) is a complex disease that is driven by numerous genetic and epigenetic alterations. Integrated oncogenomic analyses of human MM have identified candidates of genetic alterations are predicted to be involved in MM pathogenesis and progression. The biological behavior and clinical outcome in MM are dependent on these molecular determinants, which are also attractive therapeutic targets (Munshi and Avet-Loiseau 2011).

Vitucci et al. (2011) studied that gene expression profiling of human glioma to identify molecular subtypes related with the uniquely response to adjuvant therapies. Such progress may lead to a more precise classification system that accurately reflects the cellular, genetic, and molecular basis of disease, a prerequisite for identifying subsets uniquely responsive to specific adjuvant therapies and ultimately in achieving individualized clinical care of patients.

1.5 Prognostic Biomarkers and Risk Stratification

Risk stratification is likely to progress further in the near future, with the increasing availability of omics techniques allowing the routine global molecular analyses of patient-related samples. Qiu et al. (2009) have identified a number of novel biomarkers for prognosis prediction of gastric cancer by using SELDI-TOF-MS combined with sophisticated bioinformatics. By combing Clinical Data Mining Software (Pavlou and Diamandis 2010) and extensive analysis of patient samples, scientists set up systems modeling to confer additional benefit to current risk stratification and management. The term "classifier" refers to a diagnostic that translates one or more biological measurements to a set of predicted categories.

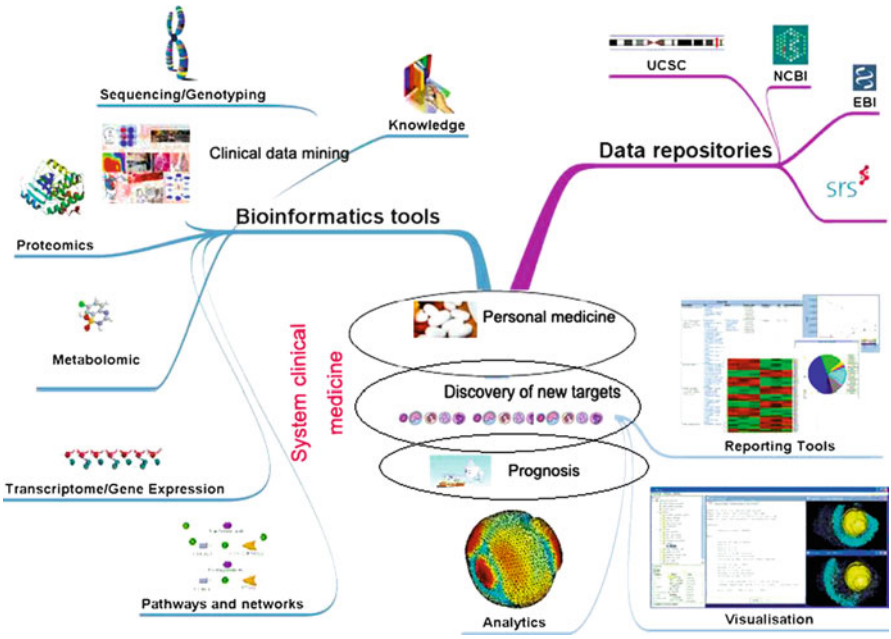


Fig. 1.4 General workflow used for omics-based research in clinical medicine

The process includes identification of the genes or proteins to be included in the classifier, selecting a mathematical way of combining the expression levels of the individual genes or proteins, and training the classifier (i.e., determining the weights and cutoff points) on a training set of data to distinguish responders from nonresponders (Simon 2003). There is substantial literature on the development of gene expression-based classifiers (West et al. 2001; Radmacher et al. 2002). Some designs for phase III clinical trials may facilitate movement to a more predictive oncology (Simon 2008). Because the number of variables (genes, proteins) available is much greater in genomics/proteomics expression profiling than the number of cases available in the training data set, traditional statistical regression model-building strategies are ineffective.

Systems clinical medicine approach to human disease is challenging, but collaborative efforts are facilitating progression toward this ultimate goal. Rapid advances are being made in analytical technologies, bioinformatics, and modeling, resulting in more accurate representation of biology with each iteration of the model development cycle. The integration of proteome-based approaches with data from genomic or genetic profiling will lead to a better understanding of different diseases, which will then contribute to the direct translation of the research findings into clinical practice (Fig. 1.3). Although different "omics data can make important contributions to human systems biology," in order for these data to benefit the wider community, it must be of good quality and properly annotated (Fig. 1.4). Standards are also important for facilitating systems model compatibility in clinical application.

References

Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, Currie I. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics. 2003;3(1):36–44.

Bonneau R. Learning biological networks: from modules to dynamics. Nat Chem Biol. 2008;4(11):658–64.

Chang PL. Clinical bioinformatics. Chang Gung Med J. 2005;28(4):201–11.

Chen H, Song Z, Qian M, Bai C, Wang X. Selection of disease-specific biomarkers by integrating inflammatory mediators with clinical informatics in AECOPD patients: a preliminary study. J Cell Mol Med. 2012;16(6):1286–97.

Dietel M, Schafer R. Systems pathology – or how to solve the complex problem of predictive pathology. Virchows Arch. 2008;453(4):309–12.

Hathout Y. Approaches to the study of the cell secretome. Expert Rev Proteomics. 2007;4(2): 239–48.

Lequin RM. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). Clin Chem. 2005;51(12):2415–8.

Li X, Qian L, Bittner ML, Dougherty ER. A systems biology approach in therapeutic response study for different dosing regimens-a modeling study of drug effects on tumor growth using hybrid systems. Cancer Inform. 2012;11:41–60.

Logan JA, Kelly ME, Ayers D, Shipillis N, Baier G, Day PJ. Systems biology and modeling in neuroblastoma: practicalities and perspectives. Expert Rev Mol Diagn. 2010;10(2):131–45.

Long PM, Stradecki HM, Minturn JE, Wesley UV, Jaworski DM. Differential aminoacylase expression in neuroblastoma. Int J Cancer Journal international du cancer. 2011;129(6):1322–30.

Matharoo-Ball B, Ball G, Rees R. Clinical proteomics: discovery of cancer biomarkers using mass spectrometry and bioinformatics approaches – a prostate cancer perspective. Vaccine. 2007;25 Suppl 2:B110–21.

Munshi NC, Avet-Loiseau H. Genomics in multiple myeloma. Clinical Cancer Res. 2011;17(6): 1234–42.

Nicholson JK. Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol. 2006;2:52.

Nicholson RL, Welch M, Ladlow M, Spring DR. Small-molecule screening: advances in microarraying and cell-imaging technologies. ACS Chem Biol. 2007;2(1):24–30.

Parisi F, Ariyan S, Narayan D, Bacchiocchi A, Hoyt K, Cheng E, Xu F, Li P, Halaban R, Kluger Y. Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. BMC Genomics. 2011;12:230.

Pavlou MP, Diamandis EP. The cancer cell secretome: a good source for discovering biomarkers? J Proteomics. 2010;73(10):1896–906.

Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 2001;14(9):609–14.

Qiu FM, Yu JK, Chen YD, Jin QF, Sui MH, Huang J. Mining novel biomarkers for prognosis of gastric cancer with serum proteomics. J Exp Clin Cancer Res. 2009;28:126.

Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol. 2002;9(3):505–11.

Saratsis AM, Yadavilli S, Magge S, Rood BR, Perez J, Hill DA, Hwang E, Kilburn L, Packer RJ, Nazarian J. Insights into pediatric diffuse intrinsic pontine glioma through proteomic analysis of cerebrospinal fluid. Neuro Oncol. 2012;14(5):547–60.

Seng P, Drancourt M, Gouriet F, La Scola B, Fournier PE, Rolain JM, Raoult D. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis. 2009;49(4):543–51.

Shi XJ, Geng F, Jiao Z, Cui XY, Qiu XY, Zhong MK. Association of ABCB1, CYP3A4*18B and CYP3A5*3 genotypes with the pharmacokinetics of tacrolimus in healthy Chinese subjects: a population pharmacokinetic analysis. J Clin Pharm Ther. 2011;36(5):614–24.

Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. Br J Cancer. 2003;89(9):1599–604.
Simon R. The use of genomics in clinical trial design. Clin Cancer Res. 2008;14(19):5984–93.
Sparkman OD. Review of the 48th ASMS conference on mass spectrometry and allied topics held in Long Beach, California June 11–15, 2000. J Am Soc Mass Spectrom. 2000;11(10):921.
Teitz T, Stanke JJ, Federico S, Bradley CL, Brennan R, Zhang J, Johnson MD, Sedlacik J, Inoue M, Zhang ZM, et al. Preclinical models for neuroblastoma: establishing a baseline for treatment. PLoS One. 2011;6(4):e19133.
Thelen JJ, Miernyk JA. The proteomic future: where mass spectrometry should be taking us. Biochem J. 2012;444(2):169–81.
Vitucci M, Hayes DN, Miller CR. Gene expression profiling of gliomas: merging genomic and histopathological classification for personalised therapy. Br J Cancer. 2011;104(4):545–53.
Wang X, Liotta L. Clinical bioinformatics: a new emerging science. J Clin Bioinformatics. 2011;1(1):1.
West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson Jr JA, Marks JR, Nevins JR. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A. 2001;98(20):11462–7.
Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M, Glanville J, Gray A, Harris A, Johnston K, et al. Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy. Health Technol Assess. 2006;10(34):iii–iv. ix–xi, 1–204.
Xiang Y, Liu Y, Lee ML. Ultrahigh pressure liquid chromatography using elevated temperature. J Chromatogr A. 2006;1104(1–2):198–202.
Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A. 2005;102(4):1029–34.
Zhao S, Iyengar R. Systems pharmacology: network analysis to identify multiscale mechanisms of drug action. Annu Rev Pharmacol Toxicol. 2012;52:505–21.
Zhou Q, Gallo JM. The pharmacokinetic/pharmacodynamic pipeline: translating anticancer drug pharmacology to the clinic. AAPS J. 2011;13(1):111–20.

Xiangdong Wang, M.D., Ph.D., Professor, China and Sweden
Dr. Wang works as a distinguished professor of respiratory medicine at Fudan University, director of Biomedical Research Center, Fudan University Zhongshan Hospital, and adjunct professor of clinical bioinformatics at Lund University, Sweden. His main research is focused on the role of clinical bioinformatics in the development of disease-specific biomarkers and dynamic network biomarkers, the molecular mechanism of organ dysfunction, and the potential therapies.

Dr. Wang was appointed as the principal scientist, global disease advisor, medical monitor and director, and chairman of director board in a number of pharmaceutical companies, for example, Astra Draco, AstraZeneca, PPT, and CatheWill. He worked on pharmacology profiles of target identification and validation, drug screening and optimization, drug PK and PD profile, and translation between discovery and development in areas of respiratory diseases, inflammation, and cancer.

In addition, Dr. Wang serves as the executive vice-president and chairman of executive committee of International Society for Translational Medicine, deputy president of National Professional Society of Insurance and Health in China, senior advisor of Chinese Medical Doctor Association, and director of National

Program of Doctor-Pharmaceutist communication. Dr. Wang acts as an editor in chief of *Journal of Clinical Bioinformatics* and *Journal of Epithelial Biology & Pharmacology* and Asian editor of *Journal of Cellular and Molecular Medicine*. He was also appointed as the president advisor of Wenzhou Medical University, the first hospital adjunct professor of molecular bioscience at North Carolina State University, honor fellow of Romania Academy of Medicine, and visiting professor of Kyoto Prefectural University of Medicine, Japan.

Duojiao Wu, Ph.D., Project Director, China Dr. Wu is a senior researcher and project director in Biomedical Research Center and Shanghai Key Laboratory of Organ Transplantation, Fudan University Zhongshan Hospital. Her research focuses on immunologic rejection and tolerance induction of organ transplantation. She utilizes and integrates the large amount of data generated by genomics and proteomics to describe the complex and dynamic protein interaction networks in immunologic rejection, with the ultimate goal of predicting the behavior of the system. She has published more than 20 peer-reviewed articles as first author and undertook a number of projects as a project manager. She has a Chinese national patent of diagnostic kit of acute rejection in renal transplant.

Chapter 2
Proteomics Defines Protein Interaction Network of Signaling Pathways

Shitao Li

Abstract Protein interactions play fundamental roles in signaling transduction. Analysis of protein–protein interaction (PPI) has contributed numerous insights to the understanding of the regulation of signal pathways. Different approaches have been used to discover PPI and characterize protein complexes. In addition to conventional PPI methods, such as yeast two-hybrid (YTH), affinity purification coupled with mass spectrometry (AP-MS) is emerging as an important and popular tool to unravel protein complex and elucidate protein function through the interaction partners. With the AP-MS method, protein complexes are prepared first by affinity purification directly from cell lysates, followed by characterization of their components by mass spectrometry. In contrast to most PPI methods, AP-MS reflects PPI under near physiological conditions in the relevant organism and cell type. AP-MS is also able to probe dynamic PPI dependent on protein posttranslational modifications, which is common for signal transduction. AP-MS mapping protein interaction network of various signal pathways has dramatically increased in recent years. Here, I'll present the strategies toward obtaining an interactome map of signal pathway and the methodology, detailed protocols, and perspectives of AP-MS.

Keywords Protein interactions • Signaling transduction • Mass spectrometry • Affinity purification • Interaction network • Dynamic

S. Li, Ph.D. (✉)
Department of Microbiology and Immunobiology,
Harvard Medical School, Boston, MA 02115, USA
e-mail: lishitao@hotmail.com

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_2,

2.1 Introduction

Protein interaction plays essential role in cell structure and function. In a simplified diagram of a signaling pathway, upon interaction of a ligand, the receptor alters its conformation, such as dimerization, phosphorylation, and ubiquitination, leading to recruitment of intracellular molecules and subsequent activation of downstream signal cascades. Each level of the signaling cascades requires protein interaction to work as a well-assembled, multifunctional protein complex essential for signal transduction. The functionality of proteins relies on their ability to interact with one another, whereas pathogenic conditions can reflect the perturbations of these protein interactions.

Numerous protein–protein interaction (PPI) methods have been developed, but only a few of them are used for large-scale PPI detection, including yeast two-hybrid (YTH), protein fragment complementation assay (PCA), luciferase-mediated interactome (LUMIER), mammalian protein–protein interaction trap (MAPPIT), protein array, and affinity purification coupled with tandem mass spectrometry (AP-MS). The YTH system is the first assay for analysis of large-scale protein–protein interactions and widely accepted method (Fields and Song 1989). In YTH system, interested gene (bait, X) is fused to the DNA-binding (DB) domain of a transcription factor such as Gal4 (DB-X), while the interacting protein (prey, Y) is fused to an activation domain (AD) such as Gal4-AD (AD-Y). Physical interaction between X and Y brings AD and DB together, which reconstitutes the transcription factor and subsequently activates the downstream reporter genes (Fields and Song 1989). Like the YTH, PCA requires that bait and prey are each fused with incomplete fragments of a third protein, which acts as a reporter. Interaction between bait and prey proteins brings the fragments of reporter protein in close enough proximity to allow them to form a functional reporter protein (Rossi et al. 1997). When fluorescent proteins are reconstituted, the PCA is called bimolecular fluorescence complementation assay (Kerppola 2009). LUMIER is basically a co-immunoprecipitation assay, in which bait is linked to an epitope for purification and prey protein is fused to renilla or firefly luciferase for detection (Barrios-Rodiles et al. 2005). In the MAPPIT, bait and prey proteins are linked to signaling deficient cytokine receptor chimeras. Interaction of bait and prey restores JAK–STAT cascade after the receptor has been stimulated with ligand, which leads to STAT3-dependent reporter gene activation (Eyckerman et al. 2001). Protein microarray is a microscopic array glass slide on which interested proteins have been affixed at separate locations in an ordered manner using a variety of available chemical linkers (MacBeath 2002). Protein microarrays are typically high-density arrays that are used to identify novel proteins or protein–protein interactions. Antibody microarrays are the most common analytical microarray.

AP-MS is biochemical purification of protein complexes followed by characterization of their components by mass spectrometry. However, unlike the methods discussed above, AP-MS is not designed for one-to-one protein interaction (i.e., binary interaction). Instead, AP-MS detects multi-protein complexes. As with

AP-MS, gene of interests is tagged with desirable epitope for affinity purification. Various tags have been developed, such as FLAG tag, HA tag, glutathione S-transferase (GST) tags, the calmodulin-binding peptide, the streptavidin-binding peptide, or the in vivo biotinylation of the target tagged peptide using coexpression of the BirA ligase (Waugh 2005). With affinity tag, protein complexes are enriched first by affinity purification. One early developed AP-MS is to use the tandem affinity purification (TAP) tag (Puig et al. 2001). The original TAP tag is composed of a protein A tag and a calmodulin-binding peptide for two sequential enrichment purifications. In the first purification step, the protein complex is isolated from the cell lysate using immunoglobulin gamma (IgG) resin with high protein A affinity. After protein complex is cleaved from the protein A tag with TEV protease, the eluate undergoes second purification on an immobilized calmodulin column.

To date, AP-MS has been performed in combination with other techniques, such as biochemical fractionation and chemical cross-linking, for characterization of protein complex. Combining biochemical fractionations, like size fractionation, with AP-MS can provide a more precise characterization of multi-protein complexes according to the factions. For example, a combination of TAP purification with standard gel filtration has allowed for a better characterization of RNA polymerase II complex (Mueller and Jaehning 2002). Cross-linker is used for detecting weak interactions, such as membrane complex, which may be interrupted by detergents in lysis buffer. A combination of TAP with in vivo cross-linking with formaldehyde was used to identify novel proteasome interactors (Tagwerker et al. 2006). AP-MS can also be combined with quantitative proteomics approaches, such as SILAC and ICAT, to better understand the dynamics of protein complex assembly. Stable isotope labeling by amino acids in cell culture (SILAC) is an approach for in vivo incorporation of a label into proteins for mass spectrometry (MS)-based quantitative proteomics (Ong et al. 2002). Isotope-coded affinity tags (ICAT) are complementary to SILAC and measure dynamic changes in complexes isolated from tissues or organisms that cannot be metabolically labeled (Gygi et al. 1999). Both entail labeling the samples with isotope labels that allow the mass spectrometer to distinguish between identical proteins in separate samples. Differentially labeled samples are combined and analyzed together, and the differences in the peak intensities of the isotope pairs accurately reflect difference in the abundance of the corresponding proteins.

Given the fundamental importance of protein interactions, systematically mapping protein–protein interaction (PPI) in various species has dramatically increased in recent years. Using high-throughput YTH, proteome-wide physical interaction maps have been generated for several organisms: *Saccharomyces cerevisiae* (Fromont-Racine et al. 1997; Uetz et al. 2000; Ito et al. 2001), *Caenorhabditis elegans* (Walhout et al. 2000; Reboul et al. 2003; Li et al. 2004), *Drosophila melanogaster* (Giot et al. 2003; Guruharsha et al. 2011), and human (Guruharsha et al. 2011; Rual et al. 2005). Virus–host protein interactomes were also explored, such as severe acute respiratory syndrome (SARS)-coronavirus (Pfefferle et al. 2011), Kaposi sarcoma herpesvirus (KSHV), and Varicella zoster virus (VZV) (Uetz

et al. 2006; Rozen et al. 2008). In addition to global mapping, protein interaction networks of several important signal pathways, such as MAPK (Bandyopadhyay et al. 2010), TGFβ (Tewari et al. 2004), SMAD (Colland et al. 2004), and PI3K-mTOR (Pilot-Storck et al. 2010), have been investigated.

In addition to YTH, AP-MS is another widely used PPI tool to map protein interactomes. Due to many advantages that will be discussed later, AP-MS mapping protein interaction network of various signal pathways has dramatically increased in recent years. Global-wide interactomes have been established in *Escherichia coli* (Hu et al. 2009) and *Mycoplasma pneumonia* (Kuhner et al. 2009), *Saccharomyces cerevisiae* (Krogan et al. 2006; Gavin et al. 2006; Ho et al. 2002), *Drosophila melanogaster* (Guruharsha et al. 2011), and HIV–host interactome (Jager et al. 2012). In vertebrate, this approach has so far been used to define proteomic subspaces or specific signal pathways: antiviral innate immunity pathway (Li et al. 2011), autophagy pathway (Behrends et al. 2010), deubiquitinase interactome (Sowa et al. 2009), endoplasmic reticulum-associated protein degradation network (ERAD) (Christianson et al. 2012), TNF pathway (Bouwmeester et al. 2004), proteasome interaction network (Guerrero et al. 2008), and disease-related protein network (Ewing et al. 2007).

Systematic identification of protein interactions within an organism will facilitate systems-level studies of biological processes. Current binary PPI networks are mainly generated by high-throughput yeast two-hybrid. Due to the small overlap of these maps, it has been assumed that these maps are of low quality containing many false positives (Parrish et al. 2006). Recent efforts to map interactions using AP-MS illustrate the promise to measure specific protein interactions in vivo (instead of in yeast) and provide a more powerful tool to model the in vivo interactome. First, I discuss the advantages of AP-MS versus YTH, and then focus the details of the methodology, applications, and perspectives of AP-MS.

2.2 AP-MS Versus YTH

Despite the wide acceptance of YTH system for protein–protein interaction analysis and discovery, high-throughput YTH for protein interaction network bears several major limitations: (1) Reporter analysis method indirectly reflects protein–protein interaction which usually leads to high false positives. For example, proteins with transcriptional activity can lead to autoactivation of the reporter genes. (2) Some heterologous protein expressions are incompatible or toxic to yeast, i.e., membrane proteins which are unlikely to be appropriately assayed as a fusion with a reconstituted transcription factor in YTH. (3) YTH cannot reflect the endogenous protein interactions in the relevant organism. (4) Lots of signaling pathways in vertebrates do not exist in yeast. Thus, interactions triggered by posttranslational modifications do not occur in yeast, resulting in many intrinsic false negatives. (5) The coverage of prey library usually is not completed. In addition, in high-throughput YTH, the bait expression is not monitored. Heterologous full-length protein expression,

Table 2.1 Comparison between YTH and AP-MS

	Cell type	Interaction detection	Interaction type	Interaction level	Interaction status	Cost
YTH	Yeast	Indirect	Binary	Overexpression	Static	Relatively cheap
AP-MS	Relevant species	Direct	Multi-complexes	Endogenous	Static/dynamic	Expensive

especially high-molecular-weight protein, expects to have low expression level in yeast.

Although both YTH and AP-MS detect protein–protein interaction, they have several distinct differences (Table 2.1). AP-MS couples affinity purification with mass spectrometry and requires more labor works and sophisticated equipments. Basically, baits can be expressed in any cell line, which investigator is interested in. After antibiotic selection, bait expression levels are monitored in stable cell lines by western blot, and cell line expressing low bait protein level (close to endogenous level) is usually chosen for following affinity purification. Since the bait expression is close to the counterpart endogenous protein level, we expect the purified complex reflects the endogenous protein interactions under physiological conditions. AP-MS also can be used to detect dynamic protein interactions dependent on protein posttranslational modification by signal stimulation. Unlike YTH detecting one-to-one interaction (aka binary interaction), AP-MS analyzes the entire bait complex and provides all prey information in one run. However, the purified complex represents a mix of direct and indirect binding partners since the nature of the interactions identified in AP-MS data cannot be determined to be either direct or indirect. Last, protein abundance and specificity in different cell lines also limits the detection of protein complex. For example, MIB1 and MIB2 have comparable affinity with TBK1, but we did not detect MIB2 in TBK1 complex in 293T cells by AP-MS. Using real-time PCR, we found MIB1 predominantly expressed in 293T cell line (Li et al. 2011). Taken all together, AP-MS overcomes the limitations of YTH discussed above except several disadvantages over YTH: high cost, indirect interaction, and cell type specificity.

2.3 Methodology of AP-MS

The pipeline of AP-MS from gene construction to interaction network mapping is shown in Fig. 2.1 (Li et al. 2011). In brief, interested gene is tagged with desirable epitopes such as FLAG, GST, His, and biotin. Depending on the purification strategy, one or two tags (usually tandem tags) are adopted. These vectors should carry one antibiotic resistance gene for mammalian cell stable line selection.

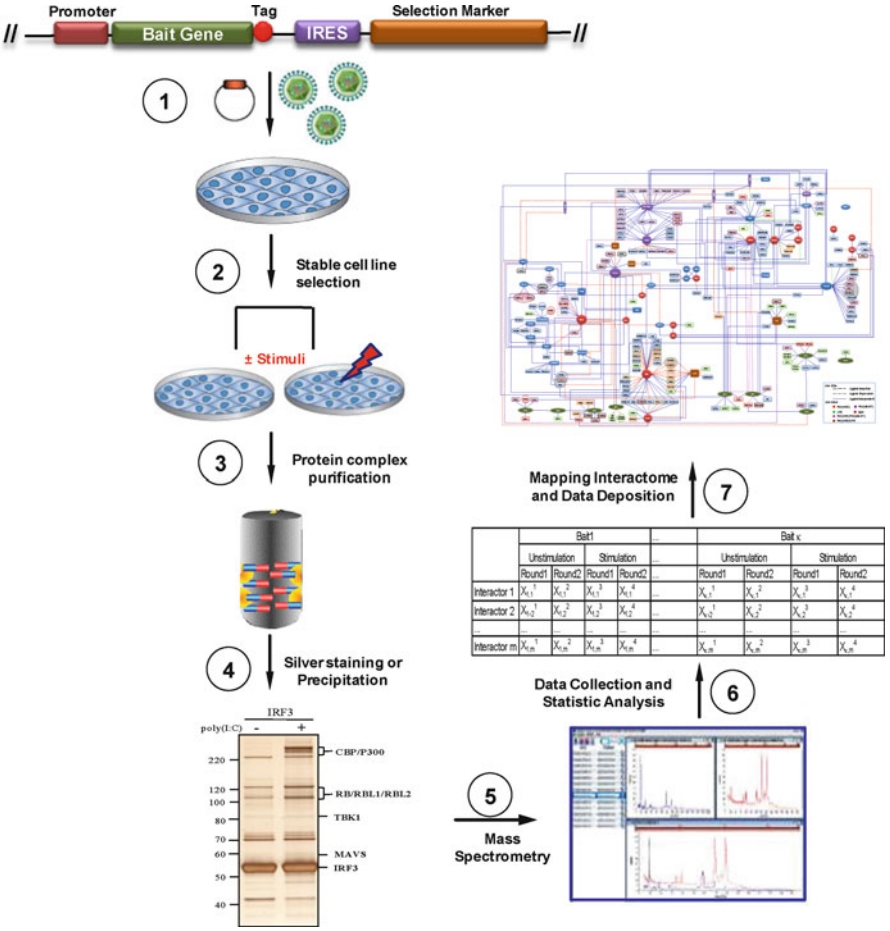


Fig. 2.1 Schematic illustration of the experimental pipeline from gene construction, stable cell line selection, and protein complex affinity purification and identification to data analysis and interactome mapping

After transfection or infection into the desirable mammalian cell line, cells are selected by designated antibiotics to obtain stably and close to endogenous protein expression. Protein complexes are precipitated from lysates of bulk cells by using various immobilized matrixes, such as resin conjugated with antibody. Protein complexes are then eluated from the matrixes after several washing steps to remove nonspecific interactors. Protein complex is either separated on gel following silver staining or precipitated. Sliced gel bands or solution samples are analyzed by mass spectrometry. After data collection and statistical analysis, protein interaction network is generated and ready for validation and further function analysis.

2.3.1 Vector

To purify protein complex closing to physiological level, cell line stably expressing tagged bait is a prerequisite. Therefore, antibiotic resistance gene should be included in the vector for stable cell line selection. Genes of interest also needs to be tagged in-frame with an epitope (at either the N or C terminus), which is used to affinity purify the tagged protein (aka bait) along with its interacting partners (aka prey). Any affinity tag can be used for AP-MS in theory, and most successful tags developed to date are FLAG, HA, S-tag, and tandem affinity purification (TAP) tag. Each purification tag has advantages and disadvantages, and the appropriate technique should be selected depending on the goals of the experiment. For example, a single FLAG or HA epitope only adds 8–11 amino acids (Li et al. 2011), while the TAP tag adds a >20-kDa tag (Krogan et al. 2006) which may cause more nonspecific binding. Because tag may interfere with protein expression or interaction, both N-terminal and C-terminal fusion could be tested for optimal AP-MS. For example, membrane protein may need to put the tag on the C-terminal or after signal peptide on the N-terminus. Furthermore, two kinds of purification methods (single and tandem purification) are used for AP-MS, which requires bait fused with single or double epitopes, respectively.

2.3.2 Purification

Depending on the number of tags on the vector, there are one-step and two-step purification methods for specific protein complex, cell line, or organism. Originally developed for yeast, the first TAP tag consists of calmodulin-binding peptide (CBP), followed by tobacco etch virus protease (TEV protease) cleavage site and protein A with high affinity to immunoglobulin gamma (IgG). Protein complex is first purified from the cell lysate on an IgG affinity resin and cleaved from the protein A tag with TEV protease. The eluate is then enriched in a second affinity purification step on an immobilized calmodulin column. Several variants of TAP with different combinations of tags, such as FLAG-HA double tags, are developed.

Usually, one-step purifications on average preserve weaker or more transient protein–protein interactions in the price of a higher number of nonspecific binding proteins. Conversely, the tandem procedure tends to yield cleaner results, but weak interactions can be lost. FLAG and HA double tags are most commonly applied for tandem purification of protein complexes. We compared the effect of tandem tag versus single tag purification on the yield of total prey and HCIP by examining four protein complexes purified by single purification with FLAG versus a two-step purification with FLAG followed by HA (Li and Dorf 2013). MS analysis revealed that the number of total interactors was dramatically reduced in all protein complexes (TBK1, NAP1, IRF3, and SINTBAD) isolated by TAP purification. However, the ratio of HCIP to total prey did not increase. Consistently, more HCIP were detected

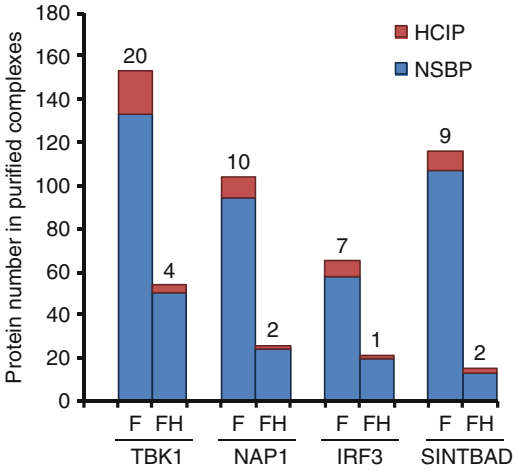


Fig. 2.2 Comparison of one-step and tandem purifications. Protein numbers identified in *TBK1*, *NAP1*, *IRF3*, and *SINTBAD* complexes are depicted. *F* stands for FLAG affinity purification, and *FH* indicates FLAG and HA tandem purification

by single-step affinity purification (Fig. 2.2). In brief, tandem purification reduces the NSBP at the price of HCIP loss. Due to on average more than 90% of proteins as nonspecific binding protein in one-step purification, researchers prefer to tandem affinity purification to get a cleaner background if they only study on a few protein complexes. However, if the study is to map the protein interaction network of a specific signaling pathway, NSBP from one-step purification can be excluded by statistical analysis of the whole database.

2.3.3 Mass Spectrometric Protein Identification

In most proteomics experiments, the purified proteins are separated by one-dimensional SDS-PAGE and stained with a mass spectrometry-compatible dye such as silver, SYPRO ruby, or Coomassie. SDS-PAGE separation removes unwanted contaminants such as buffer components from the protein sample, and the sample complexity is decreased by separating the proteins according to molecular weight. Moreover, it also can be used to compare bands distribution with and without stimulation. In some cases, like IRF3 complexes shown in Fig. 2.1, unique bands are only found in the bait complex with stimulation, indicating these interacting proteins are dependent on ligand stimulation.

Individual protein bands of interest are excised, or the entire lane is cut into approximately 1-mm^3 pieces. Gel pieces were then subjected to an in-gel trypsin digestion procedure to produce peptides for mass spectrometry analysis. But the extraction efficiency of peptides from a gel is low and dependent on the primary structure of the peptide. As an alternative approach to in-gel digestion, protein mixtures can be digested in solution without prior separation (Behrends et al. 2010).

Because buffer components, such as detergents, interfere with the mass spectrometry ionization process, protein samples need to be precipitated with trichloroacetic acid (TCA), washed, and redissolved in a digestion buffer. The main advantages of solution digestion are the reduction of the time and a higher recovery of peptides compared to in-gel digestion. However, bear in mind that some proteins like membrane proteins are resistant to be redissolved.

The peptide mixture can be directly introduced into the mass spectrometer or separated by HPLC before mass spectrometric analysis (LC-MS). The two primary mass spectrometry methods developed for identification of proteins are electrospray ionization (ESI) (Fenn et al. 1989) and matrix-assisted laser desorption/ionization (MALDI) (Hillenkamp et al. 1991). Electrospray ionization mass spectrometry is a desorption ionization method. A sample solution is sprayed from a small tube into a strong electric field in the presence of a flow of warm nitrogen to assist desolvation. The droplets formed evaporate in a region maintained at a vacuum of several torr causing the charge to increase on the droplets. The multiply charged ions then enter the analyzer. The most obvious feature of an ESI spectrum is that the ions carry multiple charges, which reduces their mass-to-charge ratio compared to a singly charged species. This advantage allows mass spectra to be obtained for large molecules. A major disadvantage is that this technique cannot analyze mixtures very well. The other most used technique, MALDI, is a two-step process. First, desorption is triggered by a UV laser beam. Matrix material heavily absorbs UV laser light, leading to the ablation of upper layer (~micron) of the matrix material. A hot plume produced during the ablation contains many species: neutral and ionized matrix molecules, protonated and deprotonated matrix molecules, matrix clusters, and nanodroplets. The second step is ionization (more accurately protonation or deprotonation). In the most common instrumental designs, ESI and MALDI are performed with mass spectrometers capable of tandem mass spectrometry (MS/MS) experiments. Ion traps, quadrupole time-of-flight instruments (Q-TOF), Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers (FTMS), and the Orbitrap are the most common types of instrumentation now used in high-end protein analysis.

2.3.4 Quantification and Dynamics

Most protein interactomes only represent as static entities, which however only poorly captures the dynamics of complex composition. There has been increasing efforts to detect dynamic views of interactomes using various modified AP-MS. Systematic methods to map dynamic changes include semi-quantification based on total spectral counts or ion intensities of precursor peptide (MS1) or fragment ions (MS2) and use of isotopic labeling approaches to obtain more accurate relative quantification. Relative quantification methods such as the stable isotope labeling by amino acids in cell culture (SILAC) detect differences in protein abundance among samples using nonradioactive isotopic labeling. Although relative quantitation

is more costly and time-consuming, and less sensitive to experimental bias than label-free quantitation, it entails labeling the samples with stable isotope labels that allow the mass spectrometer to distinguish between identical proteins in separate samples. Differentially labeled samples are combined and analyzed together, and the differences in the peak intensities of the isotope pairs accurately reflect difference in the abundance of the corresponding proteins. Thus, relative quantitation may discover the dynamic interactions by comparing the change of identical protein abundances from same bait cells with and without extracellular stimulation. Absolute quantitation of proteins is also developed by using isotopic peptides entails spiking known concentrations of synthetic, heavy isotopologues of target peptides into an experimental sample (Mirgorodskaya et al. 2012). However, the cost of absolute quantitation is too high and not realistic for large-scale interactome mapping.

As quantitative methods become more robust, there will be increasing demand for detection of dynamic protein interaction upon extracellular stimulation. For example, we revealed that ~20% protein interactions are dependent on ligand stimulation, such as viral dsRNA mimics poly(dI:dC), in the *h*uman *i*nnate *i*mmunity *i*nteractome for type *I i*nterferon (HI5) (Li et al. 2011). Another example in insulin pathway, Glatter et al. defined the interaction network of insulin receptor/target of rapamycin pathway in Drosophila (Glatter et al. 2011). They found that 22% of the detected interactions were regulated by insulin. In addition to the quantitative power of mass spectrometry, it is also crucial to establish a stable cell line sensitive to stimulations. When overexpressed in cells, bait protein may not respond to stimuli as sensitive as the corresponding endogenous protein.

2.3.5 Data Collection and Analysis

In most cases, the raw data files are first processed by the software controlling the respective mass spectrometry instrument. The generated data sets are then searched against a protein database using search engines such as MASCOT (Hirosawa et al. 1993) or SEQUEST (MacCoss et al. 2002). A valid approach for validation of the chosen parameters is to search the obtained data sets against a decoy protein database. The data also need to be further filtered by setting specific thresholds such as a minimum peptide length or a specific number of peptides to consider a protein identification. Mass spectrometry has some intrinsic problems, such as the common problem of carryovers between mass spectrometry runs. To circumvent the carryover problem in mass spectrometry, we usually analyze the repeated sample in different batch. The carryovers in two independent AP-MS of the same bait will not be possible to show up twice. The record of each batch of MS runs will also help to discriminate the carryovers.

In addition to mass spectrometry, affinity purification also has its own inherent false positives and false negatives, which is critical general limitation encountered in the interpretation of the AP-MS due to lack of binary interaction information. False positives are nonspecific binding proteins and contaminants found in purified bait complex. Several types of false positives are present in typical affinity purified protein samples. The most

common ones are from researchers' hands when they perform purification and handle samples. These contaminants usually are keratin proteins and easy to remove from the dataset. There are also other various kinds of nonspecific binding proteins: (1) proteins binding to affinity matrices, like STK38 and PRMT5; (2) proteins bind to affinity tag, like KIF11 binding to FLAG tag; (3) abundant proteins (e.g., actin, tubulin); (4) proteins prefer binding to specific domain, like ribosomal proteins binding to baits with nucleic acid-binding domain; (5) and heat-shock proteins for protein folding. Therefore, it is important to use cell line stably expressing baits at near physiological levels to avoid NSBPs, as transient overexpression may probably result in protein aggregation and improper intracellular localization. To discriminate NSBP from the protein complex, repetition of AP-MS is mandatory. In our experiences, NSBPs are dramatically different in two independent AP-MS of the same bait. Proper controls including cells expressing GFP with the same epitope will be also useful to exclude NSBPs. Last, large database with the same affinity tag and the same cell line background from high-throughput study will be a good resource for identification of NSBPs and HCIPs. If a protein is often isolated with many unrelated bait proteins, it is easily recognized through analysis of the high-throughput data. However, systematic large-scale experiment does not allow for the subjective and individual evaluation of their results, which means the removal of potential contaminating proteins cannot be based on judging individual purifications. Therefore, statistic tools for analysis of database are required to filter out nonspecific proteins and yield high-confidence interacting proteins.

For statistical analysis of AP-MS data, three main parameters are protein abundance, uniqueness (the frequency of observed protein in database), and reproducibility. Total spectral counts (TSC) have gained acceptance as a practical, label-free, semiquantitative measure of protein abundance in proteomics study. Several computational tools have been developed for the processing of AP-MS data, like CompPASS (Sowa et al. 2009), SAINT (Breitkreutz et al. 2010), and MiST (Jager et al. 2012).

We designed a simplified method for analysis of AP-MS data, combining three main parameters: protein abundance, uniqueness (the frequency of observed protein in the database), and reproducibility. Total spectral counts (TSC) have gained acceptance as a practical, label-free, semiquantitative measure of protein abundance for proteomics studies. We adopted the *z*-score statistic to compare protein abundance because *z*-score calculates the probability of TSC occurring within a normal distribution. However, *z*-score does not reflect reproducibility. In our protocol, each protein complex is tested in 4 MS runs, so reproducibility can be readily factored into the analysis. *z*-Score also does not analyze information about prey occurrence (i.e., prey uniqueness). To explore the likelihood that an interaction is specific, we set a value of prey occurrence at <5%. We now propose a simple 3-stage scoring system to identify HCIP. This algorithm combines *z*-score plus prey occurrence and reproducibility (ZSPORE) (Li and Dorf 2013). In the ZSPORE scoring system, each interaction must pass all three criteria to merit classification as HCIP. The flowchart of ZSPORE is shown as in Fig. 2.3, and a detailed description is provided in Sect. 2.4.6. Taken together, the ZSPORE method combines three parameters (*z*-score based on TSC, prey occurrence, and reproducibility) and is a simple, efficient, and robust way to analyze AP-MS data.

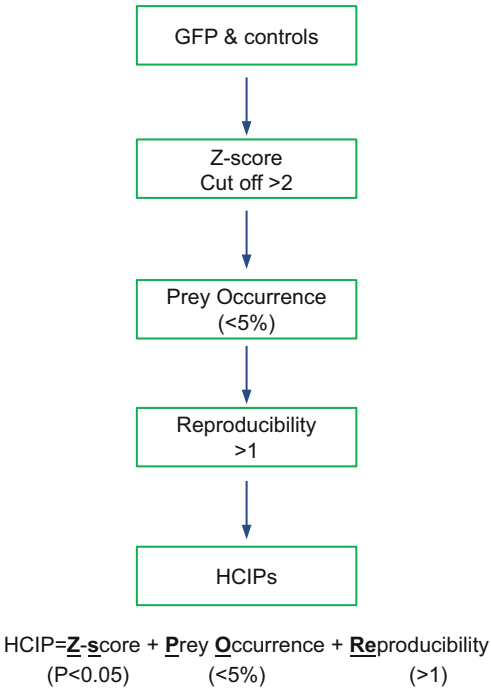


Fig. 2.3 Flowchart of ZSPORE analysis

As with any large screening database, AP-MS also has false negatives, like lacking many known protein–protein interactions documented previously. There are several reasons why a known interaction fail to be found in AP-MS. First, statistical analysis tool may filter out the known interaction as a nonspecific binding. Second, the nature and location of the tag might interfere bait protein function and disrupt its interactions. Third, to parallel comparison, all AP-MS experiments are performed in a same single condition. The generic conditions of affinity purification may be too harsh to preserve some protein interactions, such as the buffer for membrane proteins should be different from other ones. Fourth, the known protein interaction depends on different stimulation. Some proteins may be involved in several pathways and have different interactors in response to the relevant stimulation. Last, the absence of detection is often due to the protein expression level in the specific cell type, especially when the cells have relative low abundances of the protein.

2.3.6 Network Mapping and Analysis

To visualize the protein interaction network formed by HCIPs and baits, graphic representation of two protein interactions basically consists of drawing two circles (nodes) linked by a line (edge). All interactions are combined to generate a map of

the protein interaction network or interactome. A common protein interactome displays a few highly connected nodes forming hubs or subnetwork, while most nodes have a few edges. Several software are developed for graphic mapping protein interaction network. Cytoscape is the most used open-source software platform, which can be used to visualize complex networks derived from AP-MS data. Cytoscape is available for free download at http://www.cytoscape.org. A lot of plug-ins are also available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web (Smoot et al. 2011). For comprehensive and dynamic visualization of the network, various kinds of attributes can be applied to the node and the edge by representation of different color and line thickness.

In addition, the functional classifications of HCIPs can be analyzed by a few online programs. For example, HCIP list can be uploaded to PANTHER (Thomas et al. 2003) or DAVID (da Huang et al. 2009) via a web interface. These programs group these proteins by protein domains, molecular functions, biological processes, and signal pathways. The functional classifications may help discover common threads underlying the proteins of interest. Another approach is to obtain clues from known protein interactions to discover regulation mechanisms. Several protein–protein interaction databases are available for online search, repository, and free download, such as BioGRID, STRING, IntAct, and MINT. The BioGRID database is an online protein interaction repository with data compiled through comprehensive curation efforts. The latest version searches 31,739 publications for 510,188 raw protein and genetic interactions from major model organism species (Stark et al. 2011). The STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations. STRING quantitatively integrates interaction data from these sources for a large number of organisms and transfers information between these organisms where applicable (Szklarczyk et al. 2011). The IntAct database provides a freely available, open-source database system and analysis tools for molecular interaction data (Kerrien et al. 2012). All interactions are derived from literature curation or direct user submissions and are freely available. The MINT database focuses on experimentally verified protein–protein interactions mined from the scientific literature by expert curators (Licata et al. 2012).

AP-MS raw data also can be deposited in the Tranche repository (Smith et al. 2011), which is a distributed file system into which any sort of proteomics data may be uploaded. The data then are distributed on the internet and downloaded by anyone who has access to the hash key identifiers for the data, which may be kept private or publicly released. In summary, all these free online programs are useful and convenient research tools for mapping, analysis, and repository of AP-MS data.

2.4 Protocols

AP-MS has applied for mapping of protein interactome of various cellular signaling pathways in mammalian cells. Our lab has established an efficient AP-MS pipeline for defining protein interaction network and successfully applied in several

pathways including *h*uman *i*nnate *i*mmunity *i*nteractome for type *I* *i*nterferon (HI5) (Li et al. 2011), *mi*RNA pathway *i*nteractome (Mii), and influenza–host (iHost) protein interaction network (Li and Dorf, unpublished data). Detailed pipeline of our AP-MS is provided in this section, and how this applies on different pathways in mammalian cells will be discussed.

2.4.1 Bait Selection and cDNA Cloning

Genes known to regulate the studied signaling pathway are usually selected as primary baits. Baits cover from extracellular signals like ligands to cognate receptors on cell membrane and to signaling intermediates, kinases, and transcription factors involved in these signaling pathways and their family members. After analysis of primary bait AP-MS, some new and important HCIPs with primary baits are also chosen to be as secondary baits. Secondary baits will validate the association with primary baits but also expand the protein interaction network, provide new insights into this signaling pathway, and cross talk with other pathways.

Bait cDNAs can be tagged with various epitopes, such as FLAG or HA epitope. As we discussed earlier, commercially available anti-FLAG beads have much higher affinity than anti-HA beads. We use two mammalian expression vectors, pCMV-3Tag8 (Stratagene) and viral expression vector, pLPCX (Clontech), for transfection and infection, respectively. Vector pCMV-3Tag8 harbors a hygromycin resistance gene, while pLPCX confers cells' resistance to puromycin.

2.4.2 Establishment of Stable Cell Line and Cell Stimulation

Transfection and transduction are two common DNA delivery methods into mammalian cells. For cell lines easy to be transfected like HEK293 cells, bait constructs are directly transfected into cells. For cell lines with low transfection efficiency, such as THP-1 cell line, bait gene needs to be first packaged into retroviral virion. The following infection will allow bait gene to integrate into cell genome DNA and subsequent expression in cells. Two days after transfection and infection, cells are treated with puromycin or hygromycin for 14 days. Single colonies are picked and expanded in 6-well plates. Protein expression levels in each colony are determined by immunoblotting. Colony with protein expression close to endogenous level is picked up for AP-MS.

Most protein interactomes are descriptions of homeostasis of a specific signaling pathway, such as DUB network (Sowa et al. 2009), autophagy interaction network (Behrends et al. 2010), and ERAD interactome (Christianson et al. 2012). However, many protein interactions depend on protein posttranslational modifications induced by different stimuli. For example, we found that about 20% interactions were ligand dependent in HI5 protein interaction network (Li et al. 2011). We also noticed many

new interactions between influenza virus protein and human host after viral infection (Li and Dorf, unpublished data). Therefore, in our pipeline for AP-MS, each stable cell line is divided into two groups, and cells are treated with ligand specific for the signaling pathway or infected with virus for studying virus–host interactome.

2.4.3 Complex Purification

Each group of cells is cultured in four or five 15-cm^2 culture dishes (about 5×10^7 cells) to scale up for affinity purification. Cells are lysed in 10 ml TAP buffer (50 mM Tris HCl [pH 7.5], 10 mM $MgCl_2$, 100 mM NaCl, 0.5% Nonidet P40, 10% glycerol, phosphatase inhibitors, and protease inhibitors). After shaking on ice for 30 min, cell lysates were centrifuged for 30 min at 15,000 rpm. Supernatants are collected and precleared with 50 µl of protein A/G resin. After shaking for 1 h at 4°C, resin is removed by centrifugation. Cell lysates are added to 20 µl anti-FLAG M2 resin (Sigma) and incubated on a shaker for 12 h. Then the anti-FLAG resin is 3× washed (15 min/time) with 10 ml TAP buffer. After removing the wash buffer, the resin is transferred to a spin column (Sigma) and incubated with 40 µl 3× FLAG peptide (Sigma) for 1 h at 4°C in a shaker. Eluates are collected by centrifugation and stored at −80°C.

2.4.4 Silver Staining

Purified complexes are loaded on 4–15% NuPAGE gels (Invitrogen) and run about 1 cm^2 distance for 8 min at 200 V. Gels were stained using the SilverQuest Staining Kit (Invitrogen). Each entire stained lane was excised and rinsed twice with 50% acetonitrile.

2.4.5 Mass Spectrometry

The Taplin Biological Mass Spectrometry Facility (Harvard Medical School) performs MS analysis for our samples. Excised gel bands were cut into approximately 1-mm^3 pieces. Gel pieces are then subjected to a modified in-gel trypsin digestion procedure. Gel pieces were washed and dehydrated with acetonitrile for 10 min followed by removal of acetonitrile. Pieces were then completely dried in a speed-vac. Gel pieces were rehydrated with 50 mM ammonium bicarbonate solution containing 12.5 ng/µl modified sequencing grade trypsin (Promega, Madison, WI) at 4°C. After 45 min, the excess trypsin solution was removed and replaced with 50 mM ammonium bicarbonate solution to just cover the gel pieces. Peptides were later

extracted by removing the ammonium bicarbonate solution, followed by one wash with a solution containing 50% acetonitrile and 1% formic acid. The extracts were then dried in a speed-vac (~1 h) and stored at 4°C until analysis.

On the day of analysis, the samples were reconstituted in 5–10 μl of HPLC solvent A (2.5% acetonitrile, 0.1% formic acid). A nanoscale reverse-phase HPLC capillary column was created by packing 5 μm C18 spherical silica beads into a fused silica capillary (100-μm inner diameter *x* ~12-cm length) with a flame-drawn tip. After equilibrating the column, each sample was loaded via a FAMOS auto sampler (LC Packings, San Francisco, CA) onto the column. A gradient was formed and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile, 0.1% formic acid).

As peptides eluted, they were subjected to electrospray ionization and then entered into an LTQ Velos ion trap mass spectrometer (Thermo Fisher, San Jose, CA). Peptides were detected, isolated, and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide. Dynamic exclusion was enabled such that ions were excluded from reanalysis for 30 s. Peptide sequences (and hence protein identity) were determined by matching protein databases with the acquired fragmentation pattern by the software program SEQUEST (Thermo Fisher, San Jose, CA). The human IPI database (ver. 3.6) was used for searching. Precursor mass tolerance was set to ±2.0 Da, and MS/MS tolerance was set to 1.0 Da. A reversed-sequence database was used to set the false discovery rate at 1%. Filtering was performed using the SEQUEST primary score, Xcorr, and delta-Corr. Spectral matches were further manually examined, and multiple identified peptides (>1) per protein were required.

2.4.6 Statistical Analysis of Mass Spectrometry Data

As with many screening methods, unfiltered AP-MS data contain many nonspecific binding proteins due to some intrinsic characteristics, such as nonspecific binding to bead or tag, protein aggregation, and carryover during MS runs. We now describe a simple efficient statistic method, *z*-score plus prey occurrence and reproducibility (ZSPORE) scoring system, for identification of HCIP. Using this pipeline, we achieve a higher efficiency of AP-MS and better identification of high-confidence interacting proteins. The methods and criteria used to remove nonspecific binding proteins and identify high-confidence interacting proteins include:

(a) GFP and controls. AP-MS of GFP-FLAG and various controls, such as non-FLAG IgG conjugated resin for AP-MS, were used to identify nonspecific binding proteins in the database.
(b) *z*-Score. A *z*-score (aka a standard score) indicates how many standard deviations an element is from the mean. To calculate *z*-score, mass spectrometry data were transformed into a "stats table," where the columns are total spectral counts (TSC) from 4 MS runs, the rows are bait-associated proteins (Table 2.2). Then we calculated *z*-score of each $X_{i,j}$ (i prey interacts with j bait) based on

Table 2.2 Statistical analysis table

	Bait 1				Bait κ			
	Unstimulation		Stimulation		Unstimulation		Stimulation	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
Interactor 1	$X^1_{1,1}$	$X^2_{1,1}$	$X^1_{2,1}$	$X^2_{2,1}$	$X^1_{2\kappa-1,1}$	$X^2_{2\kappa-1,1}$	$X^1_{2\kappa,1}$	$X^2_{2\kappa,1}$
Interactor 2	$X^1_{1,2}$	$X^2_{1,2}$	$X^{s1}_{2,2}$	$X^{s2}_{2,2}$	$X^1_{2\kappa-1,2}$	$X^2_{2\kappa-1,2}$	$X^1_{2\kappa,2}$	$X^2_{2\kappa,2}$
Interactor m	$X^1_{1,m}$	$X^2_{1,m}$	$X^1_{2,m}$	$X^2_{2,m}$	$X^1_{2\kappa-1,m}$	$X^2_{2\kappa-1,m}$	$X^1_{2\kappa,m}$	$X^2_{2\kappa,m}$

the maximum total spectral counts (TSC) of 4 MS runs. For HI5 database analysis, we set the cutoff of z-score as 2.

$$z = \frac{(X - \mu)}{\sigma}$$

z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation.

(c) Prey occurrence. We considered any prey associated with a single bait as an HCIP while preys associated with all baits as NSBP. Generally, we set the bar of prey occurrence as <5%, which means one specific prey interacts less than 5% of total baits in the entire database. In HI5, we showed that preys that interact with less than 5 baits represented statistically significant interactions in HI5 dataset. So the threshold for prey occurrence in HI5 is set as 4. Due to known high interconnectivity among selected baits, bait-to-bait interactions were considered as HCIP.

(d) Reproducibility. Each prey must appear in at least 2 out of 4 MS runs.

(e) Batch reproducibility. To account for possible variations in the list of background contaminants observed in our dataset that were not identified by other statistical approaches, we intentionally sequenced each duplicate purified complex in different experiments. Any protein that did not appear in different purifications was considered an NSBP and manually removed from HCIP list.

2.4.7 Construction of Protein Interaction Map and Bioinformatics Analysis

After statistical analysis of dataset, all pairwise interactions are collected and analyzed by Cytoscape. Several important attributes, such as z-score and TSC, can be integrated into the interaction map. Except generating interaction map, the functional classifications of HCIPs also need to be analyzed. Interactors can be grouped by protein domains, molecular functions, biological processes, and signal

pathways, which may help discover common mechanism underlying the proteins of interest. To figure out the new interactions in database, several protein–protein interaction databases such as BioGRID, STRING, IntAct, and MINT can be used to identify the known interaction. However, protein interactions in new publication will not be included in these databases. The interaction information is also not completed, and many known interactions may not be found in these database. Therefore, it is important to dig out protein interaction information in curated literature. Take together, all AP-MS data must be interpreted with care and validated with additional experiments. As with any screening approach, the database does not represent a final or complete interaction network.

2.5 Perspectives

Understanding how proteins interact in complex and dynamic networks is the key to dissect the complexity of many genotype-to-phenotype relationships. The systematic mapping of physical interactions is therefore critical for post-genomic research. Comprehensive analysis of protein–protein interactions is still a challenging endeavor of functional proteomics. Since intrinsic negatives are inherent to every technique, the physical interaction data generated by AP-MS may carry many false positives and negatives. Thus, AP-MS is unlikely to grasp the entire interactome. It is also still a challenge to develop optimal computational tools to visually and computationally represent the multiple layers of data and integrate existing biological knowledge and functional data in literature with the interactome data. Since most AP-MS data represent static graph of PPI map, advanced methods have to be developed and focused on dynamic and spatial changes in PPI.

We have presented the general principles of the AP-MS approach and highlighted some recent developed technologies and successful applications on various signaling pathways. Despite of the increasing AP-MS data and analysis tools, there are still many major challenges. It includes (1) the specificity of protein complex in different cells and tissues, (2) the dynamics of protein complex with different stimulations or posttranslational modifications, (3) the absolute and relative quantitation of proteins, (4) mapping of transient or weak PPI and endogenous PPI from native cells and tissues, (5) the integration of PPI data sets with the other functional data sets, (6) the standardization and benchmarking for interactome mapping, and (7) the challenges for primary cells like neuronal cells and the detection of weak endogenous interaction. Given the different types of mass spectrometric instrumentation, ionization processes, and software platforms, the assessment of published data becomes increasingly difficult. To facilitate sharing experimental data, common standards in data acquisition, data interpretation, and data storage are required.

Many processes in a cell depend on PPI, and perturbations of these interactions can lead to diseases. Comprehensive knowledge of PPI network of signaling pathways will not only give us insights on how the cells respond to stimulation but will also provide new drug targets for therapeutic application. Moreover, many viral

and bacterial pathogens rely on host PPIs to survive in host cells and tissues and exert their damaging effects. Ultimately, such high-quality PPI networks will become invaluable resources for better understanding the mechanisms underlying major human diseases and will enable the better definition of drug targets.

References

Bandyopadhyay S, Chiang CY, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S, et al. A human MAP kinase interactome. Nat Methods. 2010;7(10):801–5.

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, et al. High-throughput mapping of a dynamic signaling network in mammalian cells. Science. 2005;307(5715):1621–5.

Behrends C, Sowa ME, Gygi SP, Harper JW. Network organization of the human autophagy system. Nature. 2010;466(7302):68–76.

Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S, et al. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat Cell Biol. 2004;6(2):97–105.

Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, Lin ZY, Breitkreutz BJ, Stark C, Liu G, et al. A global protein kinase and phosphatase interaction network in yeast. Science. 2010;328(5981):1043–6.

Christianson JC, Olzmann JA, Shaler TA, Sowa ME, Bennett EJ, Richter CM, Tyler RE, Greenblatt EJ, Harper JW, Kopito RR. Defining human ERAD networks through an integrative mapping strategy. Nat Cell Biol. 2012;14(1):93–105.

Colland F, Jacq X, Trouplin V, Mougin C, Groizeleau C, Hamburger A, Meil A, Wojcik J, Legrain P, Gauthier JM. Functional proteomics mapping of a human signaling pathway. Genome Res. 2004;14(7):1324–32.

da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

Ewing RM, Chu P, Elisma F, Li H, Taylor P, Climie S, McBroom-Cerajewski L, Robinson MD, O'Connor L, Li M, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol Syst Biol. 2007;3:89.

Eyckerman S, Verhee A, der Heyden JV, Lemmens I, Ostade XV, Vandekerckhove J, Tavernier J. Design and application of a cytokine-receptor-based interaction trap. Nat Cell Biol. 2001;3(12):1114–19.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246(4926):64–71.

Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature. 1989;340(6230):245–6.

Fromont-Racine M, Rain JC, Legrain P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet. 1997;16(3):277–82.

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440(7084):631–6.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al. A protein interaction map of *Drosophila melanogaster*. Science. 2003;302(5651):1727–36.

Glatter T, Schittenhelm RB, Rinner O, Roguska K, Wepf A, Junger MA, Kohler K, Jevtov I, Choi H, Schmidt A, et al. Modularity and hormone sensitivity of the *Drosophila melanogaster* insulin receptor/target of rapamycin interaction proteome. Mol Syst Biol. 2011;7:547.

Guerrero C, Milenkovic T, Przulj N, Kaiser P, Huang L. Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. Proc Natl Acad Sci U S A. 2008;105(36):13333–8.

Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, et al. A protein complex network of *Drosophila melanogaster*. Cell. 2011;147(3):690–703.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999;17(10):994–9.

Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem. 1991;63(24):1193A–203A.

Hirosawa M, Hoshida M, Ishikawa M, Toya T. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. Comput Appl Biosci. 1993;9(2):161–7.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature. 2002;415(6868):180–3.

Hu P, Janga SC, Babu M, Diaz-Mejia JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol. 2009;7(4):e96.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A. 2001;98(8):4569–74.

Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, et al. Global landscape of HIV-human protein complexes. Nature. 2012;481(7381):365–70.

Kerppola TK. Visualization of molecular interactions using bimolecular fluorescence complementation analysis: characteristics of protein fragment complementation. Chem Soc Rev. 2009;38(10):2876–86.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40(Database issue):D841–6.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature. 2006;440(7084):637–43.

Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, et al. Proteome organization in a genome-reduced bacterium. Science. 2009;326(5957):1235–40.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, et al. A map of the interactome network of the metazoan *C. elegans*. Science. 2004;303(5657):540–3.

Li ST, Dorf ME. Optimization and ZSPORE analysis of affinity aurification aoupled with tandem mass spectrometry in mammalian cells. J Proteomics Genomics Res. 2013 (in press).

Li S, Wang L, Berman M, Kong YY, Dorf ME. Mapping a dynamic innate immunity protein interaction network regulating type I interferon production. Immunity. 2011;35(3):426–40.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40(Database issue):D857–61.

MacBeath G. Protein microarrays and proteomics. Nat Genet. 2002;32(Suppl):526–32.

MacCoss MJ, Wu CC, Yates 3rd JR. Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal Chem. 2002;74(21):5593–9.

Mirgorodskaya OA, Korner R, Kozmin YP, Roepstorff P. Absolute quantitation of proteins by acid hydrolysis combined with amino acid detection by mass spectrometry. Methods Mol Biol. 2012;828:115–20.

Mueller CL, Jaehning JA. Ctr9, Rtf1, and Leo1 are components of the Paf1/RNA polymerase II complex. Mol Cell Biol. 2002;22(7):1971–80.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1(5):376–86.

Parrish JR, Gulyas KD, Finley Jr RL. Yeast two-hybrid contributions to interactome mapping. Curr Opin Biotechnol. 2006;17(4):387–93.

Pfefferle S, Schopf J, Kogl M, Friedel CC, Muller MA, Carbajo-Lozoya J, Stellberger T, von Dall'Armi E, Herzog P, Kallies S, et al. The SARS-coronavirus-host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors. PLoS Pathog. 2011;7(10):e1002331.

Pilot-Storck F, Chopin E, Rual JF, Baudot A, Dobrokhotov P, Robinson-Rechavi M, Brun C, Cusick ME, Hill DE, Schaeffer L, et al. Interactome mapping of the phosphatidylinositol 3-kinase-mammalian target of rapamycin pathway identifies deformed epidermal autoregulatory factor-1 as a new glycogen synthase kinase-3 interactor. Mol Cell Proteomics. 2010;9(7):1578–93.

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods. 2001;24(3):218–29.

Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, et al. *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet. 2003;34(1):35–41.

Rossi F, Charlton CA, Blau HM. Monitoring protein-protein interactions in intact eukaryotic cells by beta-galactosidase complementation. Proc Natl Acad Sci U S A. 1997;94(16):8405–10.

Rozen R, Sathish N, Li Y, Yuan Y. Virion-wide protein interactions of Kaposi's sarcoma-associated herpesvirus. J Virol. 2008;82(10):4742–50.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005;437(7062):1173–8.

Smith BE, Hill JA, Gjukich MA, Andrews PC. Tranche distributed repository and ProteomeCommons. org. Methods Mol Biol. 2011;696:123–45.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431–2.

Sowa ME, Bennett EJ, Gygi SP, Harper JW. Defining the human deubiquitinating enzyme interaction landscape. Cell. 2009;138(2):389–403.

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011;39(Database issue):D698–704.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39(Database issue):D561–8.

Tagwerker C, Flick K, Cui M, Guerrero C, Dou Y, Auer B, Baldi P, Huang L, Kaiser P. A tandem affinity tag for two-step purification under fully denaturing conditions: application in ubiquitin profiling and protein complex identification combined with in vivo cross-linking. Mol Cell Proteomics. 2006;5(4):737–48.

Tewari M, Hu, PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD, et al. Systematic interactome mapping and genetic perturbation analysis of a C. elegans TGF-β signaling network. Mol Cell. 2004;13:469–82.

Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. 2003;31(1):334–41.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature. 2000;403(6770):623–7.

Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, Berger B, Rajagopala SV, Roupelieva M, Rose D, Fossum E, et al. Herpesviral protein networks and their interaction with the human proteome. Science. 2006;311(5758):239–42.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science. 2000;287(5450):116–22.

Waugh DS. Making the most of affinity tags. Trends Biotechnol. 2005;23(6):316–20.

Shitao Li, Ph.D., USA Shitao Li is a research fellow in Department of Microbiology and Immunobiology, Harvard Medical School. He obtained his Ph.D. from Wuhan University, China. He is a recipient of Kaneb Fellowship and AAI Abstract Award (2010). Dr. Li studies on protein interaction network using proteomics approach. He has mapped a dynamic antiviral innate immunity protein interaction network and currently is working on virus-host protein interaction network. By examining the protein network, he is investigating the signaling mechanisms controlling innate antiviral immunity and new drug targets for host defense to viral infection. He published his research on several prestigious journals such as Nature, Immunity, and Molecular Cell.

Chapter 3
Protein Function Microarrays: Design, Use and Bioinformatic Analysis in Cancer Biomarker Discovery and Quantitation

Jessica Duarte, Jean-Michel Serufuri, Nicola Mulder, and Jonathan Blackburn

Abstract Protein microarrays have many potential applications in the systematic, quantitative analysis of protein function, including in biomarker discovery applications. In this chapter, we review available methodologies relevant to this field and describe a simple approach to the design and fabrication of cancer-antigen arrays suitable for cancer biomarker discovery through serological analysis of cancer patients. We consider general issues that arise in antigen content generation, microarray fabrication and microarray-based assays and provide practical examples of experimental approaches that address these. We then focus on general issues that arise in raw data extraction, raw data preprocessing and analysis of the resultant preprocessed data to determine its biological significance, and we describe computational approaches to address these that enable quantitative assessment of serological protein microarray data. We exemplify this overall approach by reference to the creation of a multiplexed cancer-antigen microarray that contains 100 unique, purified, immobilised antigens in a spatially defined array, and we describe specific methods for serological assay and data analysis on such microarrays, including test cases with data originated from a malignant melanoma cohort.

Keywords Protein microarrays • Cancer–testis antigens • Cancer biomarker discovery • Bioinformatic analysis • Pipeline

J. Duarte • J.-M. Serufuri • N. Mulder • J. Blackburn, D.Phil (✉)
Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, N3.06 Wernher Beit North Building, Anzio Road, Observatory, Cape Town 7925, South Africa
e-mail: jonathan.blackburn@uct.ac.za

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_3,

3.1 Introduction

In the postgenomic era, attention has turned towards the systematic assignment of function to proteins encoded by genomes. Bioinformatic methods are typically now used ubiquitously as an essential first step in assigning predicted function to open reading frames (Hunter et al. 2009). However, while such methods give helpful insights into possible function, there remain many examples of proteins that have closely related sequences and/or structures but which prove to have quite different functions when studied experimentally (Wise et al. 2002; Schmidt et al. 2003; Lander et al. 2001). As the number of sequenced genomes expands ever further, there is thus an ever-increasing need for experimental methods that enable the determination and/or verification of protein function in high throughput. At the forefront of this monumental task, the field of proteomics can be segregated into discovery- and systems-oriented proteomics (Macbeath 2002). Discovery-oriented proteomics is mainly concerned with documenting the abundance and localisation of individual proteins as well as building a picture of protein–protein interaction networks. This is the realm of 2-hybrid screens, 2D-gel electrophoresis and increasingly powerful more direct, isotope-labelling-based mass spectrometry methods; these latter two methods in particular are commonly used to understand the way in which expression profiles change in response to different stimuli by comparing, for example, diseased and healthy cell extracts. However, these discovery-oriented proteomic methods tell us little directly about the precise function of individual proteins or protein complexes, even when augmented by ever more sophisticated bioinformatic methods. Systems-oriented proteomics takes a different approach; rather than rediscovering each protein in each new experiment, the focus is on a predefined set of proteins – in principle up to an entire proteome, but in practice more typically a limited subset thereof – enabling the functionality of each member of that set to be dissected in great detail (Wolf-Yadlin et al. 2009). However, obtaining quantitative and genuinely comparative functional data across large sets of proteins with any degree of accuracy is technically difficult, requiring isolation of each individual protein in an assayable format. We and others have chosen to focus on protein function microarray-based methods because the parallel, high-throughput nature of microarray experiments is attractive for analysing large numbers of protein interactions, while the uniform intra-array conditions both simplify and increase the accuracy of assays (Wolf-Yadlin et al. 2009; Boutell et al. 2004; Kodadek 2001; Predki 2004; Zhu et al. 2000, 2001; Michaud et al. 2003; Fang et al. 2003). Additionally, the small volumes of ligand or reaction solution required to perform assays, typically tens to hundreds of microlitres, can provide economic advantages, for example, when using expensive recombinant proteins or labelled compounds.

The key element to such microarray experiments is that the arrayed, immobilised proteins retain their folded structure such that meaningful functional interrogation can then be carried out. There are a number of approaches to this problem, which

differ fundamentally according to whether the proteins are immobilised through non-specific, poorly defined interactions or through a specific set of known interactions. The former approach is attractive in its simplicity and is compatible with purified proteins derived from native or recombinant sources (MacBeath and Schreiber 2000; Angenendt et al. 2003) but suffers from a number of risks. Most notable among these is that the uncontrolled nature of the interactions between each protein and the surface might at best give rise to a heterogeneous population of proteins or at worst destroy activity altogether due to partial or complete surface-mediated unfolding of the immobilised protein. In practice, an intermediate situation probably most often occurs, where a fraction of the immobilised proteins either have undergone conformational change as a result of the non-specific interactions or have their binding/active sites occluded by surface attachment; these effects effectively reduce the specific activity of the immobilised protein and therefore decrease the signal-to-noise ratio in any subsequent functional assay that is sensitive to conformation. It is, therefore, important to consider the possible effects of unfolding on the intended downstream assay prior to choosing an array surface: for example, an assay in which a solution-phase kinase phosphorylates arrayed proteins may well be sensitive to disruption of the relative three-dimensional arrangement of targeting and substrate domains in the arrayed proteins (Blackburn and Shoko 2011); by comparison, an assay in which solution-phase antibodies bind to linear epitopes on the array will be unlikely to be affected by unfolding of the arrayed proteins – indeed, it may even be desirable to deliberately unfold such proteins in order to expose a greater range of potential epitopes. However, an important caveat here is that unfolded proteins are also more likely to exhibit non-specific binding to antibodies via the now-exposed hydrophobic surfaces and, across an array of diverse proteins, this non-specific binding may give rise to a high rate of false positives in serological assays.

The advantages of controlling the precise mode of surface attachment are that, providing the chosen point of attachment does not directly interfere with activity, the immobilised proteins will have a homogeneous orientation resulting in a higher specific activity and higher signal-to-noise ratio in assays, with less interference from non-specific interactions (Koopmann and Blackburn 2003). This may be of particular advantage when studying protein–small-molecule interactions or conformationally sensitive protein–protein interactions in an array format. The disadvantages of this approach though are that it is really only compatible with recombinant proteins or with families of proteins, such as antibodies, which have a common structural element through which they can be immobilised. However, in a systems-oriented approach, the disadvantage of working with recombinant proteins is largely outweighed by the problems encountered in individually purifying large numbers of active proteins from native sources. In addition, experimental approaches that facilitate high-throughput expression and purification of many different proteins in parallel have become more generally accessible over recent years, simplifying access to larger, defined collections of recombinant proteins. An important caveat here though is that it is increasingly clear that despite its ease of use, *Escherichia coli* is not an optimal host for

recombinant expression of folded, functional mammalian proteins. Furthermore, while cell-free transcription/translation-based protein microarray systems have been described (He and Taussig 2001; Ramachandran et al. 2004), it remains unclear how reproducible such arrays are or what proportion of mammalian proteins produced by such approaches are properly folded and therefore functional prior to immobilisation (Blackburn and Shoko 2011) – thus, their true utility in cancer biomarker discovery applications also remains unclear.

Having decided on and created the protein content and then fabricated a reproducible protein microarray, a typical downstream protein microarray experiment generates a large amount of raw and often noisy data that requires preprocessing via a series of computational steps in order to filter and normalise the data to yield a robust clean data set which can then be analysed using various bioinformatic tools in order to determine its biological significance (Klein and Thongboonkerd 2004; Brusic et al. 2007). However, DNA microarray software solutions in general have proved to be not well suited to the analysis of protein microarray data – for reasons that will be discussed below – and a comprehensive analysis software solution designed for protein microarrays has yet to be produced. To address this issue, our group has developed a preprocessing and quality control pipeline for raw protein microarray data (Zhu et al. 2006) which we will discuss in more detail below.

3.2 Custom Antigen Arrays as Serological Diagnosis Tools

A number of different types of protein microarrays are currently used to study the biochemical activities of proteins. Among them, the analytical microarrays are typically used to profile complex mixtures of proteins and to estimate the level of expression, binding affinities and specificities of specific components. These types of protein microarray-based experiments have interesting applications in molecular medicine, such as biomarker discovery for diagnosis, drug design and development as well as increasing understanding of pathogenesis and disease biology (Hall et al. 2007).

In the cancer field, it is well understood that individual patients generally show aberrant expression and/or post-translational modification of a selection of antigens (‘cancer antigens’), suggesting that detection and quantitation of cancer antigens should be a promising approach in, for example, disease diagnosis. In an ideal world, cancer biomarkers should be easy and inexpensive to measure to allow easy accessibility in developing countries, should be measureable in peripheral fluids (e.g. serum) to avoid unnecessarily invasive treatments and should be highly accurate for diagnostic and/or prognostic purposes to allow improved clinical management of patients (Berrade et al. 2011; Frank and Hargreaves 2003; Rifai et al. 2006).

Considering serum as a target peripheral fluid for cancer diagnosis, the first step today typically involves identification through serology (using the term in its broader

sense) of a set of candidate serum biomarkers that might correlate with disease status. In any serological analysis of cancer patient samples, there are two complementary approaches that could be taken: one obvious approach is to identify and quantify circulating tumour antigens that are aberrantly expressed in cancer; however, it is also known that aberrant protein expression and/or post-translational modification results in measureable autoimmune responses in most cancers, thus providing a second approach to serological analysis based on the identification and quantitation of circulating autoantibodies to tumour antigens.

Circulating tumour-antigen profiles may be analysed either by direct mass spectrometry-based proteomic techniques or by using antibody microarrays made up of immobilised antitumour-antigen antibodies. Alternatively, circulating autoantibody profiles can be analysed using protein microarrays made up of immobilised tumour antigens. There are pros and cons to each approach. Mass spectrometry approaches to de novo discovery of serological markers have not met with much success to date, largely due to the complexity and dynamic range of the serum proteome. Selected reaction monitoring (SRM) mass spectrometry-based assays that target known antigens are likely to be more successful but still today require significant preprocessing of serum samples (e.g. depletion of abundant proteins, tryptic digests and LC separations) that may influence downstream quantitative assays.

Antibody microarrays provide a much more direct means to analyse a wide range of different cancer antigens in what amount to miniaturised, multiplexed ELISAs. Numerous antibody microarrays are now available commercially and are being developed to be able to both quantify the individual targeted antigens and to also assess changes in post-translational modifications (e.g. glycosylation) of a given antigen between samples. The obvious current limitation in antibody microarray technology is the availability of large collections of suitable high specificity anticancer-antigen antibodies; this shortcoming is being addressed via a number of public initiatives (e.g. the Human Protein Atlas (www.proteinatlas.org), which currently describes antibodies to 12,238 human proteins). However, a less obvious limitation is that the antibodies used in antibody microarrays are typically murine in origin and have binding affinities for their target antigens in the nanomolar range, which limits the ability of antibody microarrays to detect circulating tumour antigens at the sub-nanomolar concentrations likely to be present at low tumour loads, that is, in early disease; this is less easy to address and may well prove to be a limitation for antibody microarrays into the future (note that in any such antibody-based microarray assay, the array-based signal is dependent both on the analyte concentration and also on the antibody–analyte affinity due to simple equilibrium binding considerations. Thus, as the analyte concentration significantly falls below the K_d of the antibody–analyte binding interaction, the proportion of immobilised antibody that is bound by analyte – and therefore the signal from the array – becomes nonlinear and falls off rapidly into the background noise).

Antigen microarrays by comparison seek to detect and quantify circulating cancer-specific human autoantibodies, which are typically highly specific for the tumour antigen and show picomolar or lower affinities. Antigen microarray-based serological analyses thus offer unique opportunities to find novel disease-specific serological

markers – that is, cancer-specific autoantibodies – that are not readily accessible by other proteomic technologies (Matarraz et al. 2011). More specifically, the study of autoantibody profiles in cancer patients could lead to the discovery of novel biomarkers for, inter alia, early detection of tumours, patient stratification, personalised patient treatment, development of improved therapies, and monitoring therapeutic response and disease progression. Importantly, antigen microarrays provide the enticing possibility of detecting much lower concentrations of autoantibodies than are detectable for the cancer antigens themselves (due to the intrinsically high affinity of human autoantibodies), thus in principle bringing forward in time the ability to detect the cancer biomarkers and potentially enabling presymptomatic diagnosis. As with antibody microarrays, a limited number of antigen arrays are now available commercially (e.g. the Invitrogen Human ProtoArray which contains ca. 9,000 individually purified, denatured human proteins).

However, antigen microarrays also suffer problems, primarily due the fact that often the aberrant form of any given cancer antigen that is misrecognised by the host immune system as an autoantigen is poorly defined, making it difficult to create recombinant forms of the autoantigen for reproducible protein array fabrication. The choice and customisation of the antigen content on such microarrays is thus a critical component in experimental design and will strongly influence the likelihood of downstream success (Zhu et al. 2006; Ingvarsson et al. 2007; Hultschig et al. 2006; Sanchez-Carbayo 2006; Casiano et al. 2006). One group of cancer antigens – the so-called cancer–testis antigens – therefore stands out as being of particular potential utility in serological analyses of cancers.

The cancer–testis (CT) antigen family are a group of >90 structurally and functionally unrelated proteins that show highly restricted expression only in germ cells in the adult male testis, as well as in occasionally in the adult ovary and the trophoblast of the placenta (Scanlan et al. 2002). Critically, these are all immune-privileged compartments, so the adult immune system has typically never been trained to recognise CT antigens and ‘self-antigens’. The CT antigens are however aberrantly expressed in many cancers due to the disruption of gene regulation. When this occurs these proteins are thus misrecognised as autoantigens, making them potential cancer diagnostic markers as well as vaccine targets.

The expression of different CT antigens is known to be associated with many different types of cancer. However, the absence or presence of expression of any one CT antigen is not in itself exclusively indicative. Thus, as with many other candidate biomarkers, CT antigens are therefore not thought to be individually viable as diagnostic or prognostic markers (Scanlan et al. 2002; Anderson and LaBaer 2005). Instead, it is possible that patterns of CT antigen expression may be used as correlates of specific cancer types and of disease progression, making this family of cancer antigens particularly well suited to serological study via an antigen microarray-based approach. Importantly, recombinant forms of CT antigens typically retain immunological activity, further enhancing their suitability as content for customised cancer-antigen microarrays.

In contrast to systemic autoimmune diseases, where the presence of a single autoantibody might have a diagnostic value, tumour-associated antibodies have little diagnostic value when detected individually for three reasons: (1) the frequency of a specific antigen within a cohort of patients is often relatively low; (2) several tumour-associated antigens are responsible for tumorigenesis in multiple cancer types, so the detection of the associated autoantibody can only indicate the presence of a developing tumour without enabling discrimination between different cancer types; and (3) several CT antigen-associated autoantibodies lack specificity, as they might arise from events associated with other diseases. We therefore concluded that the characterisation of autoantibody profiles against a wide panel of CT antigens via a CT antigen array would be significantly more informative than the detection of autoantibodies against individual-specific antigens (Casiano et al. 2006; Robinson 2006). Our group has therefore developed a CT antigen array ('CT100 array') for the in vitro serological assessment of autoimmune responses to cancer-specific targets.

In seeking to exemplify the utility of our CT antigen array, we have focussed initially on the observation that an increasing number of clinical trials in progress today are examining the safety and efficacy of therapeutic cancer treatments which either target specific tumour-associated antigens (e.g. the CT antigens NY-ESO-1 and MAGE A3) or aim to transiently deregulate self-recognition of molecules by targeting components of the peripheral tolerance machinery such as CTLA4 (Ueda et al. 2003). However, in therapeutic vaccine trials, the chronic nature of the disease makes it difficult to provide an early assessment of whether an individual patient is generating a therapeutically useful response following vaccination or not. T-cell-based assays for cytokine release have been widely used in the search for correlates of therapeutic response, but so far without great success.

We have therefore applied our CT antigen microarray-based approach together with robust bioinformatic algorithms for data analysis, as a generic tool with which to study the serum antibody (i.e. B-cell) responses to therapeutic cancer experimental treatments, accepting implicitly that such analyses will only be a surrogate for the T-cell responses currently desired in cancer vaccine strategies. We will exemplify this approach below by reference to use of our CT100 microarray in the assay of serum samples from malignant melanoma patients who were undergoing an experimental vaccine treatment, our goal being to identify autoantibody 'biosignatures' that could serve as potential biomarkers of therapeutic response.

In this chapter, we provide an overview of the complete process necessary for creation and use of customised antigen microarrays in serological analyses. We consider general issues that arise in antigen content generation, microarray fabrication and microarray-based assays and provide practical examples of experimental approaches that address these. We illustrate these approaches by reference to serological assay of samples from a malignant melanoma cohort using a multiplexed cancer-antigen microarray that contains 100 unique, purified, immobilised antigens in a spatially defined array. We then focus on general issues that arise in raw

data extraction, raw data preprocessing and analysis of the resultant preprocessed data to determine its biological significance, and we describe computational approaches to address these.

3.3 Protein Microarray Technology

Fundamentally, protein microarray technology is based on the immobilisation of multiple proteins onto a surface (typically glass, gold or plastic) in a spatially defined array for use as capture probes. This technology permits numerous biological questions to be addressed via the high-throughput analysis of biochemical and/or biological interactions between the arrayed proteins and other biomolecules contained in complex sample solutions, thereby generating significant volumes of proteomic information that require analysis (Hardiman 2003).

The high-throughput manipulation of proteins to create protein microarrays is considerably more challenging however than the manipulation of oligonucleotides to create DNA microarrays; this is primarily due to the very diverse physical and chemical properties of different proteins, including differing stabilities, binding affinities and the requirement often of folded 3D structure for biological activity (Hardiman 2003; Draghici 2003). Considerations regarding protein acquisition, immobilisation and assay are discussed further below.

3.3.1 Antigen Content Generation

The methods used in antigen content generation usually rely on the identification of open reading frames (ORFs) and the selection of the most appropriate plasmids for protein expression. An ORF is a protein-coding segment within the DNA sequence that encodes all the amino acids between the initiation and termination codons. This ORF is typically amplified by PCR prior to insertion into an expression plasmid. In the case of eukaryotic genes that are subject to splicing, the protein encoding ORF is usually derived from cDNA – itself prepared from mRNA – to ensure that the correct protein sequence can be expressed in recombinant form in a suitable host (Hall et al. 2007; Phizicky et al. 2003; Gray et al. 1982). Heterologous expression is generally used, even though it may lead to expression issues. For example, in more than 60% of cases, soluble proteins expressed in *Escherichia coli* (*E. coli*) show altered solubility, suggesting folding defects, and lack any post-translational modifications. Insect cells by contrast provide a relatively simple eukaryotic alternative to *E. coli* and, importantly, utilise eukaryotic co-translational folding and post-translational modification pathways to typically yield properly folded forms of recombinant eukaryotic proteins with similar post-translational modifications to mammalian cells (Phizicky et al. 2003). In the protocols developed by our group, we

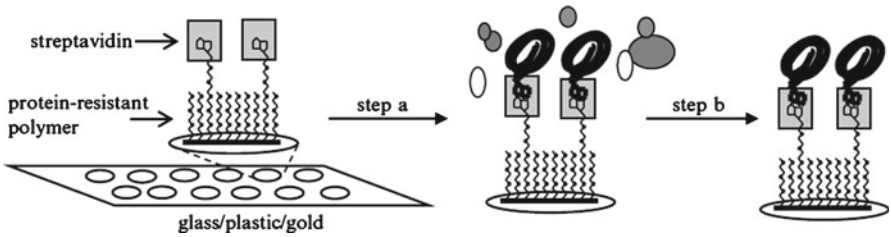


Fig. 3.1 Schematic of single-step immobilisation/purification route to array fabrication. The array surface is intrinsically 'nonstick' with respect to proteinaceous material but has a high affinity and specificity for biotinylated proteins. Crude cellular lysates containing the recombinant biotinylated proteins can then be printed onto the surface in a defined array pattern (*step a*) and all non-biotinylated proteins removed by washing (*step b*), leaving the recombinant proteins purified and specifically immobilised via the affinity tag in a single step

thus primarily use insect cells to express our human antigens of interest (Beeton-Kempen et al. submitted) (see Supplementary Material for detailed protocol).

3.3.2 Protein Immobilisation on Microarray Surfaces

The techniques of immobilisation are important both for effective concentration and orientation of immobilised proteins on the surface and also to preserve their folded conformations. There are two categories of protein immobilisation methods, covalent and non-covalent. The covalent immobilisation method is based on a covalent coupling to a cross-linker attached to the surface. By contrast, the biotin–streptavidin methodology used by our group is a non-covalent immobilisation method based on the high affinity of the biotin and streptavidin interaction. This method links biotinylated macromolecules to a surface that was previously derivatised with streptavidin via a single point of attachment (Fig. 3.1) (MacBeath and Schreiber 2000; Büssow et al. 2001) (see Supplementary Material for detailed protocol: Methodology, Sect. 1.4).

3.3.3 Detection of Binding Interactions on Microarrays

One of the main goals in protein microarray experiments is the quantitation of interactions between the probes immobilised on the slide surface and target analytes contained in the sample solution. To permit detection of this typically bimolecular interaction, molecules with specific interaction properties are labelled with either fluorescent, photochemical or radioisotope tags. The choice of one detection method over another depends on the need to reach a low signal-to-noise ratio at an affordable cost for the specific assay in question.

Since the vast majority of target analytes are not naturally coloured, fluorescent, bioluminescent or radioactive, detection of the molecular interaction on a protein microarray usually requires that some detectable molecule (a 'label') be included in the assay, either by direct conjugation to the analyte molecule itself or by conjugation to some secondary detection agent (e.g. an antibody) that can also bind to the analyte once it is specifically captured onto the array surface. Label-free biosensors (e.g. surface plasmon resonance- and quartz crystal microbalance-based biosensors) circumvent this problem, but in most manifestations are not yet truly compatible with medium- to high-density protein microarrays, plus they struggle to match label-based methods in terms of assay sensitivity (i.e. limit of detection), so will not be considered further here.

The choice of potential labels is wide: Chemiluminescence is a highly sensitive label-based detection method, but it has a relatively limited dynamic range; radioactivity-based methods are much less frequently used today because of safety concerns; fluorescent labelling is still therefore currently the most widely used detection method for protein microarray experiments since it is highly sensitive, stable, safe and effective and can be archived for future imaging. However, the direct labelling of analyte molecules might affect their ability to interact with their respective binding partners (Klein and Thongboonkerd 2004; Hall et al. 2007; Schweitzer et al. 2003). In the context of antigen microarray-based assays to measure human autoantibody profiles, the simplest mode of detection is thus via use of a fluorescently labelled antihuman IgG as a secondary detecting agent, since this will bind with high affinity to all human autoantibodies captured onto the antigen array surface, removing the need to label each target analyte in every biological sample.

3.4 Protein Microarray Data: Preprocessing

Even though microarray technologies have been around for many years, they are still subject to bias and variations. In addition to the variations introduced by probe acquisition, immobilisation and detection method, there are still a large number of environmental factors which might introduce variability in these experiments, including ambient conditions when the arrays were processed, the individual conducting the experiments, the recombinant sample differences, the variations in sample preparation, the nonuniformity in the hybridisation across an array surface, the distribution of artefacts or smears on the surface of the array and modifications in the scanner settings used for acquisition of fluorescent data. Nevertheless, a good experimental design can reduce noise and be beneficial for the downstream data analysis (Ingvarsson et al. 2007; Hultschig et al. 2006; Smyth and Speed 2003; Steinhoff and Vingron 2006; Altman 2005).

In addition to designing a standard optimised protocol to reduce noise within the microarray data, the experimental design should include the capacity to assess the quality of the microarray data, to filter out poor quality arrays and to normalise good quality data. With this consideration in mind, controls should be immobilised

onto the array surface, such as housekeeping and exogenous controls. Housekeeping controls are assumed to maintain a constant signal, either individually or collectively, within the different conditions of the experiments, while exogenous controls are those from species other than the one under study, which are generally selected to not give rise to a signal. Before proceeding with an intended array layout and assay, it is essential to identify a stable source of controls (Causton et al. 2004). In the context of an antigen microarray-based autoantibody profiling assay, simplistically one could consider arraying human IgG and sheep IgG as positive and negative controls.

3.4.1 Image Processing: Raw Data Extraction

Following a typical serological assay on an antigen microarray, in spite of several washing steps involved in our protocol, the spot quantification still remains affected by intangible factors. As illustrated in Fig. 3.2, the same lab protocol can lead to different results regarding the background of different replica arrays. The measures of background intensity usually represent the autofluorescence of the array surface at different spot locations (Causton et al. 2004).

The aim of image processing is to estimate the amount of specific anti-CT antigen autoantibody present in the serum by measuring pixel intensities across all probed spots. The image analysis software (ArrayPro Analyzer software (Media Cybernetics Inc., USA)) associated with the scanner (Tecan LS Reloaded fluorescence microarray scanner, Tecan Group Ltd., Switzerland) allows us to retrieve some statistics for the pixels measured in both the spots and their local backgrounds, for instance, the mean and median pixel distributions of the spots and their adjacent backgrounds. The scanner PMT (photomultiplier tube) gain setting helps discriminate between a weak signal and its background. By increasing the PMT gain setting, the sensitivity of the signal is improved but the selection of the optimal gain setting must provide a balance between the need to detect as many spots as possible and the necessity to avoid saturation of any of the spots.

The scanner software proceeds by matching a grid layout defined by the user (typically based on a .gal file generated by the microarray printer) to the actual image coming from the array and by locating the signal spots in order to quantify them. The spot finding can be achieved manually, automatically or semiautomatically. In the manual approach, the user adjusts a grid over the array and fits each spot individually to account for spot size variations and uneven spacing between spots. In practice, this approach is time consuming and subject to human error, especially when dealing with a large number of arrays. The automatic approach aims to identify and fit, without human intervention, the extent of each spot using specific algorithms. This approach is rapid and avoids human errors, but noise and contamination can lead to false detection of certain spots. The semiautomatic spot finding approach used by our group relies on automatic spot finding approach followed by manual curation of the grid alignment.

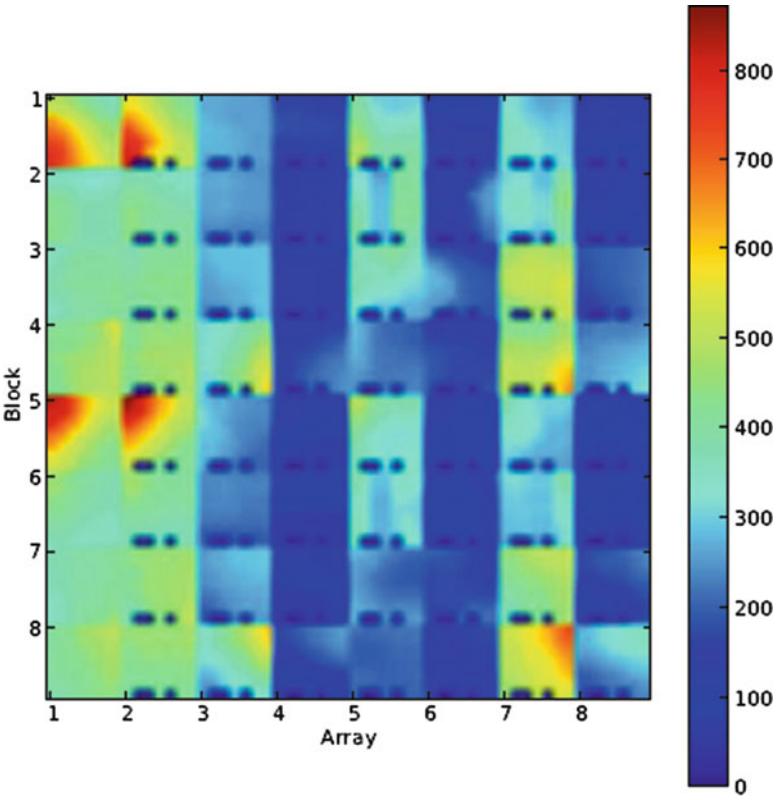


Fig. 3.2 Illustration of the variability of the background intensities of eight arrays with the application of the same experiment protocol. Here the eight blocks of each array are represented in the columns

3.4.2 Image Processing: Pixel Segmentation

Within a given spot area, individual pixels may be either true signal, background signal or erroneous signal originated from dust particles/artefacts. Typical image processing software outputs from scanned microarray images include some statistics based on pixel distributions. Image segmentation then aims to define rules to filter out unwanted pixels (see Fig. 3.3).

The simplest image segmentation rules include 'pixel filtering' and 'trimmed pixel' methods. The pixel filtering method sets an arbitrary threshold to filter out low-intensity pixels and performs statistics on the remaining pixels. The trimmed pixel method assumes that most of the pixels in the spot area are true signal, while most of those in the adjacent area belong to the background; for each spot or background intensities, the pixels falling outside defined quantiles are considered to be outliers and are therefore trimmed off (i.e. removed from further analysis) with all

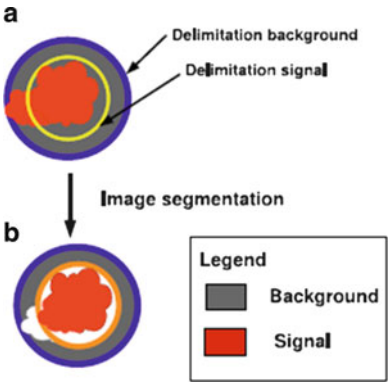


Fig. 3.3 Illustration of image segmentation. (**a**) shows the result of the spot finding process and (**b**) shows the result of image segmentation

subsequent statistics (e.g. mean and median foreground and background pixel intensities for each spot) being based on the remaining pixels. Significant differences between the effectiveness of the various segmentation methods become more noticeable as the level of artefacts on the arrays increases.

3.4.3 Image Processing: Quality Control

When evaluating the quality of an array image, there are many considerations that should be taken into account, such as:

1. Spot-to-spot variation – when looking at the signal of triplicate spots, one should expect a similar signal across all replicas, as well as uniform spots across the whole slide. Variations which may occur include spot bleeding (2 or more spots run into each other due to close proximity between spots or inappropriate spotting buffer, compromising the signal of all affected spots); pin sticking or erroneous pin calibration (inadequate cleaning of the arrayer pins and printhead between print runs could lead to the failure to print some or all intended spots, as well as non-reproducible printed spots); and washing artefacts and speckles (inadequate washing steps throughout the assay could lead to large washing artefacts which appear as negative spots or random additional smaller spots across the slide) on the array surface.
2. Spot homogeneity – when looking at the signal of an individual spot, one should expect a homogenous signal across all pixels within that spot. Variations which may occur include the 'doughnut' effect (inadequate pin height during print run and liquid residues on the pin body when immersing pin head in source plate could lead to uneven spot distribution); dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles on spots of interest, which appear as high-intensity pixels and skews the real signal) on the array surface; and temperature and humidity conditions (increased

temperature/decreased humidity may lead to the evaporation of printed spots, and humidity above 75% may lead to condensation, which would account for an uneven intensity within spots). The homogeneity between replica spots can be measured by calculating the coefficient of variation (CV), which is the ratio between the standard deviation (SD) of all pixel intensities within a spot and the mean intensity as a percentage. The CV should not exceed a value of 20%.

3. Background variation – when looking at several spots across an array, one should expect low variation of the background between the neighbourhood spots. Variations which may occur include dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles around spots of interest, which results in high local background for a specific spot and difficulty distinguishing between real signal) on the array surface.
4. Signal-to-noise ratio – when looking at several spots across an array, one should expect the spot intensity to be greatly above its local or neighbourhood background. Variations which may occur include washing artefacts and speckles (inadequate washing steps throughout the assay could lead to large washing artefacts which appear as negative spots or random additional smaller spots across the slide) and dust particles (inadequate storage and handling of slides during assays could lead to the presence of dust particles around spots of interest, which results in high local background for a specific spot and difficulty distinguishing between real signal) across the array surface. To be confident that the net spot intensity (i.e. foreground intensity minus background intensity) is significantly above background, a signal-to-noise ratio of at least 2 is used for quality assurance, with 'noise' defined as the standard deviation of the background pixels.
5. Saturated pixels – when looking at several spots across an array, no saturated pixels should be visible within the spots of interest, as these are above the scanner's reading capacity (approximately 65,000 RFU in the case of the Tecan Reloaded scanner used in our laboratory). If this occurs, the slide should be rescanned at a lower PMT gain setting until no saturation is visible across the array (Schäferling and Nagl 2006; Espina et al. 2003).

3.4.3.1 Background Correction and Subtraction

Following pixel segmentation, the background intensity for a given spot is estimated from the adjacent area of that spot and is subtracted from its foreground intensity to obtain the net intensity, that is, the true signal derived from the specific interaction being interrogated on the array – in our case, the antigen–autoantibody binding interaction. The most important factor here is to avoid including artefacts in the estimation of the background because that could arbitrarily induce overestimated background intensity and, as a result, artificially reduce the true value of the spot intensity.

Module 1 of the protein chip analysis tools (ProCAT) (Zhu et al. 2006) provides a robust way to tackle the issue of local artefacts in background signals. The ProCAT approach for background correction essentially replaces the local

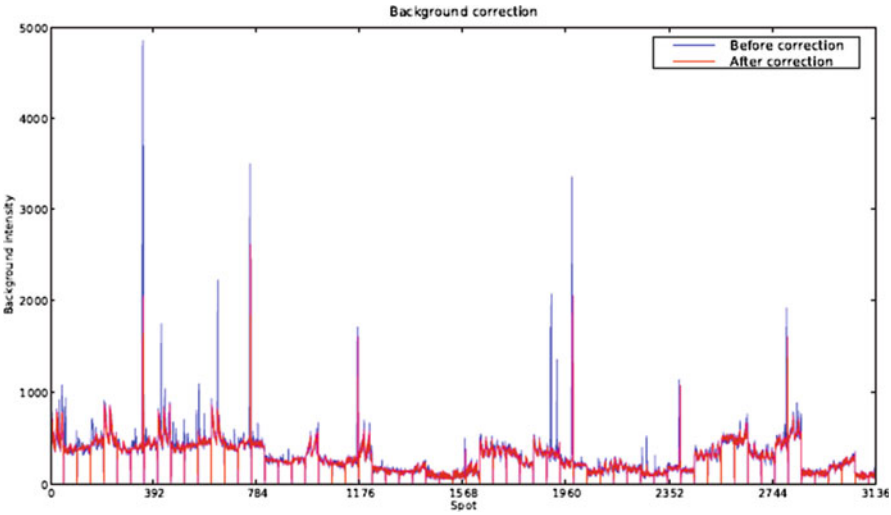


Fig. 3.4 The background correction corrects the arbitrary peak of the background intensity in a 3×3 spot window in a sample test case

median background intensity of a specific spot by the 'neighbourhood background', $0.5em\widehat{B}_{i,j}$, defined as the median background pixel intensity of a surrounding 3×3 spot window centred on the spot of interest:

$$\widehat{B}_{i,j} = \text{median}\{B_{i',j'}\}$$

$$i-1 \le i' \le i+1; j-1 \le j' \le j+1$$

where *i, j, i′* and *j′* design the row and column coordinates of spots.

This neighbourhood background correction smooths the local background (see Fig. 3.4) by reducing the effect of artefacts and noise in the background, thus enabling calculation of more accurate net intensities by subtracting the corrected neighbourhood background value from the median local foreground pixel intensity.

3.5 Protein Microarray Data: Filtering

Following background subtraction for each spot on the array, it is useful to then run a number of quality control (QC) tests to filter out noisy or defective array data prior to bioinformatic analysis. Data filtering therefore increases the data quality by flagging

questionable and low-quality arrays and/or individual spots. Our approach to this problem utilises a collection of criteria to filter out poor quality data. Among these are:

1. Flagging of spots with foreground intensity close to the saturation level (65,536RFU).
2. Flagging of triplicates based on their variability, as measured by the CV of their net intensities. N.B. When one of the triplicates is flagged, the measure of variability is then defined by the two remaining spots $S1$ and $S2$ as being equal to $(|S1 - S2|)/(S1 + S2)$.
3. Flagging of net signal intensities close to the noise level. A way to estimate the level of noise in the neighbourhood of a spot is to measure the standard deviation of the local background and to stipulate a noise threshold of 2SD of the local background pixel intensity (Tecan LS TM).

The negative controls on the array surface can in principle also enable the filtering of low-intensity spots because they reflect the cross reactivity of the detecting antibody with copurifying insect cell proteins or with the BCCP tag. However, in reality this typically proves less straightforward since both the expression level and the physical accessibility of the BCCP tag may differ in a difficult-to-quantify manner from antigen to antigen. It may therefore prove more effective in practice to define a baseline grass intensity signal for each antigen observed across a significant number of samples from healthy volunteers, when available.

3.6 Protein Microarray Data: Normalisation

Depending on the printing method used to fabricate the protein microarrays, each spot may or may not be printed with the same pin or nozzle. Where different spots are printed by different pins/nozzles, there may be subtle differences in the volumes of the printed antigens and therefore in the density of the immobilised antigens. Such differences can be corrected by so-called 'pin-to-pin' normalisation. More importantly, to correct for differences in the density of the same arrayed antigen across replica arrays, as well as to correct for any other systematic, non-biological variation between arrays, so-called 'array-to-array' normalisation is typically then applied.

The overall purpose of normalisation is thus to correct microarray data from variations in their measurements due to processes other than the targeted biological activity, thereby improving the quality of the data and allowing comparison of data originating from different arrays and different experiments (Lu et al. 2005; Oshlack et al. 2007).

3.6.1 DNA Microarray Normalisation Methods

Numerous methods for microarray normalisation have been published for 'two-colour' DNA microarrays. In two-colour arrays, two samples – control and target – are assayed on the same array with control and target analytes labelled with different fluorophores, simplifying downstream normalisation of the data. However,

in most cases protein arrays are run as single-colour assays due to concerns about differential physical occlusion effects arising from the size of the assessed macromolecules, together with the fact that the analyte molecules themselves are typically not directly labelled. Protein microarray-based data therefore typically lacks strong biological assumptions to carry out normalisation (Zhu et al. 2006). However, the methods published for normalising DNA microarrays can still inspire and support the normalisation of protein arrays and can be classified into three categories (Smyth and Speed 2003; Steinhoff and Vingron 2006; Freudenberg 2005):

1. *Scaling methods* assume that the arrays being normalised share a common statistical measure, such as the mean or median of their spot intensities or even the total intensities of their spots, and apply a common factor to each spot intensity (Freudenberg 2005). Thus, if m_j is a statistical measure on chip j that is equalised across the chips after normalisation, the scaling factor α_j on chip j is then given by

$$\alpha_j = \frac{m}{m_j}$$

   where m is the final value of m_j after normalisation.
2. *Transformation methods* rely on assumptions that allow quantitative mapping of two sets of spot intensities. The most popular methods are curve fitting, LOWESS and quantile normalisation. The curve fitting method assumes that the distribution of the normalised data set is known, and attempts to identify parameters of the distribution model; for instance, Lu et al. (2005) suggested adapting Zipf's law for the normalisation of one or two-colour DNA arrays (Draghici 2003; Lu et al. 2005). The LOWESS (locally weighted polynomial regression) method maps data from two data sets using a polynomial regression within overlapping intervals (Draghici 2003) and is most effective when most of the spots within two arrays show similar intensities (Oshlack et al. 2007). The quantile normalisation method can be applied where the assumption of a common underlying distribution seems to be justified. It is fast, easily implementable and does not require statistical modelling of the data (Freudenberg 2005; Bolstad et al. 2003).
3. *Invariant set method* relies essentially on the ability to identify a suitable set of non-differentially expressed probes or 'housekeeping' probes. The selection of the set of invariant spots might be experiment dependent, and an inappropriate choice of housekeeping probes can lead to bias in results (Freudenberg 2005; Ploner et al. 2005).

3.6.2 Protein Microarray Normalisation Methods

Custom antigen arrays however are typically not directly amenable to any of these standard DNA microarray normalisation approaches (Oshlack et al. 2007) since often a relatively small selection of specific probes show strong signals for a given sample and the identity of these probes vary between samples. This breaks most of the usual assumptions based on the comparison of equivalent signals across arrays,

unless the samples being compared display special features (Oshlack et al. 2007). Therefore, two methods are widely used for normalisation of antigen microarray data: the first relies on housekeeping controls and the second on a microarray sample pool (MSP) control (Lu et al. 2005; Oshlack et al. 2007; Wilson et al. 2003). Housekeeping controls are ideally supposed to keep a consistent signal across experimental conditions and different samples. In regard to antigen arrays, such as our CT100 array, used in assessing antibody profiles in serum, the housekeeping controls should ideally be serum independent to enable comparisons across different samples. However, identification of suitable housekeeping proteins in serum is less straightforward, as shown by the many mass spectrometry-based proteomic experiments that have documented high variability in the serum proteome; this housekeeping control-based normalisation approach therefore may be of limited value in serological assays using antigen microarrays.

Alternatively, the MSP method selects probes from a heterogeneous pool library and dilutes them at different concentrations to cover ranges similar to those covered by the probes used in the experiment; these probes must be printed in a number large enough to enable the assumption of non-differential expression between samples (Oshlack et al. 2007). Transformation methods such as LOWESS can subsequently be applied for normalisation. However, the limited size of the typical custom antigen microarrays can be a limiting factor in the usage of the MSP method and may restrict its readily application in serological assays using custom antigen microarrays.

3.6.3 Composite Normalisation Methods for Custom Antigen Microarrays

The normalisation method used by our group aims to make more efficient, effective and robust usage of a relatively small number of positive controls to correct for systematic bias in pin-to-pin and array-to-array variations. Robustness here is taken to mean the ability of the method to cope with the flagging of some positive controls, while still being based on sound biological principles.

The normalisation assumption we make here is that our positive control spots share a common underlying distribution across the chips (block, arrays, etc.) on which they are printed. This perspective provides greater flexibility than assuming that the individual positive control spots maintain the same intensities across the chips. Thereafter, our composite normalisation method combines quantile and total intensity normalisation modules to correct for systematic bias among the chips, while providing more robustness when dealing with flagged positive control spots (Causton et al. 2004; Bolstad et al. 2003).

1. *Quantile-Based Module*
   Since the positive control spots – in our CT100 array, these are Cy5-labelled, biotinylated BSA – are replica spots across different arrays, it seems reasonable to assume that they share an underlying distribution across arrays, and the quantile approach can be used to identify the corresponding housekeeping spot intensities based on their intensity distributions.

Bolstad et al. (2003) described an algorithm to carry out spot identification within the same quantile according to the following steps, where S_{ij} is the intensity of a positive control spot i on chip j:

(a) Load the positive control spot intensities S_{ij} into an $I \times J$ matrix X.
(b) Sort the spot intensities in each column j of X to get X_{sort}.
(c) Take the means across each row i of X_{sort} and get $\overline{X}_i$.

$\overline{X}_i$ is considered the underlying distribution of the positive control spot intensities across chips (Bolstad et al. 2003). This reorganisation enables more flexibility in handling outliers or flagged spots within the positive control data set.

2. *Total Intensity-Based Module*
This module assumes that post-normalisation, all arrays have a common total intensity value of their positive control spots (i.e. the sum of all the positive control spot intensities on each array should be constant) (Causton et al. 2004; Quackenbush 2001), given by $\sum_{i=1}^{N_{\text{spots}}} \overline{X}_i$. The normalisation factor α_k to normalise array k is then given by

$$\alpha_k = \frac{\sum_{i=1}^{N_{\text{spots}}} \overline{X}_i}{\sum_{i=1}^{N_{\text{spots}}} \overline{X}_{ik}}$$

where $\sum_{i=1}^{N_{\text{spots}}} \overline{X}_{ik}$ is the total intensity of all the positive control spots on array k prior to normalisation.

This is a scaling normalisation method that assumes that different arrays share a common total intensity of their housekeeping spots, while taking into account the potential existence of flagged spots within the housekeeping spots. Importantly, if a given positive control spot is identified as an outlier on one array (i.e. it is flagged for some reason), the corresponding positive control spots across all arrays are identified in the quantile module above and are then also flagged across all arrays prior to normalisation; the net consequence of this is to ensure that the same number of positive control spots are considered across all arrays during normalisation.

3.6.4 Overall Preprocessing Pipeline

Our overall workflow, illustrated in Fig. 3.5, includes the following steps:

1. Spot filtering flags spots whose fluorescent pixels comprise less than 20% of an arbitrarily defined spot area, as well as spots that show saturation in >10% of the pixels.
2. Background adjustment corrects the local background of each spot to become the median neighbourhood background of its surrounding spots (3×3 window).

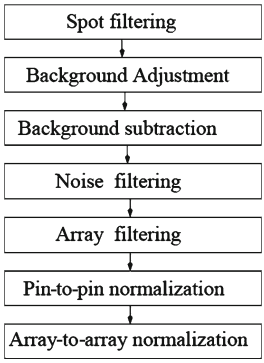


Fig. 3.5 Schematic of overall antigen microarray data preprocessing pipeline

3. Background subtraction gives the net spot intensity by subtracting the corrected neighbourhood background from the original raw median pixel intensities.
4. Noise filtration sets all net intensities to zero if they are lower than 2 SD of the background.
5. Array filtering calculates the CV of the positive controls within the array prior to normalisation and flags arrays with CVs above a user-defined threshold (we typically use a value of 30% here).
6. Pin-to-pin normalisation normalises the data based on our composite normalisation method in order to account for variations between the usage of different pins during the print run.
7. Array-to-array normalisation normalises the data, again using our composite method, in order to account for variations between arrays and to thereby allow, for example, comparison of serology data obtained on samples collected at different time points from the same patient (Safari Serufuri 2010).

3.7 Protein Microarray Data: Qualitative Clustering

Following preprocessing, quality control and normalisation of antigen microarray data, quantitative bioinformatic methods can then be applied with greater confidence to enable biological interpretation of the data.

The ultimate purpose of all clustering methods is to group or segment a set of items by taking into consideration a criterion of similarity or dissimilarity. The clustering can provide information regarding data structures as well as outliers – an outlier being an item not sufficiently similar to any other items in the data set. An additional goal of clustering can be to infer hierarchical order between clusters (Draghici 2003; Causton et al. 2004; Hastie et al. 2001; Boutros and Okey 2005; Costello and Osborne 2005).

Before choosing a clustering algorithm, one has to determine the intended type of cluster that might be expected from the data set and the most appropriate measure of similarity to capture the clusters of interest. Another essential consideration

is the performance of the selected algorithm (Hastie et al. 2001; Boutros and Okey 2005). Qualitative clustering can be based on the trend line similarities between antibody profiles of patient samples at a common time point or on trend line similarities between patients across a timecourse (e.g. trend line similarities in changes in autoantibody profiles between samples collected at different time points posttreatment).

We routinely use two clustering methods for analysis of our CT antigen array data: a factor analysis method and a *K*-means method using a Pearson correlation metric. Detailed mathematical description of these two clustering methods is beyond the scope of this chapter but a brief qualitative description is as follows:

The factor analysis method – used in an unsupervised mode – aims to investigate the number of intrinsic factors that are required to account for the correlations among variables or observations (Hastie et al. 2001). Factor analysis has been widely used in intelligence research to explain a variety of results based on different tests by identifying groups of correlated results. For instance, the performance at running, weight lifting and jumping could be explained by general athletic ability (Costello and Osborne 2005; Wikipedia 2010; Tryfos 2010).

The *K*-means method is among the most popular clustering algorithms. It is an iterative method relying on the minimisation of an objective function defining a measure of dissimilarity between the items or of their K-centroids, a centroid here being a measure by which the dissimilarity between clusters or between clusters and items can be summarised by one value (Causton et al. 2004; Hastie et al. 2001; Boutros and Okey 2005). The *K*-means method however requires a number of preconceived inputs to achieve the clustering, including the number of expected clusters K, the maximum number of iterations, the selection of the similarity metric (Pearson, Euclidian, etc.) and an initial clustering which is improved iteratively until a steady state or the maximum number of iterations is reached (Draghici 2003).

3.8 Experimental Design: Test Case of a Cancer–Testis Antigen Array for In Vitro Cancer Biomarker Discovery

We have applied our antigen microarray approach in cancer biomarker discovery projects and describe below a test case using serum samples from patients undergoing an experimental cancer treatment, the aim being to use our CT100 microarray to monitor changes in the autoimmune profiles of those patients following treatment as a possible correlate of therapeutic response.

3.8.1 Description

The ‘CT100’ array is a ‘one-colour’ CT antigen microarray designed to discover the expression profile of 100 CT- or tumour-associated (TA) antigens of interest (see Table 3.1) in specific cancer patients, as revealed by the binding of autoantibodies

Table 3.1 List of the 72 CT antigens (*yellow*) and the 28 non-CT antigens of interest (*blue*) present within each array field and comprising the 'CT100' array

Antigen identity	Antigen identity	Antigen identity	Antigen identity
BAGE2	LEMD1	SGY-1	CDK7
BAGE3	LIPI	SILV	FES
BAGE4	MAGEA1	SPAG9	FGFR2
BAGE5	MAGEA10	SPANXA1	MAPK1
CCDC33	MAGEA11	SPANXB1	MAPK3
CEP290	MAGEA2	SPANXC	PRKCZ
COL6A1	MAGEA3	SPANXD	RAF
COX6B2	MAGEA4 v2	SPO11	SRC
CSAG2	MAGEA v3	SSX1	CALM1
CT47.11	MAGEA v4	SSX2a	CDC25A
CT62	MAGEA5	SSX4	CREB1
CTAG2	MAGEB1	SYCE1	CTNNB1
CXorf48.1	MAGEB5	SYCP1	p53 S6A
DDX53	MAGEB6	THEG	p53 C141Y
MMA1	MART-1/MLANA	TPTE	p53 S15A
FTHL17	MICA	TSGA10	p53 T18A
GAGE1	NLRP4	TSSK6	p53 Q136X
GAGE2A	NXF2	TYR	p53 S46A
GAGE4	NY-CO-45	XAGE-2	p53 K382R
GAGE5	NY-ESO-1	XAGE3a v1	p53 S392A
GAGE6	OIP5	XAGE3a v2	p53 M133T
GAGE7	p53	ZNF165	p53 L344P
GRWD1	PBK	AKT1	CYP3A4
HORMAD1	RELT	CDK2	CYPR
LDHC	ROPN1	CDK4	EGFR

present in a patient's serum to the purified and spatially segregated, immobilised antigens. Uses of this CT antigen array include monitoring therapeutic responses and/or the rate of cancer progression in individual patients (Safari Serufuri 2010).

Recombinant gene cloning methods were used to clone each antigen into a relevant insect cell expression vector as a C-terminal fusion construct with the biotin carboxyl carrier protein (BCCP) tag (see Supplementary Material for detailed protocol). The BCCP tag is biotinylated in vivo in insect cells (as well as in *E. coli* and yeast) and allows single-step purification and immobilisation onto the array surface via the high specificity and affinity streptavidin–biotin interaction. Expression and biotinylation of each recombinant antigen in insect cells (see Supplementary Material for detailed protocol) was confirmed using Western blot analysis prior to printing, and crude lysates were then diluted with PBS containing 40% sucrose. Replica 'CT100' protein arrays were printed in a 4-plex format (i.e. 4 replica fields per slide) using crude cell lysates of Sf21 insect cells expressing each of the 72 CT antigens and 28 TA antigens of interest. Various controls were also included in each array field: 50 ng/μl human

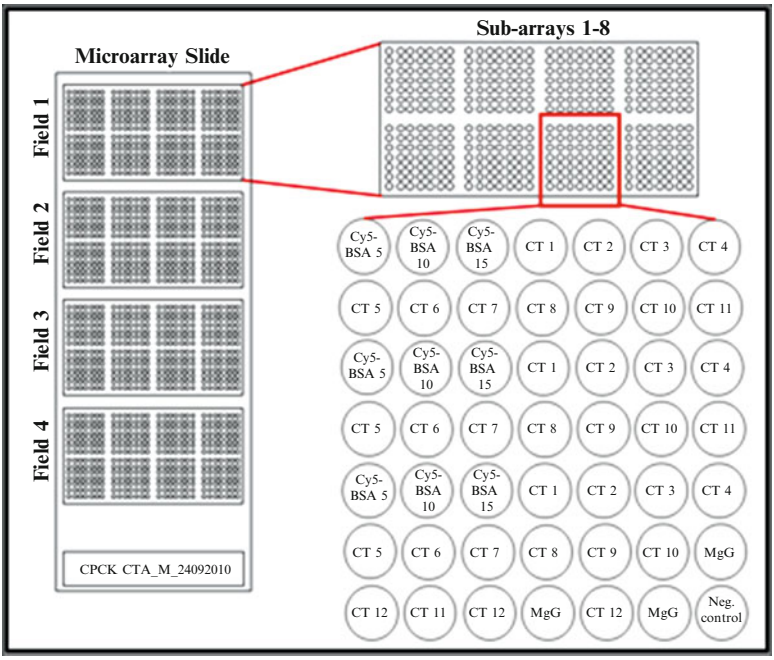


Fig. 3.6 Schematic of the CT100 array layout. Each slide contains four replicate fields each comprising all 100 antigens of interest (one field is used per patient sample assay). Each field contains eight blocks of 7×7 spots each. A representative block is shown with the *Cy5-BSA* controls, the various antigens, the *hIgG* positive control and a negative control

IgG (positive control), 200 ng/µl sheep IgG (negative control), as well as a crude insect cell lysate expressing BCCP only and no fusion protein (negative control). For slide orientation and signal normalisation, three different biotinylated Cy5-BSA concentrations were included in each sub-array (5, 10 and 15 ng/µl). Each sample and control was spotted in triplicate within each array field, while the Cy5-BSA concentration series was spotted in triplicate by each pin within each sub-array (i.e. 24 spots total per array field). Array design is shown in Fig. 3.6 below (Beeton-Kempen et al. submitted).

Each CT100 array was printed on an in-house streptavidin-coated surface (Nexterion slide H) using a Genetix QArray2 (Genetix Ltd., UK) robotic microarrayer with 8×300 µm flat-tipped solid pins (see Supplementary Material for detailed protocol). After printing, each slide was washed with prechilled blocking solution (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 1 mM DTT and 50 µM biotin) and stored at –20°C in storage buffer (same as the blocking solution except with 50% glycerol and no biotin). Under these conditions it proved possible to store these slides for up to 3 weeks prior to assay (Beeton-Kempen et al. submitted).

3.8.2 CT100 Assay Using Patient Serum

A collection of 100 serum samples from a malignant melanoma patient cohort (UCT HREC Ref number: 240/2011) was subjected to serological assay as follows:

Each serum sample was diluted 1:800 in PBS/0.1% Tween-20 and incubated on an individual CT100 array (prepared as described above, Sect. 3.8.1) at room temperature for 1 h. Each array was then washed independently with slide buffer (PBS/0.1% Tween-20). Cy5-labelled goat antihuman IgG (Invitrogen) was diluted 1:100 in slide buffer and incubated with each individual array for 1 h at RT. The individual arrays were then washed independently in slide buffer, dried and then scanned immediately using a Tecan LS Reloaded fluorescence microarray scanner (Tecan Group Ltd., Switzerland). All liquid handling steps were carried out on a Tecan HS4800 Pro automated hybridisation station (Beeton-Kempen et al. submitted). The arrays were scanned at a resolution of 20 μm using fixed gain settings of 110, 120, 125 and 135 PMT in order to determine the setting that gave the highest signal with minimal saturation across all slides.

3.8.3 CT100 Array Data Extraction

Using ArrayPro Analyzer v6.3 software (Media Cybernetics Inc., USA), a grid was semiautomatically autoaligned over the individual spots on each image file. A constant area feature finder was used across the array surface. ArrayPro was then used to extract the raw data from each of the spots (in batch processing mode), using a 200-RFU pixel threshold and a .gal file (output from the Genetix QArray2 printer) containing information about the identity of the sample in each spot. Upon extraction, ArrayPro provided both the mean and median foreground and background pixel intensities of each spot. This data was then used for further processing and analysis (Beeton-Kempen et al. submitted).

3.8.4 CT100 Array Data Processing

The net intensity for each spot was calculated by subtracting the corrected neighbourhood median background of surrounding spots from the local median foreground intensity of each spot. The average of the three replicate spots' net intensities was calculated to determine the mean signal for each sample. The data was then filtered to remove all spots displaying saturated signals, signals below the background signal plus two standard deviations noise threshold or where the signals observed occupied less than 20% of the area of the captured spot. Each array (corresponding to a unique patient serum sample) was inspected, and all array fields displaying >30% coefficient of variation across the Cy5-BSA spots were excluded from further analysis and the corresponding samples re-assayed on new arrays.

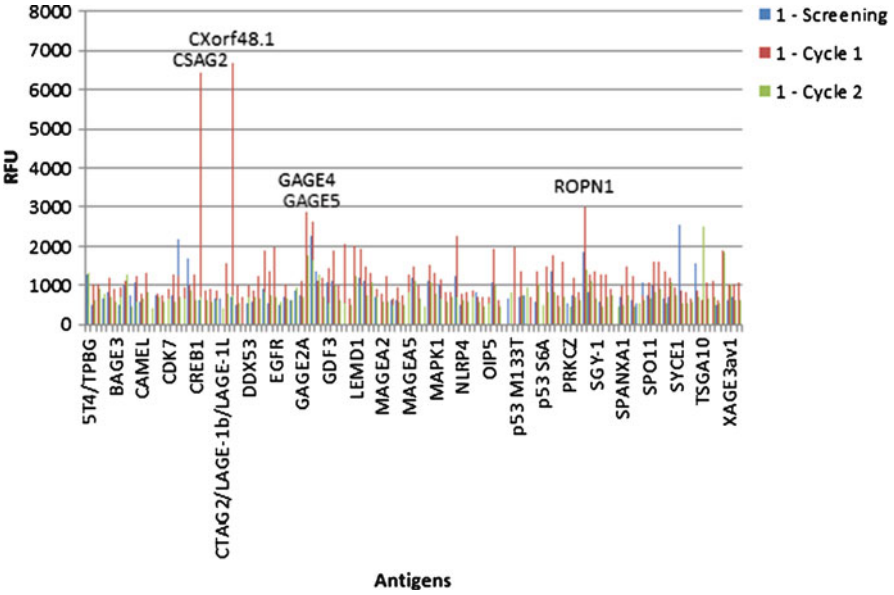


Fig. 3.7 Graph displaying the autoimmune profile of Patient 1 pre and post experimental treatment

Array-to-array normalisation was carried out using the net signal intensities of the 15 ng/µl biotinylated Cy5-BSA spots of each array and across each slide using the composite normalisation method developed by our group, as described above (Sect. 3.6.3 & 3.6.4) (Safari Serufuri 2010). All array images were also visually inspected for evidence of spot bleeding, washing artefacts or pin sticking, and relevant arrays were excluded from further analysis and the corresponding samples re-assayed on new arrays.

3.8.5 CT100 Array Results and Discussion

Serum samples were taken from patients prior to receiving an experimental treatment ('screening') and then following different cycles of treatment (cycles 1, 2, etc.). Not all patients had the same number of treatment cycles nor were these all carried out at the same time intervals. Nevertheless, distinct anti-CT antigen autoimmune responses were observed for the majority of this melanoma cohort prior to treatment, and these patterns appeared to be modified significantly in response to treatment (see, e.g. Figs. 3.7 and 3.8). Note that these two patients showed very different autoimmune profiles prior to treatment as well as differential responses to the experimental treatment. Here, erring on the side of caution, we interpret antigen signals from the array with net intensity >1,000 RFU as real and significant data. When comparing different

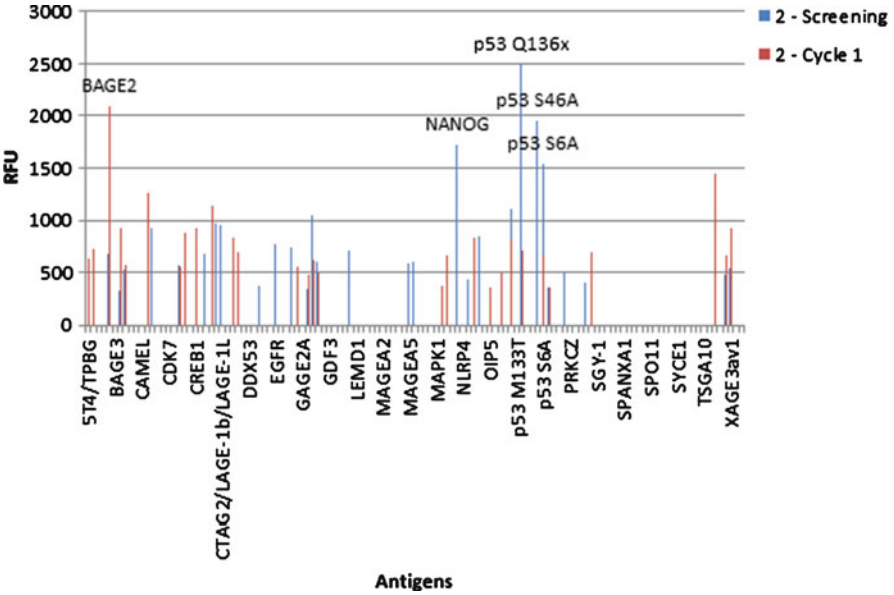


Fig. 3.8 Graph displaying the autoimmune profile of Patient 2 pre and post experimental treatment

time points for a given patient, simple statistical calculations (e.g. *t*-tests or ANOVA) – based on the triplicate repeat data for each antigen on each array – can be performed to determine whether the intensity at one time point is significantly different to that at a later time point. As expected, the number, identity and strength of anti-CT antigen autoimmune responses varied from patient to patient across our cohort, as exemplified in Figs. 3.7 and 3.8. In addition, when comparing the autoimmune profile of Patient 1–2, it is evident that Patient 1 has a noisier array, with nearly all antigens lighting up between 0 and 500 RFU, while Patient 2 has a cleaner array, with approximately only 30 antigens of interest showing strong signals above 500 RFU. The biological rationale underlying this observation remains unclear as yet, but it is noteworthy that all signals shown in Figs. 3.7 and 3.8 are significant compared to background. It is also noteworthy that we have previously verified individual anti-CT antigen responses by Western blot and have also compared our microarray data to ELISA data for individual antigens where available: in all cases, such comparisons have provided independent verification of the specificity of our protein microarray data (data not shown) (Beeton-Kempen et al. submitted). Furthermore, we have also demonstrated that our CT100 microarray platform shows linearity of response to anti-CT antigen autoantibody titres across 3–4 orders of magnitude and that it has a limit of quantitation in the range of 100 pg/ml (data not shown) (Beeton-Kempen et al. submitted).

The freely available *MultiExperiment Viewer* (MeV; v4.8.1) software was used to cluster the above data using the *K*-means algorithm using the Pearson correlation

(see Fig. 3.9). N.B. MEV was utilised solely for clustering purposes and not for normalisation or other data adjustments.

The resultant heatmap (Fig. 3.9) provides a compact visualisation of the data, which facilitates data interpretation and comparison between patients.

3.9 Conclusions

In this chapter we have highlighted the importance of both experimental and informatic protocols to minimise the influence of variations within microarray data sets due to non-specific binding, smears, artefacts or high background resulting from ineffective washing of the array surface, any of which can confound interpretation of data from a protein microarray experiment. Standardising protein microarray workflows and methodologies is essential to allow comparison of data generated at different times by the same or different laboratories.

We have emphasised the deterministic role of an experimental design and suggested the importance of the selection of specific controls, to enable normalisation of data and to assess background binding effects. With the development of appropriate preprocessing and quality control pipelines for raw array data, qualitative clustering of patient samples based on their autoantibody profiles measured on such antigen arrays becomes viable. We note that characterisation of anti-CT antigen autoantibody profiles from a single patient provides qualitative data on CT antigen expression, while comparison of anti-CT antigen autoantibody profiles across a timecourse for a given patient enables quantitative data to be generated on changes in autoantibody titres (Casiano et al. 2006). Factor analysis allows an unsupervised approach to cluster samples and provides a qualitative measure of how profiles correlate to a given cluster while also enabling the detection of outlier profiles within clusters. Compared to the *K*-means method, factor analysis is more straightforward, since it directly identifies the number of potential clusters and does not proceed by iterative steps. Factor analysis also provides more reproducible results, which is not always the case with *K*-means clustering, where final clustering might depend on the initial cluster assignment – which is randomly generated – and a sufficient number of iterations. In our experience, the limited number of control spots typically printed onto custom antigen arrays proved to be a challenge when determining the appropriate controls to generate a robust normalisation hypothesis, particularly since serum compositions are known to be strongly patient dependent. Therefore, the future utility of custom antigen microarrays in cancer biomarker discovery research will rely strongly on their design, which is ultimately linked to the available knowledge on the research question being addressed in each particular experiment. Despite these challenges though, the fast evolving protein microarray field has already enabled the identification of several serum antigens and antibody markers in cancer, although it is worth noting that validation of such candidate cancer biomarkers has in general yet to be confirmed (Matarraz et al. 2011).

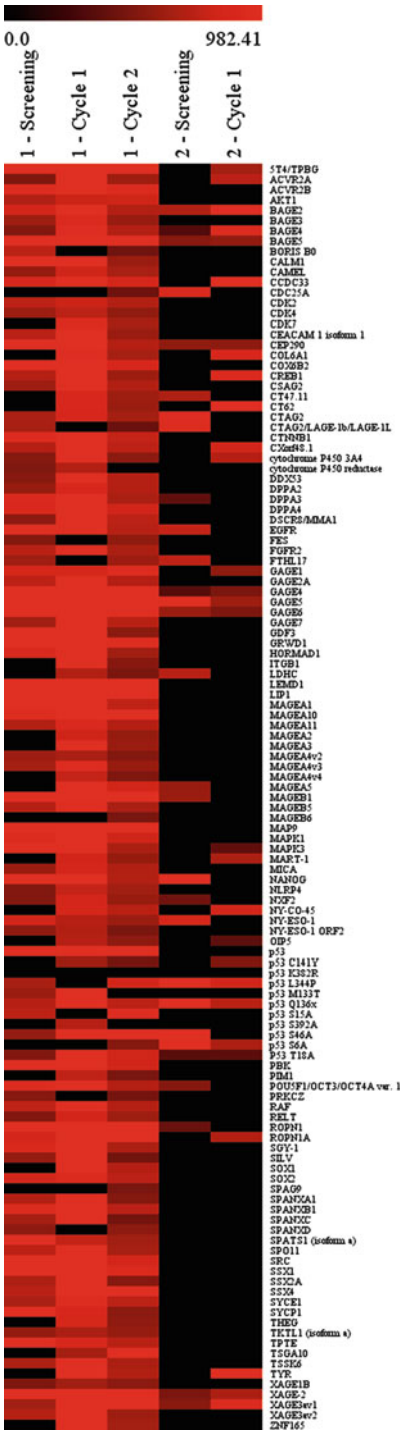


Fig. 3.9 Heatmap of Patient 1 and 2 autoimmune profiles at two time points following *K*-means clustering

Acknowledgements The authors thank Dr Aubrey Shoko, Dr Natasha Beeton-Kempen and Dr Judit Kumuthini for their help in generating the data herein. We thank the Centre for Proteomic & Genomic Research, Cape Town, for access to equipment and assistance in developing the CT100 array. JMB thanks the National Research Foundation (NRF), South Africa, for a Research Chair. The research was supported by grants from the NRF, University of Cape Town (UCT) and Marion Beatrice Waddel.

Supplementary Material

Methodology

Cloning of Cancer/Testis Antigen Genes

In total, 100 proteins were cloned and expressed for printing on the CT100 array. Seventy-two of these were CT antigen proteins, while the remaining twenty-eight were other cancer-associated proteins/proteins of interest. All antigens were cloned into baculoviral expression vectors and expressed in insect cells.

The following procedure was carried out for insect cell-expressed proteins. The gene encoding the *E. coli* biotin carboxyl carrier protein (BCCP) domain – amino acids 74–156 of the *E. coli accB* gene (Athappilly and Hendrickson 1995; Chapman-Smith and Cronan 1999) – was amplified by PCR from an *E. coli* genomic DNA preparation and cloned downstream of a viral polyhedrin promoter in an E. coli vector to create the transfer vector pJB1. This *E. coli* transfer vector system is derived from pTriEx-1.1 (Novagen). Flanking this *polh*-BCCP expression cassette were the baculoviral 603 and 1,629 genes (Zhao et al. 2003), which enabled the subsequent homologous recombination of the construct into a replication-deficient baculoviral genome.

Synthetic genes for each of the antigens of interest were obtained from Origene, Open Biosystems or GeneService. PCR primers were designed for each CT antigen cDNA such that the stop codon would be removed, enabling it to be cloned into the pJB1 transfer vector upstream of and in-frame with the 3′-BCCP tag via ligation-independent cloning methods, replacing the ORF region between the *Spe* I and *Nco* I sites of pJB1 in the process (all primers synthesised by IDT, UK) (Yang et al. 1993). Each resulting transfer vector thus encoded an individual antigen fused to a C-terminal BCCP tag.

The PCR amplification of each synthetic gene; ligation-independent cloning of these products into a BCCP tag-containing transfer vector, pJB30; and transformation of this vector into *E. coli* DH5α were all carried out according to standard recombinant DNA protocols (Sambrook et al. 2001) and are accordingly not described here in detail. Successful PCR amplification of each antigen was confirmed by gel electrophoresis, while successful cloning was determined by sequencing the relevant regions of each transfer vector (i.e. the region containing the ligated PCR

products as well as the junctions between these and the BCCP tags) according to standard protocols and verified against the RefSeq database.

Maintenance and Co-transfection of Sf21 Cells

A replication-incompetent baculovirus vector, bacmid pBAC10:KO_{1629} (Zhao et al. 2003), was propagated in *E. coli* HS996 cells, and bacmid DNA was prepared according to standard procedures. pBAC10/KO_{1629} was then linearised by restriction with *Bsu*361 (New England Biolabs) for 5 h at 37°C, after which *Bsu*361 was heat killed at 80°C for 15 min. Five hundred nanograms of undigested pJB1 transfer vector was then combined with 500 ng linearised bacmid, and the total volume made up to 12 μl with water. Twelve microlitres Lipofectin (diluted 2:1 in H_2O) was then added to this DNA mix, and the tube was incubated at room temperature for 30 min. One millilitre serum-free media (InsectXpress, Lonza) was added to the Lipofectin/DNA mixture. A 6-well plate containing 1×10^6 Sf21 cells/well (Invitrogen) was prepared and incubated at 27°C for 1 h to allow the cells to adhere. Excess media were aspirated from the Sf21 cells and replaced with the Lipofectin/DNA/serum-free mix. The transfected cells were incubated at 27°C overnight. The media was then replaced with 2 ml InsectXpress media supplemented with 2% FBS and incubated at 27°C without agitation for a further 72 h. Cells were resuspended by physical agitation and then pelleted by centrifugation at 1,000 × *g* for 10 min. The supernatant containing recombinant baculovirus was transferred to a fresh tube and stored at 4°C; this was the P_0 stock. The general baculoviral system used here was adapted from the work of Prof Ian Jones (Reading University, UK) (Zhao et al. 2003).

Recombinant baculoviral particles were amplified according to standard procedures. Briefly, a 6-well plate was set up with 1×10^6 Sf21 cells/well and incubated at 27°C for 1 h. Excess media were removed and replaced with 500 μl of P_0 virus plus 500 μl InsectXpress media supplemented with 2% FBS and incubated at 27°C without agitation for a further 72 h. P_1 virus was harvested as described above. A 150-ml tissue culture flask was seeded with 20 ml of 1×10^6 Sf21 cells/ml and incubated at 27°C for 1 h. Excess media were removed and replaced with 500 μl of P_1 virus plus 3 ml InsectXpress media supplemented with 2% FBS and incubated at 27°C for 1 h, after which a further 25 ml InsectXpress media supplemented with 2% FBS were added and cells incubated without agitation for 72 h. P_2 virus was harvested as described above. The titre of the P_2 viral stock was determined by a SybrGreen-based quantitative PCR assay versus a stock of known titre determined by plaque assay. Stocks that were found to have low titre were re-amplified.

Expression of BCCP-Tagged CT Antigens

A 24-well deep well plate containing 6×10^6 Sf21 cells/well suspended in 3 ml InsectXpress media supplemented with 2% FBS and 50 μM biotin was used. 200 μl

of P_2 virus was added, and the plate was incubated at 27°C for 72 h with agitation. Cells were harvested by centrifugation of the plate prior to lysis. Cells were gently resuspended and washed in 3 ml of PBS buffer for 5 min, the plate was recentrifuged and the supernatant was discarded; this was repeated three times in total. Pellets were gently resuspended in 350 µl of freezing buffer (25 mM HEPES, 50 mM KCl, pH 7.5) ensuring thorough mixing of the cells. Cells were aliquoted in 50 µl volumes and stored at −80°C until required for cell lysis. For cell lysis, aliquots were thawed and 50 µl lysis buffer (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 250 U/ml protease inhibitor cocktail and 1 mM DTT plus 10 U Benzonase (Novagen)) was added to each; this was then incubated on ice with agitation for 30 min. Cell debris was removed by centrifugation at 13,000 × *g* for 30 min at 4°C, the supernatant collected and then stored on ice for up to 24 h prior to printing.

The protein concentration of the soluble, crude protein extract was determined by Bradford assay (Bradford 1976) to confirm that effective cell lysis had occurred. Antigen expression and biotinylation were analysed by Western blot according to standard protocols (Sambrook et al. 2001). Antigen expression was confirmed using a mouse anti-c-myc antibody (Sigma-Aldrich) at 1:5,000 followed by a 1:25,000 dilution of goat anti-mouse IgG HRP conjugate (KPL). For more rapid processing, dot blots were sometimes used (same conditions as for Western blots) to assess expression prior to array fabrication. Biotinylation of the antigens was confirmed using a streptavidin–HRP conjugate probe (GE Healthcare) at 1:10,000.

Fabrication of Protein Microarrays

Preparation of Streptavidin-Coated Slides for Printing

A Nexterion Slide H microarray slide (Schott, Germany) was equilibrated to room temperature and removed from the foil package. A 1 mg/ml streptavidin solution was made up in 150 mM of Na_2HPO_4 buffer (pH 8.5). The microarray surface was immersed in approximately 5 ml of the streptavidin solution for 1 h at room temperature. The slide was removed from the streptavidin solution (which can be reusable successively up to 10 times) and then washed for 1 h at room temperature in 10 ml of 150 mM Na_2HPO_4 buffer (pH 8.5) containing 50 mM of ethanolamine to deactivate any remaining amine-reactive groups. The slide was washed for 3 × 5 min in 10 ml wash buffer and then for 5 min in 10 ml water. The slide was then placed in a 50-ml Falcon tube and centrifuged at 1,000 × *g* for 5 min at 20°C until dry. Streptavidin-coated slides were placed into slide boxes, sealed in Ziploc bags and stored at −20°C.

As a QC test, one streptavidin-coated slide per batch was incubated for 1 h with a solution of Cy5-biotinylated BSA (10 µg/ml in PBS), washed and scanned; this demonstrated that with this procedure, we can readily achieve CVs of 2–3% across the print area of the slide surface, judged by analysis of a virtual grid of 576 evenly distributed spots.

CT Antigen Microarray Fabrication

The expression and biotinylation of the various antigens were confirmed using SDS-PAGE- or dot blot based Western blot analysis prior to printing and crude lysates were then diluted with PBS containing 40% sucrose (sucrose was included to increase the surface tension and to reduce spreading of printed droplets). Forty microlitres of the crude protein extract for each BCCP-tagged protein to be arrayed was transferred into individual wells of a 384-well V-bottom plate. The plate was centrifuged at 4,000 rpm for 2 min at 4°C to pellet any cell debris that may have carried over from cell lysis. The plate was then stored on ice prior to the microarray print run, and during printing it was kept at 4°C.

Replica CT100 arrays were printed in a 4-plex format (i.e. 4 replica arrays per slide), using crude cell lysates. Each of the 72 CT antigens and the 28 TA antigens were printed in triplicate within each array. Several different controls were also included in each array. The positive controls included 50 ng/μl biotinylated human IgG (Rocklands Immunochemicals Inc.). The negative controls included biotinylated 200 ng/μl sheep IgG (Rocklands Immunochemicals Inc.) and an 'empty vector' lysate control consisting of a crude insect cell lysate containing the BCCP-tag alone with no recombinant fusion partner. In addition, three different concentrations (5, 10 and 15 ng/μl) of biotinylated Cy5-BSA were included in each sub-array for slide orientation and signal normalization purposes.

Each CT100 array was printed on home-made streptavidin-coated microarray slides (prepared as above) using a Genetix QArray2 robotic arrayer (Genetix Ltd., UK) equipped with 8 × 300 μm flat-tipped solid pins. Each array was printed as a set of eight 7 × 7 blocks, with each block printed by a different pin. The printing procedures were carried out at room temperature, while the source plate was kept at 4°C, and the atmosphere in the print chamber was humidified to ~50%. The arrays were printed using the following key QArray2 settings: inking time = 500 ms, microarraying pattern = 7 × 7, 500 μm spacing, maximum stamps per ink = 1, number of stamps per spot = 2, printing depth = 150 μm, water washes = 60 s wash and 0 s dry, ethanol wash = 10 s wash and 1 s dry.

After printing, each slide was washed for 30 min with 50 ml prechilled blocking solution (25 mM HEPES pH 7.5, 20% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA, 1 mM DTT and 50 μM biotin) and then stored at −20°C submerged in storage buffer (25 mM HEPES pH 7.5, 50% glycerol, 50 mM KCl, 0.1% Triton X-100, 0.1% BSA and 1 mM DTT).

Verification of Immobilisation of BCCP-Tagged Proteins to Array Surface

Following standard protocols for Western blots, it is possible to verify the successful immobilisation of biotinylated proteins to the array surfaces, as follows. Mouse anti-c-myc antibody was diluted 1:1,000 in 1 ml PBST containing 5% fat-free milk powder. The protein array was removed from wash buffer and equilibrated in PBST at room temperature for 5 min. The PBST was drained away and 5 ml antibody solution

was added to the array, which was then incubated with gentle agitation at room temperature for 30 min. The array was washed for 3×5 min with 1 ml of PBST. Goat anti-mouse antibody-HRP conjugate was diluted 1:1,000 in 1 ml milk/PBST. The antibody solution was added to the array, and the array was incubated with gentle agitation at room temperature for 30 min. The array was washed for 3×5 min with 1 ml of PBST and then submerged in 5 ml of chemiluminescent detection reagents (Pierce). After 1 min, the slide was placed in a 50-ml Falcon tube and centrifuged for 30 s to dry. In a dark room, the array was placed against autoradiography film for varying lengths of time before developing the film.

References

Altman N. Replication, variation and normalization in microarray experiments. Appl Bioinformatics. 2005;4:1–23.

Anderson KS, LaBaer J. The sentinel within: exploiting the immune system for cancer biomarkers. J Proteome Res. 2005;4:1123–33.

Angenendt P, Glökler J, Sobek J, Lehrach H, Cahill DJ. Next generation of protein microarray support materials: evaluation for protein and antibody microarray applications. J Chromatogr. 2003;1009:97–104.

Athappilly FK, Hendrickson WA. Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. Structure. 1995;3:1407–19.

Beeton-Kempen N, Duarte JG, Shoko A, Safari Serufuri J-M, Cebon J, Blackburn JM. Monitoring melanoma patient responses to therapeutic vaccination using a cancer/testis antigen protein microarray. Manuscript submitted.

Berrade L, Garcia AE, Camarero JA. Protein microarrays: novel developments and applications. Pharm Res. 2011;28:1480–99.

Blackburn JM, Shoko A. Protein function microarrays for customised systems-oriented proteomic analysis. In: Korf U, editor. Protein microarrays: methods and protocols, Methods in molecular biology. Springer protocols. New York: Humana Press; 2011. Chapter 21. ISBN 978-1-61779-285-4.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.

Boutell JM, Hart DJ, Godber BLJ, Kozlowski RZ, Blackburn JM. Functional protein microarrays for parallel characterisation of p53 mutants. Proteomics. 2004;4:1950–8.

Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005;6:331–43.

Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem. 1976;72:248–54.

Brusic V, Marina O, Wu CJ, Reinherz EL. Proteome informatics for cancer research: from molecules to clinic. Proteomics. 2007;7:976–91.

Büssow K, Konthur Z, Lueking A, Lehrach H, Walter G. Protein array technology: potential use in medical diagnostics. Am J Pharmacogenomics. 2001;1:1–7.

Casiano CA, Mediavilla-Varela M, Tan EM. Tumor-associated antigen arrays for the serological diagnosis of cancer. Mol Cell Proteomics. 2006;5:1745–59.

Causton HC, Quackenbush J, Brazma A. Microarray gene expression data analysis: a beginners guide. 1st ed. Malden: Blackwell Publishing; 2004.

Chapman-Smith A, Cronan JE. The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity. Trends Biochem Sci. 1999;24:359–63.

Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Prac Assess Res Eval. 2005;10:1–9.

Draghici S. Data analysis tools for DNA microarrays. 2nd ed. Boca Raton: Chapman & Hall; 2003.
Espina V, Mehta AI, Winters ME, Calvert V, Wulfkuhle J, Petricoin III EF, et al. Protein microarrays: molecular profiling technologies for clinical specimens. Proteomics. 2003;3: 2091–100.
Fang Y, Lahiri J, Picard L. G protein-coupled receptor microarrays for drug discovery. Drug Discov Today. 2003;8:755–61.
Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev. 2003;2:566–80.
Freudenberg JM. Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. Leipzig Bioinformatics Working Paper. 2005;3:1–120.
Gray MR, Colot HV, Guarente L, Rosbash M. Open reading frame cloning: identification, cloning, and expression of open reading frame DNA. Proc Natl Acad Sci U S A. 1982;79:6598–602.
Hall DA, Ptacek J, Snyder M. Protein microarray technology. Mech Ageing Dev. 2007;128:161–7.
Hardiman G. Microarray technologies – an overview. Pharmacogenomics. 2003;4:251–6.
Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 1st ed. New York: Springer; 2001.
He M, Taussig MJ. Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). Nucleic Acids Res. 2001;29:73–3.
Hultschig C, Kreutzberger J, Seitz H, Konthur Z, Bussow K, Lehrach H. Recent advances of protein microarrays. Curr Opin Chem Biol. 2006;10:4–10.
Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:211–15.
Ingvarsson J, Larsson A, Sjo AG, Truedsson L, Jansson B, Borrebaeck CAK, et al. Design of recombinant antibody microarrays for serum protein profiling: targeting of complement proteins research articles. J Proteome Res. 2007;6:3527–36.
Klein JB, Thongboonkerd V. Overview of proteomics. In: Thongboonkerd V, Klein JB, editors. Proteomics in nephrology. Basel: Karger; 2004. p. 1–10.
Kodadek T. Protein microarrays: prospects and problems. Chem Biol. 2001;8:105–15.
Koopmann J-O, Blackburn J. High affinity capture surface for matrix-assisted laser desorption/ ionisation compatible protein microarrays. Rapid Commun Mass Spectrom. 2003;17:455–62.
Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
Lu T, Costello CM, Croucher PJP, Häsler R, Deuschl G, Schreiber S. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005;6:1–13.
Macbeath G. Protein microarrays and proteomics. Nat Genet. 2002;32:526–32.
MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. Science. 2000;289:1760–3.
Matarraz S, González-González M, Jara M, Orfao A, Fuentes M. New technologies in cancer. Protein microarrays for biomarker discovery. Clin Transl Oncol. 2011;13:156–61.
Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, et al. Analyzing antibody specificity with whole proteome microarrays. Nat Biotechnol. 2003;21:1509–12.
Oshlack A, Emslie D, Corcoran LM, Smyth GK. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. Genome Biol. 2007;8:2.1–8.
Phizicky E, Bastiaens PIH, Zhu H, Snyder M, Fields S. Protein analysis on a proteomic scale. Nature. 2003;422:208–15.
Ploner A, Miller LD, Hall P, Bergh J, Pawitan Y. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. BMC Bioinformatics. 2005;6:1–20.
Predki PF. Functional protein microarrays: ripe for discovery. Curr Opin Chem Biol. 2004;8:8–13.
Quackenbush J. Computational analysis of microarray data. Genetics. 2001;2:418–27.
Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, Lau AY, et al. Self-assembling protein microarrays. Science. 2004;305:86–90.
Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24:971–83.
Robinson WH. Antigen arrays for antibody profiling. Curr Opin Chem Biol. 2006;10:67–72.

Safari Serufuri J-M. Development of computational methods for Custom protein arrays analysis. A case study on a 100 protein (“ CT100 ”) cancer/testis antigen array. Masters thesis, University of Cape Town. 2010.
Sambrook J, Russel DW, Macallum P. Molecular cloning – a laboratory manual. 3rd ed. Cold Spring Harbour: Cold Spring Harbour Laboratory Press; 2001.
Sanchez-Carbayo M. Antibody arrays: technical considerations and clinical applications in cancer. Clin Chem. 2006;52:1651–9.
Scanlan MJ, Gure AO, Old LJ, Chen Y-t. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev. 2002;188:22–32.
Schäferling M, Nagl S. Optical technologies for the read out and quality control of DNA and protein microarrays. Anal Bioanal Chem. 2006;385:500–17.
Schmidt DMZ, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, et al. Evolutionary potential of (b/a)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. Biochemistry. 2003;42:8387–93.
Schweitzer B, Predki P, Snyder M. Microarrays to characterize protein interactions on a whole-proteome scale. Proteomics. 2003;3:2190–9.
Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31:265–73.
Steinhoff C, Vingron M. Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform. 2006;7:166–77.
Tecan LS™ Series Laser Scanner: how to set the correct gain in the LS scanner. http://www.tecan.com
Tryfos P. Notes on Factor analysis. 2010. http://www.yorku.ca/ptryfos/f1400.pdf
Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature. 2003;423:506–11.
Wikipedia. Factor analysis in psychometrics. 2010. http://en.wikipedia.org/wiki/Factor_analysis
Wilson DL, Buckley MJ, Helliwell CA, Wilson IW. New normalization methods for cDNA microarray data. Bioinformatics. 2003;19:1325–32.
Wise E, Yew WS, Babbitt PC, Gerlt JA, Rayment I. Homologous (b/a)8 -barrel enzymes that catalyze unrelated reactions: orotidine 5′-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase. Biochemistry. 2002;41:3861–9.
Wolf-Yadlin A, Sevecka M, MacBeath G. Dissecting protein function and signaling using protein microarrays. Curr Opin Chem Biol. 2009;13:398–405.
Yang Y-S, Watson WJ, Tucker PW, Capra JD. Construction of recombinant DNA by exonuclease recession. Nucleic Acids Res. 1993;21:1889–93.
Zhao Y, Chapman DAG, Jones IM. Improving baculovirus recombination. Nucleic Acids Res. 2003;31:1–5.
Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, et al. Analysis of yeast protein kinases using protein chips. Nat Genet. 2000;26:283–9.
Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, et al. Global analysis of protein activities using proteome chips. Science. 2001;14(293):2101–5.
Zhu X, Gerstein M, Snyder M. ProCAT: a data analysis approach for protein microarrays. Genome Biol. 2006;7:110.

Jonathan Blackburn, D.Phil, Professor, South Africa Prof. Blackburn currently holds the South African Research Chair in Applied Proteomics & Chemical Biology and is head of the African Network for Drugs and Diagnostic Innovation (ANDI) Centre of Excellence in Proteomics and Genomics. He previously held a Royal Society University Research Fellowship at the Department of Biochemistry, Cambridge University, and is an EPSRC visiting research fellow at the University of Manchester.

He was founder and Research Director of the Centre for Proteomic and Genomic Research in Cape Town, was the academic founder and chief scientific officer of a UK biotechnology company, Sense Proteomic Ltd, and was chief scientist of Procognia Ltd. He obtained his DPhil degree in chemistry from the University of Oxford under the supervision of Prof. Sir Jack Baldwin, FRS, and carried out postdoctoral research at the Medical Research Council UK with Prof. Sir Alan Fersht, FRS. Prof. Blackburn serves in a number of national and international committees including the National Health Research Committee (South Africa) and the Biotechnology Subcommittee of the International Union of Pure and Applied Chemistry. He sits on the editorial advisory boards of the Journal of Proteome Research, Journal of Proteome Science and Computational Biology, and Expert Review of Proteomics. His academic expertise ranges from mechanistic enzymology, protein biochemistry, molecular biology, and proteomics to the creation of novel biomolecules by in vitro evolution. He is currently particularly interested in applications of protein microarray and mass spectrometry technologies in diagnostic marker discovery and validation, in the high-throughput study of protein-drug interactions, as well as in studying the effects of polymorphic variation on protein function.

Chapter 4
Proteomics and Cancer Research

Elena Lopez Villar and William Chi-Shing Cho

Abstract Cancer presents high mortality and morbidity globally, largely due to its complex and heterogeneous nature as well as the lack of effective biomarkers. There is an urgent need to identify clinically relevant markers for better diagnosis, prognosis, monitoring treatment efficacy, and accelerating the development of novel targeted therapeutics. Proteomics study of cancer may identify and characterize functional proteins that drive the transformation of malignancy as well as discover biomarkers to detect early-stage cancer, predict prognosis, determine therapy efficacy, identify novel drug targets, and ultimately develop personalized medicine. The technology platforms for proteomics analysis have advanced considerably over the last few years. Driven by these advancements in technology, a number of potential biomarkers have been identified in the tissues, blood, and body fluids.

Keywords Biomarker • Cancer • MALDI • Mass spectrometry • Proteomics

4.1 Introduction

Cancer is not a single disease but rather an accumulation of several events, mainly genetic and proteomic, arising in a single cell over a long period. Therefore, a top priority in the cancer field is the identification of these events. This can be achieved by characterizing cancer-associated genes and their protein products.

E.L. Villar, Ph.D.
Department of Oncohematology of Children, Hospital Universitario Niño Jesús, Madrid, Spain
e-mail: elena.lopez.villar@gmail.com

W.C.-S. Cho, Ph.D., RCMP, Chartered Scientist (UK), FHKIMLS, FHKSMDS, FIBMS (UK) (✉)
Department of Clinical Oncology, Queen Elizabeth Hospital, Kowloon, Hong Kong
e-mail: williamcscho@gmail.com

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_4,

Identifying the molecular alterations which distinguish any particular cancer cell from a normal cell will ultimately help to define the nature and predict the pathologic behavior of that cancer cell. In addition, it will indicate the responsiveness to treatment of that particular tumor. Moreover, the understanding of the profile of molecular changes in any specific cancer will be extremely useful as the correlation of the resulting phenotype of that cancer with molecular events will become possible. As a general rule, achieving these goals and knowledge will provide an opportunity for discovering new biomarkers for early cancer detection and developing prevention approaches. Thus, this will also help us identify new targets for therapeutic development (Mandong and Ngbea 2011; Ectors and Verh 2011; Semiglazova et al. 2011; Ostapenko and Ostapenko 2011; Junker 2011; Ozkan et al. 2011).

Mass spectrometry (MS) technology includes methods and tools that enable research including, but not limited to, instrumentation, techniques, devices, and analysis tools. The identification and definition of the molecular profiles of cancer require the development and dissemination of high-throughput molecular analysis technologies as well as elucidation of all of the molecular species embedded in the genome and proteome of cancer and normal cells. Moreover, the main challenge in cancer control and prevention is early detection. This could then enable effective interventions and therapies contributing to reduction in mortality and morbidity. At a specific time, biomarkers serve as molecular signposts of the physiologic state of a cell. These signposts are the result of genes, their products (proteins), and other organic chemicals made by the cell. Biomarkers could prove to be vital for the identification of early cancer and those subjects at risk of developing cancer as a normal cell progresses through the complex process of transformation to a cancerous state (Wright et al. 2012; Zhong et al. 2012; Schirle et al. 2012; Bouwman et al. 2012; Catusse et al. 2011; López et al. 2011a, b, c). This chapter discusses ongoing research in proteomics and MS basic concepts to identify molecular signatures such as protein biomarkers and their relevance in cancer diagnoses.

4.2 Genes and Proteins in Eukaryote Cellular Regulation

It is well known that eukaryotic organisms include a wide area of species: plants, fungi, animals, and humans. Moreover, although we show great diversity from plants and fungi, we all share the same characteristics of biochemistry, cellular organization, and molecular biology. Furthermore, for many years cellular biology has been intensively studied by examining the DNA and RNA levels, and today the complete human genome has been mapped (Venter et al. 2001). The discovery of restriction enzymes as the tools to cut DNA and RNA initiated a whole new world of experiments, giving DNA/RNA research a boost (Meselson and Yuan 1968). However, neither DNA nor RNA gives us the entire picture of the cellular organism. Not all DNA is translated into RNA and the proteins formed by RNA translation can carry various posttranslational modifications (PTMs), which cannot be predicted from the study of RNA. Proteins constitute the

majority of the effector molecules in the cell and the study of proteins is therefore very important to achieve a better understanding of cellular activities.

In the organism, the proliferation and activities of cells need to be strictly regulated. Intercellular signaling determines the exact differential stage and functions of each cell, and defects in this signaling can result in abnormalities such as uncontrolled proliferation and loss of apoptosis leading to diseases such as cancer. Several diseases result from deviances from normal cellular behavior, which emphasizes the need to understand the cellular system and the dynamic of living cells. All cellular activities are due to biochemical communication within and between cells. The cells communicate by means of nucleic acids, steroids, fatty acid derivatives, retinoids, lipids, carbohydrates, proteins, etc. Proteins have a predominant role acting as catalysts in biochemical reactions, as signal integrators, transducing motion and forming large multifunctional complexes. They are involved in processes such as cell proliferation, metabolism, development, and defense. Therefore, the elucidation of protein structures and functions is necessary for the understanding of the living cell (Bayley and Devilee 2012; Schuettengruber et al. 2011; Zhu et al. 2011; Guo et al. 2011; Nicholls et al. 2011).

4.3 Basic Concepts of Proteomics and MS

4.3.1 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions (Anderson and Anderson 1998; Blackstock and Weir 1999). Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term proteomics was first coined in 1997 to make an analogy with genomics (James 1997). The word proteome is a blend of protein and genome and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student (Wilkins et al. 1996; Thomson 1913). The proteome is the entire complement of proteins (Wilkins et al. 1996), including the modifications made to a particular set of proteins, and produced by an organism or system. This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes (Wilkins et al. 1996) (Fig. 4.1).

4.3.2 MS

MS implies a powerful tool in the study of chemical and biological compounds. By the measurement of the mass-to-charge ratio (m/z) of a sample molecule, exact details on the mass and structure can be obtained, which in turn can give the scientist useful information about the function, interactions, and regulation of the particular molecule. A century ago, Thomson developed the very first mass spectrometer (Thomson 1913).

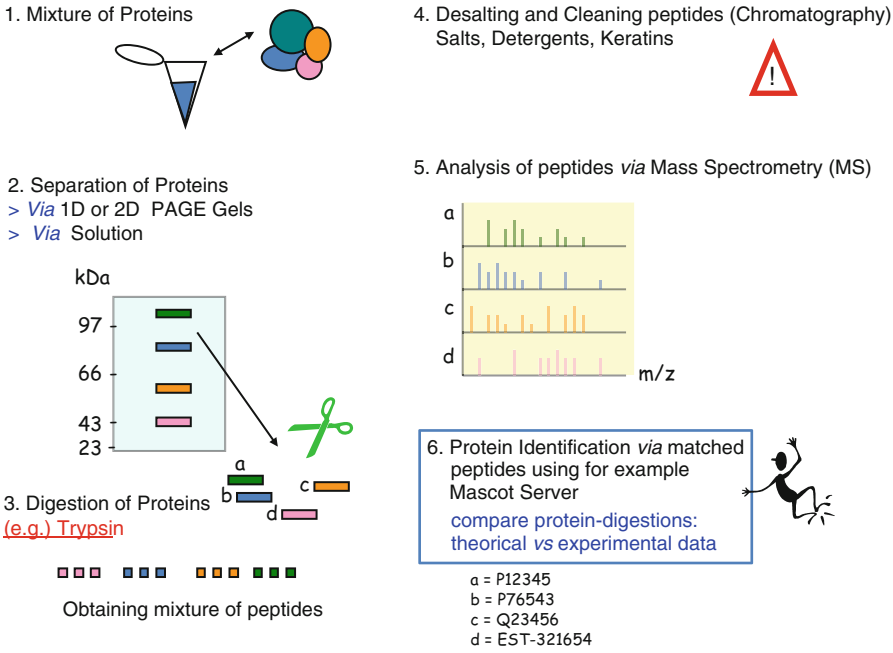


Fig. 4.1 Identifying proteins *via* mass spectrometry. The mixture of proteins (or just one protein) must be digested to obtain peptides. The resulting peptides have to be cleaned and desalted *via* chromatography (e.g., POROS R2) to avoid salts and detergents, which artifact the mass spectrometry analysis. Subsequently, the desalted and cleaned peptides are injected into the mass spectrometer. Finally, the matched peptides allow the identification of the proteins using databases (e.g., Mascot Server) (Courtesy Dr. K. Kjerno, Group Meetings 2007, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

A mass spectrometer instrument consists of (a) an ion source, in which the sample molecules are ionized; (b) a mass analyzer, which separates the ions according to their *m*/*z*; and (c) a detector, where the ions are registered according to the *m*/*z* values. Also, the mass spectrometer is connected to visualize the resulting spectra and data. Thus, proteomic strategies together with MS analyses can be used to perform clinical proteomics (Adams et al. 1988).

4.3.3 What Is a Mass Spectrum

A mass spectrum is a plot of an intensity *vs m*/*z* of a separated chemical collection. The mass spectrum of a given sample is the distribution pattern of the components of that collection, whether atoms or molecules, based on their *m*/*z*.

The *X*-axis of the plot is the *m*/*z* which is the quantity obtained by dividing the mass number of an ion by its charge number. For mass analyzers such as time-of-flight (TOF) (Bush and Lehman 1995), the direct *X*-axis measurement is the time series of the ions measured by the detector. For such cases, the spectra must be calibrated with known

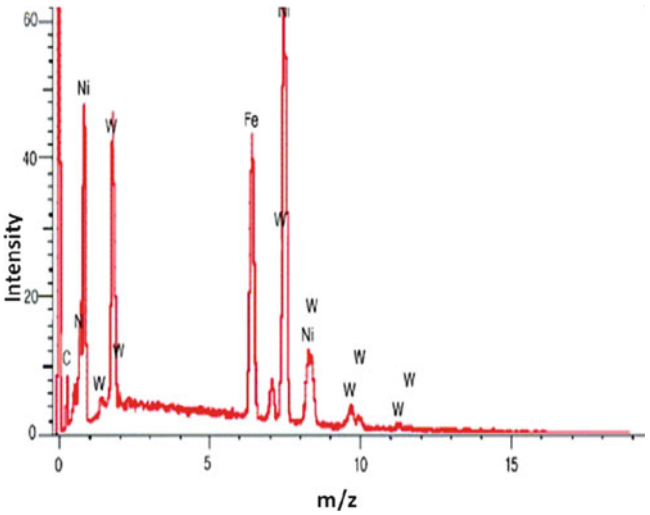


Fig. 4.2 This figure illustrates typical mass spectra. The mass spectrometer shows the mass-charge ratio (*m*/*z*) and the intensity of the detected ions according to the analyzed sample molecule (Courtesy Dr. K. Kjerno, Group Meetings 2007, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

standards in order to transform the *X*-axis from a time series into an *m*/*z*. The values for the standards are used to generate the parameters for the equations relating the time of flight to *m*/*z*. After these parameters are determined, the *m*/*z* for the unknown sample can be calculated from their TOF. For example, with the Fourier transform ion cyclotron resonance mass spectrometer (Marshall and Verdum 1990; Williams 1995), the frequency measurements gathered by the detector plates undergo fast Fourier transformation before they are mass calibrated (http://www.asms.org/whatisms/p5.html). Fourier transform ion cyclotron resonance MS is a type of mass analyzer (or mass spectrometer) for determining the *m*/*z* of ions based on the cyclotron frequency of the ions in a fixed magnetic field (Moorhouse and Sharma 2011; Carbonnelle and Nassif 2011; Tsybin et al. 2011; Allmer 2011; Vestal 2011) (Fig. 4.2).

The *Y*-axis of a mass spectrum represents the signal intensity of the ions and has arbitrary units. In most forms of MS, the signal intensity of an ion current does not represent relative abundance accurately but somewhat correlates loosely with it. Signal intensity is dependent on certain factors, the nature of the molecules being analyzed, how they ionize, the buffers' interaction, and the sample interaction (Caprioli et al. 1996).

4.3.4 Method of Interpretation of Mass Spectra

Samples can sometimes be highly difficult to analyze due to the amount of restrictions as well as variables that play roles in the output. Many factors can play roles in how a

mass spectrum is interpreted; these factors may include the following: even electron *vs* odd electron species, positive *vs* negative ion mode, and intact protein *vs* fragmented peptide ions.

Basis of resolution as well as peak height is reliant on the amount of sample being used and the amount of separation done prior to the mass spectroscopy. Based on certain heights and areas or peaks, structures can be determined.

Not all mass spectra can be interpreted similarly, due to the varying nature of the mass analyzers and ionization methods available. For example, some mass spectrometers break the analyte molecules into fragments; others observe the intact molecular masses with little fragmentation. A mass spectrum can represent many different types of information based on the type of mass spectrometer and the specific experimental conditions applied; however, all plots of intensity *vs* *m*/*z* are referred to as mass spectra (Gaskell 1986).

4.3.5 Normalization Techniques Used in MS

The process of removing statistical error from data created from repeated measurements is called normalization. In MS, normalization techniques are used to remove systematic biases from peptide samples. These biases can arise from various sources including protein degradation, measurement errors, and variation in loading samples. The common normalization techniques used on MS data require that the data is transformed from the linear to the log scale. Doing so allows the values to conform to the normal distribution and reduces the likelihood of masking more relevant proteins with less relevant ones (http://www.itl.nist.gov/div897/sqg/dads/HTML/euclidndstnc.html) (http://www.absoluteastronomy.com/topics/Normalization_statistics) (Waller 1972) (Fig. 4.3).

4.3.6 Annotation of Data

As described in the section above, there are a number of factors which are vital for meaningful interpretation of the data from a MS experiment. Without this metadata about the variables of the experiment, it is difficult to use a mass spectrum to generate an assessment. Combine this with the variance among types of mass spectrometers, and the utility of a reporting standard emerges.

In the vein of minimum information about a microarray experiment (MIAME; http://www.psidev.info/index.php?q=node/91; The Minimum Information about a Proteomics Experiment, Reporting guidelines for proteomics), the Human Proteome Project (HUPO, http://www.hupo.org/) proteomics standards initiative developed minimum information about a proteomics experiment in MS. The standard requires metadata to be recorded about general information such as the machine manufacturer and model, variables of the ion source for electrospray ionization and matrix-assisted

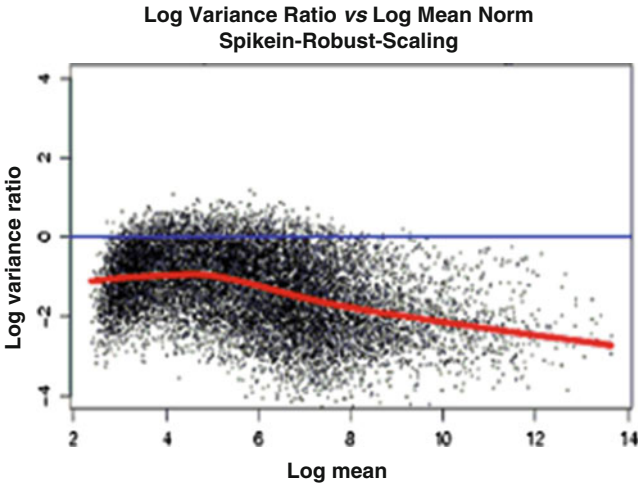


Fig. 4.3 This figure shows that the ratio of variance is on the *Y*-axis across a set of replicates after quartile normalization, divided by the variance of the scale-normalized values. The mean levels are on the *X*-axis. Both axes are on the log scale (Courtesy Dr. K. Kjerno, Group Meetings 2007, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

laser desorption/ionization (MALDI), mass analyzers and detectors involved, and post-processing involved in peak list generation and annotation. In addition to the mass spectrometer operation, post-processing methods that are often used in other fields of signal processing are frequently applied to mass spectra in order to generate a more useful spectrum.

Further work has been done to regulate the expression of mass spectroscopy data *via* the establishment of a controlled vocabulary. Developed by HUPO under the proteomics standards initiative, the controlled vocabulary provides ontology, comprised of applicable terms and transitions, which will better control the representation of MS data by regulating the vocabulary used in their description (McLafferty and Turecek 1993; Zimmer et al. 2006; Taylor et al. 2007).

4.3.7 MALDI

MALDI was invented in 1987 by Karas and Hillenkamp (1988). Later, Tanaka and coworkers (1988) applied the technique on a variety of biological macromolecules, which gave him the Nobel Prize in Chemistry for 2002.

MALDI is a technique commonly used in the analysis of large, nonvolatile biomolecules. When working with peptides, proteins, nucleotides, and oligosaccharides, MALDI is often the chosen ionization technique. Different parameters influence this ionization process, for example, laser wavelengths, pulse energies and lengths, matrix-analyte combinations and ratios, and salts, which all vary from experiment to experiment (Zenobi and Knochenmuss 1998).

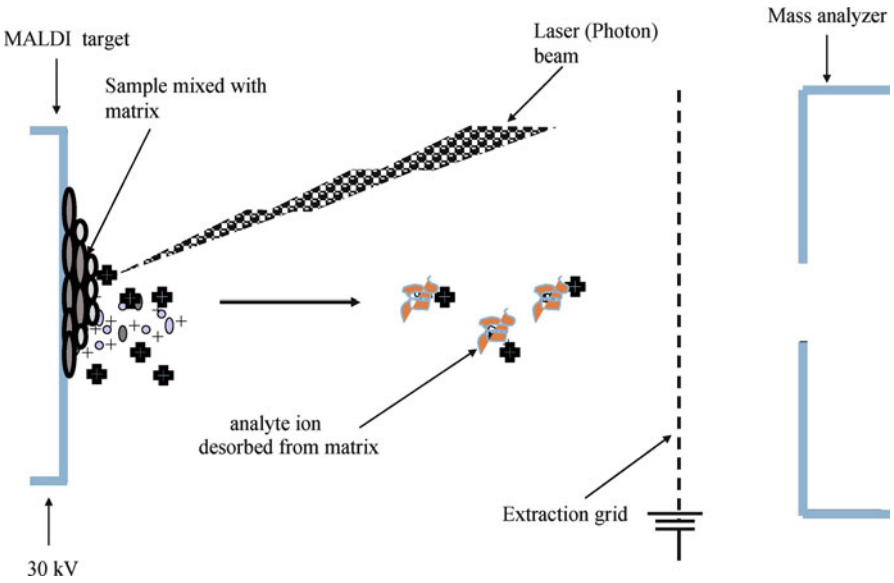


Fig. 4.4 The illustration of the principle of matrix-assisted laser desorption/ionization (*MALDI*) (Courtesy Dr. TE. Thingholm, Master Thesis 2005, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

The matrix molecules play a critical role in the ionization procedure in several ways: (a) they absorb laser light at a wavelength different from the analyte molecules; (b) the matrix molecules are added in large molecular excess, which separates the analyte molecules from each other and prevents analyte cluster formation; and (c) the matrix molecules function as a proton donor (in positive ion mode) and a proton acceptor (in negative ion mode) (Reflex II (Bruker) 1995) (Fig. 4.4).

The introduction of a matrix made it possible to desorb protein molecules of several hundred kilodalton (kDa) as intact ions as opposed to an upper limit of ~1,000 Da for biopolymers and 9,000 Da for synthetic polymers using LD-MS (Hillenkamp et al. 1991; Karas and Krüger 2003). Some of the most commonly used matrices are listed in Table 4.1 and their chemical structures are shown in Fig. 4.5.

Matrices are classified as either hot or cold matrices depending on the degree of fragmentation they cause to the sample. For example, 2,5-dihydroxybenzoic acid does not give rise to much fragmentation and is therefore considered to be a cold matrix, whereas α-cyano-4-hydroxycinnamic acid is a hot matrix (de Hoffmann and Stroobant 2001).

4.3.8 ESI

ESI is the method of choice when working with large biomolecules and highly complex samples due to its ability to be online coupled to liquid chromatography (LC).

Table 4.1 The most commonly used matrices for MALDI together with the chemical names and their applications

Matrix	Chemical name	Application
ATT	6-Aza-2-thiothymine	Oligonucleotides
Gentisic acid (DHB)	2,5-Dihydroxybenzoic acid	Peptides, proteins and carbohydrates
HPA or 3HPA	3-Hydroxypicolinic acid	Oligonucleotides
PA	Picolinic acid	Oligonucleotides
Sinapic acid (SA)	3,5-Dimethoxy-4-hydroxycinnamic acid	Higher mass biopolymers
Trihydroxyacetophenone	2,4,6-Trihydroxyacetophenone	Oligonucleotides, peptides
α-Cyano (HCCA)	α-Cyano-4-hydroxycinnamic acid	Peptides, proteins and organic compounds

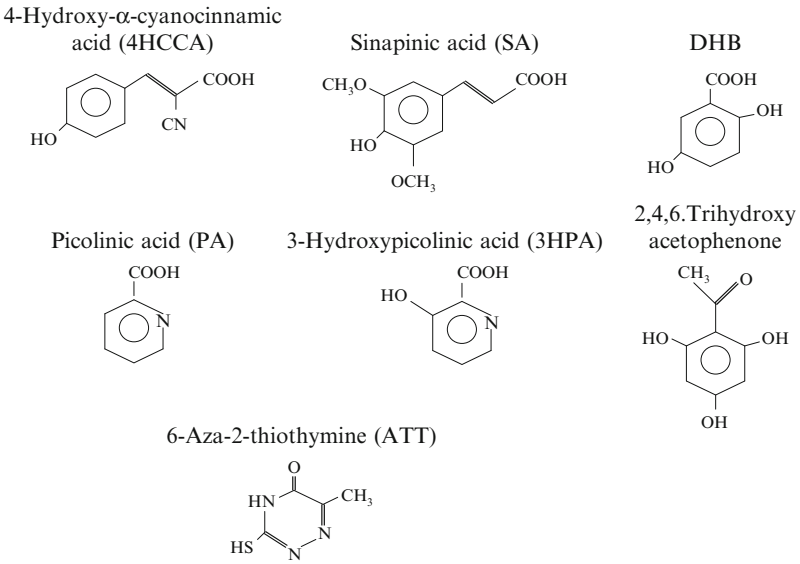


Fig. 4.5 The chemical structures of the matrices (Courtesy Dr. TE. Thingholm, Master Thesis 2005, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

In 1968, Dole introduced the concept of electrospray, and in 1988, Fenn and coworkers (Yamashita and Fenn 1984; Fenn et al. 1989) developed electrospray for ionization in MS. John Fenn received the Nobel Prize in 2003.

ESI is a liquid-phase ion source in which the nonvolatile analyte molecules are analyzed directly from the liquid phase. The analyte molecules are diluted in an aqueous solution, which is then forced through a fine capillary at a very low flow rate (0.1–10 μL/min). In positive ion mode, the analyte molecules are sprayed at low pH to improve the formation of positive ions, whereas higher pH conditions are used for negative ion mode. Proteins and peptides are usually analyzed using positive

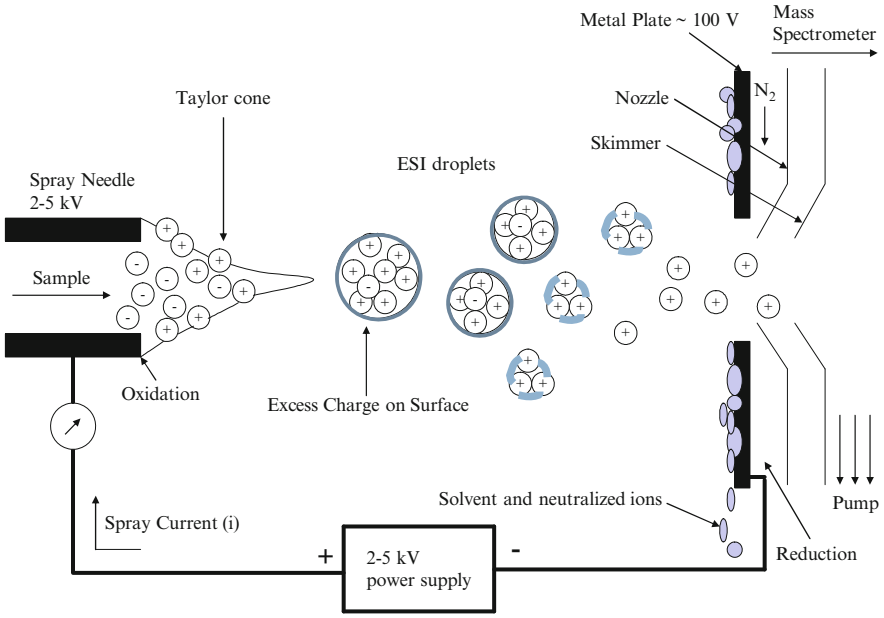


Fig. 4.6 The electrospray ionization process. The analyte solution is forced through the capillary, which has been supplied with high voltage. A Taylor cone is created due to the electric field between the capillary and the counter electrode, forming charged droplets of analyte ions and solvent. As these droplets travel toward the mass spectrometer, the solvent evaporates creating analyte ions. When the solution that comprises the Taylor cone reaches the Rayleigh limit, at which point the Coulombic repulsion of the surface charge is equal to the surface tension of the solution, charged droplets are formed at the tip of the capillary (Courtesy Dr. TE. Thingholm, Master Thesis 2005, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

ion mode, whereas oligonucleotides and oligosaccharides are analyzed using negative ion mode. The number of positive charges obtained by a molecule is related to the number of basic sites on the molecule; PTMs also influence this.

A high voltage (1–5 kV) is applied to the capillary creating an electric field gradient between the capillary and the counterelectrode. The charge of the voltage, whether it is positive or negative, depends on the analyte molecules being analyzed. The electric field gradient results in an accumulation of charge at the surface of the liquid, which forces the liquid to protrude from the tip of the capillary producing a Taylor cone. When the solution that comprises the Taylor cone reaches the Rayleigh limit, at which point the Coulombic repulsion of the surface charge is equal to the surface tension of the solution, charged droplets are formed at the tip of the capillary (Cech and Enke 2001) (Fig. 4.6).

These droplets carry an excess of positive or negative charge. The charged droplets are pulled toward the mass analyzer by the opposite charge at the counter electrode giving a fine spray of droplets (nebulization), meanwhile generating charged analyte molecules by a mechanism not yet fully understood. In 1968, Dole and coworkers (1968) suggested the coulomb fission mechanism, saying that the evaporation of the solvent from the formed droplets leads to an increase in the charge density. This causes

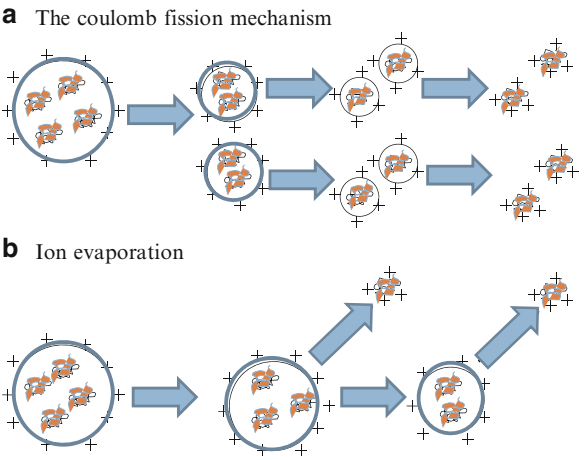


Fig. 4.7 *Section A* illustrates the coulomb fission mechanism proposed by Dole and coworkers in 1968, whereas *Section B* illustrates the ion evaporation mechanism proposed by Iribarne and Thomson in 1976 (Courtesy Dr. TE. Thingholm, Master Thesis 2005, Protein Research Group, PR Group of Odense University of Denmark, http://www.sdu.dk)

the droplets to split into smaller and smaller droplets producing free and charged analyte molecules (ions), which can then enter the mass analyzer (Fig. 4.7A). Iribarne and Thomson (1976) suggested an alternative mechanism known as ion evaporation in which the evaporation of solvent leads to an increase in the charge density of the droplets, which in turn causes Coulombic repulsion to overcome the liquid's surface tension, leading to the release of ions from the droplets' surfaces (Fig. 4.7B).

4.4 Sample Preparation: A Critical Step in Clinical Proteomic and MS Research Studies

When using proteomic and MS tools, sample preparation is one of the most crucial processes in proteomic analysis and biomarker discovery in solubilized samples. Chromatographic or electrophoretic proteomic technologies are also available for separation of cellular protein components. There are, however, considerable limitations in currently available proteomic technologies as none of them allows for the analysis of the entire proteome in a simple step because of the large number of peptides and because of the wide concentration dynamic range of the proteome in clinical blood samples. The results of any undertaken experiment depend on the condition of the starting material. Therefore, proper experimental design and pertinent sample preparation is essential to obtain meaningful results, particularly in comparative clinical proteomics in which one is looking for minor differences between experimental (diseased) and control (nondiseased) samples (Hernández-Borges et al. 2007; Guilak et al. 2005). Homogenization is one of the preparation steps employed for preparation of biological samples for proteomic analysis and includes such processes as mixing, stirring, dispersing, or emulsifying in order to change the sample's physical, but not chemical, properties. Homogenization for proteomics

incorporates five main categories: mechanical, ultrasonic, pressure, freeze-thaw, and osmotic/detergent lyses. Mechanical homogenization for tissues and cells can be accomplished by devices such as rotor-stator and open blade mills (e.g., Waring blender and Polytron), or pressure cycling technology such as French presses. Rotor-stator homogenizers can homogenize samples in volumes from 0.01 mL to l20 L depending on the tip and motor used. For optimum results, the tissue should be cut into slices, the size of which is slightly smaller than the diameter of the applied stator, as larger samples may clog the generator's inlet, making it impossible to achieve effective homogenization. Depending on the chemical resistance of a cutting tool, it is possible to homogenize samples under acidic or basic conditions in order to prevent degradation by endogenous enzymes. Heat transfer to the processed mixture is low to moderate and the process usually requires external cooling. Sample loss is minimal compared to pressure cycling technology, where by means of a pressure-generating instrument (Pressure Bioscience, West Bridgewater, MA), alternating cycles of high and low pressure are applied to induce cell lysis (Bodzon-Kulakowska et al. 2007; Rabilloud 1996).

Related to protein solubilization, proteins in biological samples are generally found in their native state associated with other proteins and often integrated as a part of large complexes, or into membranes. Once isolated, proteins in their native state are often insoluble. Breaking interactions involved in protein aggregation (e.g., disulfide hydrogen bonds, van der Waals forces, ionic and hydrophobic interactions) enables disruption of proteins into a solution of individual polypeptides, thereby promoting their solubilization. However, because of the great heterogeneity of proteins and sample-source-related interfering contaminants in biological extracts, simultaneous solubilization of all proteins remains a challenge. Integration of proteins into membranes and their association and complex formation with other proteins and/or nucleic acids hamper the process significantly. No single solubilization approach is suitable for every purpose, and each sample and condition requires unique treatment. Sample solubilization can be improved by agitation or ultrasonification, but an increase in temperature must be avoided. The selection of the appropriate solubilization protocol and buffers has especially been facilitated by the availability of commercial kits, although it is somewhat more expensive than routine reagent methods (Cañas et al. 2007; Görg et al. 2004).

To avoid protein modifications, aggregation, or precipitation resulting in occurrence of artifacts and subsequent protein loss, sample solubilization process necessitates the use in the sample buffer of (1) chaotropes (e.g., urea, thiourea, charged guanidine hydrochloride) that disrupt hydrogen bonds and hydrophilic interactions enabling proteins to unfold with ionizable groups exposed to solution; (2) ionic, nonionic, and zwitterionic detergents (SDS, CHAPS, or Triton X-100); (3) reducing agents that disrupt bonds between cysteine residues and thus promote unfolding of proteins (DTT/dithioerythritol, tributylphosphine, or tris-carboxy ethyl phosphine); and (4) protease inhibitors (Thadikkaran et al. 2005).

Although there is no general procedure to select an appropriate detergent, nonionic and zwitterionic detergents such as CHAPS and Triton X series are less denaturing than ionic detergents and have been used to solubilize proteins for

functional studies. On the other hand, ionic detergents are strong solubilizing agents that lead to protein denaturation. However, sodium cholate and deoxycholate are soft detergents compatible with native protein extraction, although variables like buffer composition, pH, salt concentration, temperature, and compatibility of the chosen detergent with the analytical MS procedure and how to remove it (e.g., by dialysis) are all crucial factors that need to be considered. Usually, tissue disruption and cell lyses require the combination of detergent and mechanical methodologies (Görg et al. 2004). The proper use of the above reagents, together with optimized cell disruption method, dissolution, and concentration techniques, collectively determines the effectiveness of proteome solubilization methodologies.

All the previously detailed information and also coupled to the use/study of blood, as a biospecimen in discovery research (a commonly used biospecimen which is highly complex and has a wide dynamic range of protein concentrations), makes it is very difficult to discover (measure) low-abundance proteins (potential biomarkers). One solution to this problem is to develop and apply nanotechnology in clinical proteomics as well as the throughput of analytical measurement systems while lowering their cost. Not only does nanotechnology have the potential of fulfilling many criteria required for the advancement of clinical proteomics, essential changes in the physicochemical properties of substances on their conversion to the nanostructured state, but it has also made it possible to create efficient systems for drug delivery to targets.

Moreover, blood cells offer unique insights into disease processes. Therefore, erythrocytes, granulocytes, monocytes, lymphocytes, and platelets are of special interest for clinical proteomics. Blood is a liquid organ and isolated blood cells reflect the environment and genome of the individual flow. Cytometry is currently widely used as an analytical tool for clinical cell analysis directly from anticoagulated whole blood and also for cell sorting to generate pure populations of cells from heterogeneous and highly integrated mixtures as are found in the majority of biological environments. Elispot, slide-based cytometry, and tissue arrays together with high-content screening microscopy are further upcoming techniques in cytoproteomics. The major challenge for this type of preanalytical standardization is related to the use of fresh samples, either for direct multiparameter analysis of cellular proteomics in whole blood or body fluids without preseparation or for cell sorting and enrichment strategies for subsequent proteomic and functional genomics analysis (Thadikkaran et al. 2005).

4.5 Relevance of Proteomics-MS Analyses in Clinical Research Studies

Improved biomarkers are of vital importance for cancer detection, diagnosis, and prognosis. While significant advances in understanding the molecular basis of disease are being made in genomics, proteomics will ultimately delineate the functional

units of a cell: proteins and their intricate interaction networks and signaling pathways in health and disease.

Much progress has been made to characterize thousands of proteins qualitatively and quantitatively in complex biological systems by use of multidimensional sample fractionation strategies, MS, and protein microarrays. Comparative/quantitative analysis of high-quality clinical biospecimen (e.g., tissue and biofluids) of human cancer proteome landscape can potentially reveal protein/peptide biomarkers responsible for this disease by means of their altered levels of expression, PTMs, as well as different forms of protein variants. Despite technological advances in proteomics, major hurdles still exist at every step of the biomarker development pipeline (Hassanein et al. 2011; Anderson 2005; Rifai et al. 2006; García-Foncillas et al. 2006; Bouchal et al. 2009; Wiener et al. 2004; Geiger et al. 2010; Anderson and Hunter 2006; Wang et al. 2009a; Lee et al. 2009; Pierobon et al. 2009; Ramachandran et al. 2008).

In the post-genome era, the field of proteomics incited great interest in the pursuit of protein/peptide biomarker discovery especially since MS demonstrated the capability of characterizing a large number of proteins and their PTMs in complex biological systems, in some instances even quantitatively. Technological advances such as protein/antibody chips, depletion of multiple high-abundance proteins by affinity columns, and affinity enrichment of targeted protein analytes, as well as multidimensional chromatographic fractionation, have all expanded the dynamic range of detection for low-abundance proteins by several orders of magnitude in serum or plasma, making it possible to detect the more abundant disease-relevant proteins in these complex biological matrices (Beirne et al. 2009; Kelleher et al. 2009; Wang et al. 2009b; Whiteaker et al. 2007; Ernoult et al. 2010; Nirmalan et al. 2010; Krishhan et al. 2009; Cha et al. 2010; Anderson et al. 2011). However, plasma- and cell-extract-based discovery research studies aimed to identify low-abundance proteins (e.g., some kinases) are extremely difficult. Therefore, it is necessary to develop significant technological improvements related to identifying these low-abundance, yet high biological impact, molecules. Moreover, if these protein kinases to be studied contain PTMs, it is important to know that spatial and temporal factors can decrease the efficiency of our study (e.g., many kinases are regulated by phosphorylation of the activation loop, which then directly reflects cellular kinase activity).

Furthermore, proteomics has been widely applied in various areas of science, ranging from the deciphering of molecular pathogenesis of diseases, the characterization of novel drug targets, to the discovery of potential diagnostic and prognostic biomarkers, where technology is capable of identifying and quantifying proteins associated with a particular disease by means of their altered levels of expression (Bateman et al. 2010; Kristiansen et al. 2008; An and Lebrilla 2010) and/or PTMs (Choudhary and Mann 2010; Madian and Regnier 2010; Iwabata et al. 2005) between the control and disease states (e.g., biomarker candidates). This type of comparative (semiquantitative) analysis enables correlations to be drawn between the range of proteins, their variations and modifications produced by a cell, tissue, and biofluids, and the initiation, progression, therapeutic monitoring, or remission of a disease state.

PTMs including phosphorylation, glycosylation, acetylation, and oxidation, in particular, have been of great interest in this field as they have been demonstrated as being linked to disease pathology and are useful targets for therapeutics.

In addition to MS-based large-scale protein and peptide sequencing, other innovative approaches including self-assembling protein microarrays (Ceroni et al. 2010) and bead-based flow cytometry (Wong et al. 2009) to identify and quantify proteins and protein-protein interaction in a high-throughput manner have furthered our understanding of the molecular mechanisms involved in diseases.

In summary, clinical proteomics has come a long way in the past decade in terms of technology/platform development, protein chemistry, and bioinformatics to identify molecular signatures of diseases based on protein pathways and signaling cascades. Hence, there is great promise for disease diagnosis, prognosis, and prediction of therapeutic outcome on an individualized basis. However, without correct study design and implementation of robust analytical techniques, the efforts and expectations to make biomarkers a useful reality in the near future can easily be hindered.

4.6 Tissue Biomarker Studies

Among the most common diseases worldwide, cancer remains a major threat to public health and there is an urgent need to identify novel biomarkers for diagnosis, prognosis, and prediction of response to anticancer treatment.

4.6.1 Diagnostic Tissue Biomarkers

Esophageal squamous cell carcinoma is among the top ten most frequent malignancies worldwide. Using the isobaric tags for relative and absolute quantitation (iTRAQ) approach, Pawar et al. (Pawar et al. 2011) have identified several novel protein biomarkers for esophagus squamous cell carcinoma, including PSAP, PLEC1, and PDIA4. These biomarker candidates were further validated to be overexpressed by immunohistochemical labeling using tissue microarrays.

Junrong et al. (2011) used two-dimensional gel electrophoresis (2-DE) coupled with electrospray ionization quadrupole TOF-MS/MS analysis to identify differentially expressed proteins among the hepatocellular carcinoma (HCC) tumor center, tumor margin, and non-tumorous liver tissues. Immunostaining suggested an increase tendency of CIB1 expression from non-tumorous liver tissue to tumor center, whereas knockdown of CIB1 expression by RNA interference led to the significant suppression of the cell growth in hepatoma cells. These data suggest that CIB1 may be used as a potential diagnostic factor and possibly an attractive therapeutic target for HCC.

Pancreas ductal adenocarcinoma (PDAC) is also a deadly malignancy with poor early diagnosis and no effective therapy. To explore the accessible proteins

overexpressed in PDAC, fresh human normal and PDAC tissues were ex vivo biotinylated, isolated, and analyzed using the two-dimensional (2D) nano-high-performance liquid chromatography (HPLC)-MS/MS method. TGFBI, LTBP2, and ASPN were found to be significantly upregulated in a large group of clinical PDAC samples compared to the corresponding normal and inflammatory tissues. These proteins bear the potential to be of clinical value for diagnostic and therapeutic applications in human PDAC (Turtoi et al. 2011).

Using label-free quantitative proteomics to analyze the insoluble fractions from colorectal cancer (CRC) patients, Yang et al. (2012) have identified a panel of protein markers (KRT5, JUP, TUBB, and COL6A1) for CRC. These proteins give specific network information for CRC, which may increase our understanding of the membrane environment in CRC and provide direction for diagnosis through molecular biomarker targeting.

There was also a comparative proteomic study on biopsies from patients with ovarian cancer to identify potential diagnostic biomarkers in tumor tissue. Evaluating by 2-DE and MS/MS analysis, calgranulin and S10A8 showed overexpressions in carcinoma tissue. These two proteins may serve as diagnostic biomarkers for ovarian cancer (Cortesi et al. 2011).

4.6.2 Prognostic Tissue Biomarkers

HCC is characterized by a multistage process of tumor progression. Differential tissue proteome analysis using LC-MS/MS has identified a cytoskeletal protein, TLN1, to be upregulated in HCC. The TLN1 expression levels in HCC nodules were significantly associated with the dedifferentiation of HCC, and TLN1 upregulation may be related to the higher rate of portal vein invasion in HCCs. These findings suggest that TLN1 may serve as a novel prognostic marker of HCC (Kanamori et al. 2011).

Another proteomics profile analysis of HCC tissues with different metastatic capabilities revealed that NDRG1 was correlated with metastasis and recurrence in HCC patients after liver transplantation. NDRG1-positive expression had poor prognosis compared with NDRG1-negative expression, either for shorter disease-free survival or overall survival. HCC cells in in vitro experiments with small interfering RNA against NDRG1 significantly suppressed its proliferation, colony formation, invasion, and migration ability. These findings suggest that NDRG1 is an important molecule in controlling HCC metastasis and thus may serve as a novel biomarker for predicting HCC recurrence after liver transplantation (Cheng et al. 2011).

There was a study comparing pairs of fresh frozen sections of Dukes B CRC and normal colorectal mucosa using a combination of 2-DE and HPLC-MS/MS. CH60, S10A9, and TCTP showed the greatest degree of significant overexpressions in primary CRC compared with normal colonic mucosa. A two-protein signature consisting of 14-3-3β and ALDH1 was identified as an independently prognostic marker in a multivariate model (O'Dwyer et al. 2011).

4.6.3 Predictive Tissue Biomarkers

Targeted therapies have optimal activity against a specific subset of tumors that depend upon the targeted molecule or pathway for growth, survival, or metastasis. Caccia et al. (2011) have found several cell-line secretome proteins (VASN, CD109, and BGH3) related to TGF-β signaling in thyroid cancer cells, which were sensitive to dasatinib and RPI-1 treatments. These secretome proteins may be potential biomarkers for measuring the effectiveness of thyroid cancer therapies.

Reversed-phase protein array is a promising technology for quick and simultaneous analysis of many patient samples allowing relative and absolute protein quantifications. Using reverse-phase protein array, Gonzalez-Angulo et al. (2011) have developed a ten-protein (ER, PR, Bcl2, GATA3, EIG121, EGFR, HER2, HER2p1248, CCNB1, and CCNE1) biomarker panel that could predict the recurrence-free survival and pathological complete response in breast cancer patients receiving neoadjuvant taxane and anthracycline-taxane-based systemic therapy.

4.7 Noninvasive Biomarker Studies

Noninvasive biomarkers, such as those from blood or urine, are ideal for cancer detection, stratification, monitoring, and prognosis.

4.7.1 Noninvasive Biomarkers for Cancer Diagnosis

Discovery of diagnostic biomarkers is the key to improve the early detection of human cancers. However, most biomarker studies use biological samples collected at or after diagnosis which are often advanced tumor stage. Opstal-van Winden et al. (2011) analyzed the pre-diagnostic breast cancer serum proteome with surface-enhanced laser desorption/ionization (SELDI)-TOF-MS, iTRAQ, and 2D nano-LC-MS/MS. AFAM, APOE, and isoform 1 of ITIH4 were found to be significantly higher, whereas A2MG and CERU were significantly lower in breast cancer. Their results showed that serum protein profiles were altered up to 3 years before breast cancer detection.

Remy-Martin et al. (2012) coupled surface plasmon resonance imaging with MALDI-TOF-MS in a hyphenated technique which established a new method to characterize potential cancer marker in human plasma. This new method had excellent analytical performance when detecting LAG3 protein in breast cancer, with its specificity >10 and reliability of 100% LAG3 identification with high significant mascot score >87.9. This rapid, collective, and automated on-chip MS analysis has many potential applications in proteomics.

Antibody-based proteomics provides a strategy for the systematic generation of antibodies against all human proteins to combine with protein profiling in cancer

samples using microarrays, immunohistochemistry, and immunofluorescence. Targeted antibody arrays are strongly contributing to the identification of protein cancer biomarker candidates and functional proteomic analyses (Pontén et al. 2011). Most HCC is generated from chronic hepatitis and cirrhosis. Hsieh et al. (2011) initiated their search in the interstitial fluid of tumor *via* differential gel electrophoresis and antibody arrays. Serum sERBB3 was identified as a marker for early HCC patients with chronic hepatitis and cirrhosis. Serum sERBB3 had a better performance than α-fetoprotein in the discrimination of early HCC from chronic hepatitis or cirrhosis; combination of sERBB3 and α-fetoprotein further improved the accuracy.

On the other hand, Ku86 has been found to be overexpressed in HCC patients. Nomura et al. (2012) assessed the diagnostic value of serum anti-Ku86 in the early detection of hepatitis C virus-related HCC patients. Serum levels of anti-Ku86 antibodies were significantly elevated in HCC patients compared to those in liver cirrhosis patients. Receiver operating characteristic analyses indicated that anti-Ku86 had a better sensitivity than α-fetoprotein, suggesting serum anti-Ku86 may be a potential biomarker for the early detection of hepatitis C virus-related HCC.

Analyzing the serum of gallbladder cancer by 2-DE and MALDI-TOF-MS, Tan et al. (2011) have found that the expressions of HPT and S10AA proteins were higher in gallbladder cancer compared to that in healthy volunteers. Further investigation revealed that patients with high expressions of HPT and S10AA were linked to late-stage disease and poor clinical prognosis. These two proteins may serve as new potential serum biomarkers for gallbladder cancer diagnosis and prognosis.

Targeted proteomics is an emerging technology that is playing an increasingly important role to facilitate disease biomarker development. Applying a selected reaction monitoring-based targeted proteomics platform to directly detect candidate biomarker proteins in plasma, three candidate proteins (GELS, LUM, and TIMP1) demonstrated an area under the curve value >0.75 in distinguishing pancreatic cancer from the chronic pancreatitis and healthy age-matched controls (Pan et al. 2012).

An MS-based plasma biomarker discovery workflow was also developed to facilitate biomarker discovery. Plasma from either healthy volunteers or pancreatic cancer patients was 8-plex iTRAQ labeled, fractionated by 2D reversed-phase chromatography, and subjected to MALDI-TOF/TOF-MS. Analysis of patient plasma prior to treatment identified PRDX2 with significant change in pancreatic cancer patients (Zhou et al. 2012).

Jayapalan et al. (2012) have profiled the sera of patients with prostate cancer and benign prostatic hyperplasia using the gel- and lectin-based proteomics methods. Significant differential expressions of APOA2, CO3 β chain fragment, ITIH4 fragment, TTHY, A1AT, and KNG1 light chain were demonstrated between the two groups of patients' samples. These data suggest the potential use of these serum proteins as complementary biomarkers to effectively discriminate prostate cancer from benign prostatic hyperplasia.

A 29-plex array platform was constructed to examine the gastric adenocarcinoma and control samples. Eleven proteins (EGFR, pro-APOA1, APOA1, TTHY, RANTES, D-dimer, VTNC, IL6, A2MG, CRP, and PAI1) were selected as classifiers

in two algorithms. These algorithms differentiated between the majority of gastric adenocarcinoma and control serum samples with >88% accuracy. They can supplement clinical gastroscopic evaluation of symptomatic patients to enhance diagnostic accuracy (Ahn et al. 2012).

Shield-Artin et al. (2012) combined immunodepletion, liquid-phase isoelectric focusing, one-dimensional differential in gel electrophoresis, MALDI-TOF-MS, and LC-MS/MS to identify differentially expressed proteins in the plasma of symptomatic ovarian cancer patients. They have identified several well-established biomarkers (CRP, HPT, A2MG, and A1A2) and two new ovarian cancer candidate biomarkers (MMRN2 and S10A9). The cancer-associated differential expressions of CRP and S10A9 were further confirmed by Western blot and ELISA.

Li et al. (2012) reported a refined workflow that combined ZipTip desalting, acetonitrile precipitation, HPLC separation, and MALDI-TOF-MS analysis for the profiling, purification, and identification of the targeted serum proteins found by SELDI-TOF-MS. Using this workflow, the isoforms of the human SAA family with or without truncations at their N-terminals were found in the sera of patients with various types of advanced-stage (stages III–IV) cancers. SAA is an acute-phase protein that is synthesized under the regulation of inflammatory cytokines during both acute and chronic inflammation (Liu 2012). It has been reported as a biomarker for various cancer types, such as nasopharyngeal (Cho et al. 2010), gastric (Sasazuki et al. 2010), lung (Cho et al. 2004), and renal cell (Vermaat et al. 2012) cancers.

4.7.1.1 Noninvasive Biomarkers for Cancer Prognosis

Egler et al. (2011) have developed an integrated proteomic approach to identify plasma biomarkers for high-risk neuroblastoma. Seven candidate biomarkers (SAA, APOA1, IL-6, EGF, MDC, sCD40L, and CCL11) were identified. These biomarkers were then used to create a multivariate classifier of high-risk neuroblastoma, which showed a specificity of 90% and a sensitivity of 81% for detecting post-diagnosis longitudinal samples that have active disease. Further validation of these biomarkers may improve the outcomes of patients by developing a simple blood test for the detection of relapse prior to the development of clinically evident disease.

The early detection of metastasis in CRC patients could improve their survival rate after curative surgery. Tsai et al. (2012) reported the use of Cy-dye labeling combined with multidimensional fractionation and MS as a proteomics-based approach for the identification of CRC metastasis-associated biomarkers in plasma samples collected from CRC patients upon diagnosis. The increased plasma levels of GELS were found in >80% of CRC patients with distal metastases, and a significant increase of GELS in plasma samples of stage IV *vs* stages I–III CRC patients was found before treatment.

The most reliable approach to diagnose ovarian cancer used to rely on multiple, time-consuming, and expensive investigative tools. Recently, serum protein biomarker HE4 has been approved by the Food and Drug Administration (United States) for monitoring recurrence or progression of epithelial ovarian cancer.

Reliable clinical evidence demonstrated that HE4 (used alone or in combination with CA125) substantially improved the accuracy of screening and/or disease monitoring of ovarian cancer (Montagnana et al. 2011).

Urine is a gold mine for biomarker discovery; nevertheless, with multiple proteins being in low amounts, urine proteomics becomes challenging. The study of Zoidakis et al. (2012) applied IMAC fractionation and LC-MS/MS for the discovery of urinary protein biomarkers for aggressive bladder cancer. PROF1 was found to be differentially expressed in the urine from patients with invasive to noninvasive bladder cancer and benign controls. By tissue microarray analysis, PROF1 was further found to have a strong correlation with poor prognosis and increased mortality.

4.7.2 Noninvasive Predictive Biomarkers

Prediction of the responses to neoadjuvant chemotherapy can improve the treatment of patients with advanced breast cancer. Hyung et al. (2011) have performed profiling of N-glycosylated proteins in the serum from advanced breast cancer patients to discover serum biomarkers of chemoresistance using a label-free LC-MS/MS method. They demonstrated that a multivariate classification model of six proteins (AHSG, APOB, C3, C9, CP, and ORM1) could predict responses to neoadjuvant chemotherapy and further predict relapse-free survival of patients.

4.8 Conclusive Remarks and Future Directions

Accessible, sufficient, and reliable biomarkers are desirable as they can reflect various states of the cancer cells and they can be used for diagnosis, prognosis, risk stratification, and therapeutic monitoring. Understanding proteomes and the protein-mediated interactions underlying their complexity and diversity is critical for the development of more reliable and robust diagnostic platforms, which are anticipated to enable personalized medicine.

Proteomics technologies not only provide insights into the heterogeneity of cancer but also open new avenues for treatment through the identification of signaling molecules in the proliferation and survival of neoplastic cells. Utilizing proteomics technologies can provide clues regarding tumor classification as well as identify clinical biomarkers and pathologic targets for the development of personalized treatments. The identified protein markers may guide clinical decision-making and advance diagnostic and therapeutic options for the treatment of various cancers. During the last few years, proteomics has facilitated biomarker discovery by coupling high-throughput techniques with novel nanosensors. It is enlightening to see the improvement of sensitivity and selectivity by using nano-proteomics approaches as novel sensors (Dasilva et al. 2012). On the other hand, the introduction of well-organized protein biomarker validation process can also accelerate the development of effective protein-based diagnostics.

Acknowledgements Special thanks for the support from International Society for Translational Medicine (www.istmed.org) and Professor Xiangdong Wang.

References

Adams F, Gijbels R, Van Grieken R. Inorganic mass spectrometry. New York: Wiley; 1988.

Ahn HS, Shin YS, Park PJ, Kang KN, Kim Y, Lee HJ, Yang HK, Kim CW. Serum biomarker panels for the diagnosis of gastric adenocarcinoma. Br J Cancer. 2012;106(4):733–9.

Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert Rev Proteomics. 2011;8(5):645–57.

An HJ, Lebrilla CB. A glycomics approach to the discovery of potential cancer biomarkers. Methods Mol Biol. 2010;600:199–213.

Anderson L. Candidate-based proteomics in the search for biomarkers of cardiovascular disease. J Physiol. 2005;563:23–60.

Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis. 1998;19(11):1853–61.

Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics. 2006;5:573–88.

Anderson KS, Sibani S, Wallstrom G, Qiu J, Mendoza EA, Raphael J, Hainsworth E, Montor WR, Wong J, Park JG, Lokko N, Logvinenko T, Ramachandran N, Godwin AK, Marks J, Engstrom P, Labaer J. Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer. J Proteome Res. 2011;10:85–96.

Bateman NW, Sun M, Hood BL, Flint MS, Conrads TP. Defining central themes in breast cancer biology by differential proteomics: conserved regulation of cell spreading and focal adhesion kinase. J Proteome Res. 2010;9:5311–24.

Bayley JP, Devilee P. The Warburg effect in 2012. Curr Opin Oncol. 2012;24(1):62–7.

Beirne P, Pantelidis P, Charles P, Wells AU, Abraham DJ, Denton CP, Welsh KI, Shah PL, du Bois RM, Kelleher P. Multiplex immune serum biomarker profiling in sarcoidosis and systemic sclerosis. Eur Respir J. 2009;34:1376–82.

Blackstock WP, Weir MP. Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol. 1999;17(3):121–7.

Bodzon-Kulakowska A, Bierczynska-Krzysik A, Dylag T, Drabik A, Suder P, Noga M, Jarzebinska J, Silberring J. Methods for samples preparation in proteomic research. J Chromatogr B Analyt Technol Biomed Life Sci. 2007;849(1–2):1–31.

Bouchal P, Roumeliotis T, Hrstka R, Nenutil R, Vojtesek B, Garbis SD. Biomarker discovery in low-grade breast cancer using isobaric stable isotope tags and two-dimensional liquid chromatography-tandem mass spectrometry (iTRAQ-2DLC-MS/MS) based quantitative proteomic analysis. J Proteome Res. 2009;8:362–73.

Bouwman J, Vogels JT, Wopereis S, Rubingh CM, Bijlsma S, van Ommen B. Visualization and identification of health space, based on personalized molecular phenotype and treatment response to relevant underlying biological processes. BMC Med Genomics. 2012;5:1.

Bush KL, Lehman TA. Guide to mass spectrometry. New York: VCH Publishers; 1995.

Caccia D, Zanetti Domingues L, Micciché F, De Bortoli M, Carniti C, Mondellini P, Bongarzone I. Secretome compartment is a valuable source of biomarkers for cancer-relevant pathways. J Proteome Res. 2011;10(9):4196–207.

Cañas B, Piñeiro C, Calvo E, López-Ferrer D, Gallardo JM. Trends in sample preparation for classical and second generation proteomics. J Chromatogr A. 2007;1153(1–2):235–58.

Caprioli RM, Malorni A, Sindona G. Mass spectrometry in the biomolecular sciences. Dordrecht: Kluwer; 1996.

Carbonnelle E, Nassif X. Applications of MALDI-TOF-MS in clinical microbiology laboratory. Med Sci (Paris). 2011;27(10):882–8.

Catusse J, Meinhard J, Job C, Strub JM, Fischer U, Pestsova E, Westhoff P, Van Dorsselaer A, Job D. Proteomics reveals potential biomarkers of seed vigor in sugarbeet. Proteomics. 2011;11(9):1569–80.

Cech NB, Enke CG. Practical implications of some recent studies in electrospray ionization fundamentals. Mass Spectrom Rev. 2001;20:362–87.

Ceroni A, Sibani S, Baiker A, Pothineni VR, Bailer SM, LaBaer J, Haas J, Campbell CJ. Systematic analysis of the IgG antibody immune response against varicella zoster virus (VZV) using a self-assembled protein microarray. Mol Biosyst. 2010;6:1604–10.

Cha S, Imielinski MB, Rejtar T, Richardson EA, Thakur D, Sgroi DC, Karger BL. In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. Mol Cell Proteomics. 2010;9:2529–44.

Cheng J, Xie HY, Xu X, Wu J, Wei X, Su R, Zhang W, Lv Z, Zheng S, Zhou L. NDRG1 as a biomarker for metastasis, recurrence and of poor prognosis in hepatocellular carcinoma. Cancer Lett. 2011;310(1):35–45.

Cho WC, Yip TT, Yip C, Yip V, Thulasiraman V, Ngan RK, Yip TT, Lau WH, Au JS, Law SC, Cheng WW, Ma VW, Lim CK. Identification of serum amyloid A protein as a potentially useful biomarker to monitor relapse of nasopharyngeal cancer by serum proteomic profiling. Clin Cancer Res. 2004;10(1 Pt 1):43–52.

Cho WC, Yip TT, Cheng WW, Au JS. Serum amyloid A is elevated in the serum of lung cancer patients with poor prognosis. Br J Cancer. 2010;102(12):1731–5.

Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11:427–39.

Cortesi L, Rossi E, Della Casa L, Barchetti A, Nicoli A, Piana S, Abrate M, La Sala GB, Federico M, Iannone A. Protein expression patterns associated with advanced stage ovarian cancer. Electrophoresis. 2011;32(15):1992–2003.

Dasilva N, Díez P, Matarraz S, González-González M, Paradinas S, Orfao A, Fuentes M. Biomarker discovery by novel sensors based on nanoproteomics approaches. Sensors (Basel). 2012;12(2):2284–308.

de Hoffmann E, Stroobant S. Mass spectrometry, principles and applications. 2nd ed. Chichester/New York: Wiley; 2001.

Dole M, Mack LL, Hines RL, Mobley RC, Ferguson LD, Alice MB. Molecular beams of macroions. J Chem Phys. 1968;49(5):2240–9.

Ectors N, Verh K. International and national initiatives in biobanking. Acad Geneeskd Belg. 2011;73(1–2):5–40.

Egler RA, Li Y, Dang TA, Peters TL, Leung E, Huang S, Russell HV, Liu H, Man TK. An integrated proteomic approach to identifying circulating biomarkers in high-risk neuroblastoma and their potential in relapse monitoring. Proteomics Clin Appl. 2011;5(9–10):532–41.

Ernoult E, Bourreau A, Gamelin E, Guette C. A proteomic approach for plasma biomarker discovery with iTRAQ labeling and OFFGEL fractionation. J Biomed Biotechnol. 2010;2010:927917.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246:64–72.

García-Foncillas J, Bandrés E, Zárate R, Remírez N. Proteomic analysis in cancer research: potential application in clinical use. Clin Transl Oncol. 2006;8:250–61.

Gaskell SJ. Mass spectrometry in biomedical research. Chichester: Wiley; 1986.

Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods. 2010;7:383–5.

Gonzalez-Angulo AM, Hennessy BT, Meric-Bernstam F, Sahin A, Liu W, Ju Z, Carey MS, Myhre S, Speers C, Deng L, Broaddus R, Lluch A, Aparicio S, Brown P, Pusztai L, Symmans WF, Alsner J, Overgaard J, Borresen-Dale AL, Hortobagyi GN, Coombes KR, Mills GB. Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics. 2011;8(1):11.

Görg A, Weiss W, Dunn MJ. Current two-dimensional electrophoresis technology for proteomics. Proteomics. 2004;4(12):3665–85.

Guilak F, Alexopoulos LG, Haider MA, Ting-Beall HP, Setton LA. Zonal uniformity in mechanical properties of the chondrocyte pericellular matrix: micropipette aspiration of canine chondrons isolated by cartilage homogenization. Ann Biomed Eng. 2005;33(10):1312–18.

Guo GS, Zhang FM, Gao RJ, Delsite R, Feng ZH, Powell SN. DNA repair and synthetic lethality. Int J Oral Sci. 2011;3(4):176–9.

Hassanein M, Rahman JS, Chaurand P, Massion PP. Advances in proteomic strategies toward the early detection of lung cancer. Proc Am Thorac Soc. 2011;8(2):183–8.

Hernández-Borges J, Borges-Miquel TM, Rodríguez-Delgado MA, Cifuentes A. Sample treatments prior to capillary electrophoresis-mass spectrometry. J Chromatogr A. 2007;1153(1–2):214–26.

Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem. 1991;63(24):1194A–201.

Hsieh SY, He JR, Yu MC, Lee WC, Chen TC, Lo SJ, Bera R, Sung CM, Chiu CT. Secreted ERBB3 isoforms are serum markers for early hepatoma in patients with chronic hepatitis and cirrhosis. J Proteome Res. 2011;10(10):4715–24.

Hyung SW, Lee MY, Yu JH, Shin B, Jung HJ, Park JM, Han W, Lee KM, Moon HG, Zhang H, Aebersold R, Hwang D, Lee SW, Yu MH, Noh DY. A serum protein profile predictive of the resistance to neoadjuvant chemotherapy in advanced breast cancers. Mol Cell Proteomics. 2011;10(10):M111.011023.

Iribarne JV, Thomson BA. On the evaporation of small ions from charged droplets. J Chem Phys. 1976;64(6):2287–94.

Iwabata H, Yoshida M, Komatsu Y. Proteomic analysis of organ-specific post-translational lysine-acetylation and -methylation in mice by use of anti-acetyllysine and -methyllysine mouse monoclonal antibodies. Proteomics. 2005;5:4653–64.

James P. Protein identification in the post-genome era: the rapid rise of proteomics. Q Rev Biophys. 1997;30(4):279–331.

Jayapalan JJ, Ng KL, Razack AH, Hashim OH. Identification of potential complementary serum biomarkers to differentiate prostate cancer from benign prostatic hyperplasia using gel- and lectin-based proteomics analyses. Electrophoresis. 2012;33(12):1855–62.

Junker A. Cancer treatment in the elderly. Med Monatsschr Pharm. 2011;34(12):462–4.

Junrong T, Huancheng Z, Feng H, Yi G, Xiaoqin Y, Zhengmao L, Hong Z, Jianying Z, Yin W, Yuanhang H, Jianlin Z, Longhua S, Guolin H. Proteomic identification of CIB1 as a potential diagnostic factor in hepatocellular carcinoma. J Biosci. 2011;36(4):659–68.

Kanamori H, Kawakami T, Effendi K, Yamazaki K, Mori T, Ebinuma H, Masugi Y, Du W, Nagasaka K, Ogiwara A, Kyono Y, Tanabe M, Saito H, Hibi T, Sakamoto M. Identification by differential tissue proteome analysis of talin-1 as a novel molecular marker of progression of hepatocellular carcinoma. Oncology. 2011;80(5–6):406–15.

Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. Anal Chem. 1988;60:2299–301.

Karas M, Krüger R. Ion formation in MALDI: the cluster ionization mechanism. Chem Rev. 2003;103:427–39.

Kelleher MT, Fruhwirth G, Patel G, Ofo E, Festy F, Barber PR, Ameer-Beg SM, Vojnovic B, Gillett C, Coolen A, Kéri G, Ellis PA, Ng T. The potential of optical proteomic technologies to individualize prognosis and guide rational treatment for cancer patients. Target Oncol. 2009;4:235–52.

Krishhan VV, Khan IH, Luciw PA. Multiplexed microbead immunoassays by flow cytometry for molecular profiling: basic concepts and proteomics applications. Crit Rev Biotechnol. 2009;29:29–43.

Kristiansen TZ, Harsha HC, Grønborg M, Maitra A, Pandey A. Differential membrane proteomics using 18O-labeling to identify biomarkers for cholangiocarcinoma. J Proteome Res. 2008;7:4670–7.

Lee J, Soper SA, Murray KK. Microfluidic chips for mass spectrometry-based proteomics. J Mass Spectrom. 2009;44:579–93.

Li J, Xie Z, Shi L, Zhao Z, Hou J, Chen X, Cui Z, Xue P, Cai T, Wu P, Guo S, Yang F. Purification, identification and profiling of serum amyloid A proteins from sera of advanced-stage cancer patients. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;889–890:3–9.

Liu C. Serum amyloid A protein in clinical cancer diagnosis. Pathol Oncol Res. 2012;18(2):117–21.
López E, Wesselink JJ, López I, Mendieta J, Gómez-Puertas P, Muñoz SR. Technical phosphoproteomic and bioinformatic tools useful in cancer research. J Clin Bioinforma. 2011a;1:26.
López E, López I, Sequí J, Ferreira A. Discovering and validating unknown phospho-sites from p38 and HuR protein kinases in vitro by phosphoproteomic and bioinformatic tools. J Clin Bioinforma. 2011b;1(1):16.
López E, López I, Ferreira A, Sequí J. Clinical and technical phosphoproteomic research. Proteome Sci. 2011c;9:27.
Madian AG, Regnier FE. Profiling carbonylated proteins in human plasma. J Proteome Res. 2010;9:1330–43.
Mandong BM, Ngbea JA. Cancer prevention strategies. Niger J Med. 2011;20(4):399–405.
Marshall AG, Verdum FT. Fourier transform in NMR, optical and mass spectrometry. Amsterdam: Elsevier; 1990.
McLafferty FW, Turecek F. Interpretation of mass spectra. 4th ed. Sausalito: University Science; 1993.
Meselson M, Yuan R. DNA restriction enzyme from *E. coli*. Nature. 1968;217(134):1110–14.
Montagnana M, Danese E, Giudici S, Franchi M, Guidi GC, Plebani M, Lippi G. HE4 in ovarian cancer: from discovery to clinical application. Adv Clin Chem. 2011;55:1–20.
Moorhouse MJ, Sharma HS. Recent advances in i-gene tools and analysis: microarrays, next generation sequencing and mass spectrometry. Indian J Biochem Biophys. 2011;48(4):215–25.
Nicholls C, Li H, Wang JQ, Liu JP. Molecular regulation of telomerase activity in aging. Protein Cell. 2011;2(9):726–38.
Nirmalan NJ, Hughes C, Peng J, McKenna T, Langridge J, Cairns DA, Harnden P, Selby PJ, Banks RE. Initial development and validation of a novel extraction method for quantitative mining of the formalin-fixed, paraffin-embedded tissue proteome for biomarker investigations. J Proteome Res. 2010;10:896–906.
Nomura F, Sogawa K, Noda K, Seimiya M, Matsushita K, Miura T, Tomonaga T, Yoshitomi H, Imazeki F, Takizawa H, Mogushi K, Miyazaki M, Yokosuka O. Serum anti-Ku86 is a potential biomarker for early detection of hepatitis C virus-related hepatocellular carcinoma. Biochem Biophys Res Commun. 2012;421(4):837–43.
O'Dwyer D, Ralton LD, O'Shea A, Murray GI. The proteomics of colorectal cancer: identification of a protein signature associated with prognosis. PLoS One. 2011;6(11):e27718.
Opstal-van Winden AW, Krop EJ, Kåredal MH, Gast MC, Lindh CH, Jeppsson MC, Jönsson BA, Grobbee DE, Peeters PH, Beijnen JH, van Gils CH, Vermeulen RC. Searching for early breast cancer biomarkers by serum protein profiling of pre-diagnostic serum; a nested case-control study. BMC Cancer. 2011;11:381.
Ostapenko V, Ostapenko A. Significance and specifics of surgical treatment in locally advanced breast cancer. Vopr Onkol. 2011;57(5):578–83.
Ozkan S, Ozkan M, Armay Z. Cultural meaning of cancer suffering. J Pediatr Hematol Oncol. 2011;33 Suppl 2:S102–4.
Pan S, Chen R, Brand RE, Hawley S, Tamura Y, Gafken PR, Milless BP, Goodlett DR, Rush J, Brentnall TA. Multiplex targeted proteomic assay for biomarker detection in plasma: a pancreatic cancer biomarker case study. J Proteome Res. 2012;11(3):1937–48.
Pawar H, Kashyap MK, Sahasrabuddhe NA, Renuse S, Harsha HC, Kumar P, Sharma J, Kandasamy K, Marimuthu A, Nair B, Rajagopalan S, Maharudraiah J, Premalatha CS, Kumar KV, Vijayakumar M, Chaerkady R, Prasad TS, Kumar RV, Kumar RV, Pandey A. Quantitative tissue proteomics of esophageal squamous cell carcinoma for novel biomarker discovery. Cancer Biol Ther. 2011;12(6):510–22.
Pierobon M, Calvert V, Belluco C, Garaci E, Deng J, Lise M, Nitti D, Mammano E, De Marchi F, Liotta L, Petricoin E. Multiplexed cell signaling analysis of metastatic and nonmetastatic colorectal cancer reveals COX2-EGFR signaling activation as a potential prognostic pathway biomarker. Clin Colorectal Cancer. 2009;8:110–17.

Pontén F, Schwenk JM, Asplund A, Edqvist PH. The Human Protein Atlas as a proteomic resource for biomarker discovery. J Intern Med. 2011;270(5):428–46.
Rabilloud T. Solubilization of proteins for electrophoretic analyses. Electrophoresis. 1996;17(5):813–29.
Ramachandran N, Raphael JV, Hainsworth E, Demirkan G, Fuentes MG, Rolfs A, Hu Y, LaBaer J. Next-generation high-density self-assembling functional protein arrays. Nat Methods. 2008;5:535–8.
Reflex II (Bruker) User´s gide. Bremen: Bruker-Franzen Analytik GMBH; 1995.
Remy-Martin F, El Osta M, Lucchi G, Zeggari R, Leblois T, Bellon S, Ducoroy P, Boireau W. Surface plasmon resonance imaging in arrays coupled with mass spectrometry (SUPRA-MS): proof of concept of on-chip characterization of a potential breast cancer marker in human plasma. Anal Bioanal Chem. 2012;404(2):423–32.
Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24:971–83.
Sasazuki S, Inoue M, Sawada N, Iwasaki M, Shimazu T, Yamaji T, Tsugane S. Japan Public Health Center-Based Prospective Study Group. Plasma levels of C-reactive protein and serum amyloid A and gastric cancer in a nested case-control study: Japan Public Health Center-based prospective study. Carcinogenesis. 2010;31(4):712–18.
Schirle M, Bantscheff M, Kuster B. Mass spectrometry-based proteomics in preclinical drug discovery. Chem Biol. 2012;19(1):72–84.
Schuettengruber B, Martinez AM, Iovino N, Cavalli G. Trithorax group proteins: switching genes on and keeping them active. Nat Rev Mol Cell Biol. 2011;12(12):799–814.
Semiglazova TI, Semiglazov VV, Filatova LV, Gershanovich ML, Chudenko VA, Latipova DK, Luk'ianchikova VS, Dashian GA, Paltuev RM. Novel target therapies used in breast cancer management. Vopr Onkol. 2011;57(5):592–600.
Shield-Artin KL, Bailey MJ, Oliva K, Liovic AK, Barker G, Dellios NL, Reisman S, Ayhan M, Rice GE. Identification of ovarian cancer-associated proteins in symptomatic women: a novel method for semi-quantitative plasma proteomics. Proteomics Clin Appl. 2012;6(3–4):170–81.
Tan Y, Ma SY, Wang FQ, Meng HP, Mei C, Liu A, Wu HR. Proteomic-based analysis for identification of potential serum biomarkers in gallbladder cancer. Oncol Rep. 2011;26(4):853–9.
Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom. 1988;2:151.
Taylor CF, Paton NW, Lilley KS, Binz PA, Julian Jr RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates 3rd JR, Hermjakob H. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007;25(8):887–93.
Thadikkaran L, Siegenthaler MA, Crettaz D, Queloz PA, Schneider P, Tissot JD. Recent advances in blood-related proteomics. Proteomics. 2005;5:3019–34.
Thomson JJ. Rays of positive electricity and their implication to chemical analysis. London: Longmans Green; 1913.
Tsai MH, Wu CC, Peng PH, Liang Y, Hsiao YC, Chien KY, Chen JT, Lin SJ, Tang RP, Hsieh LL, Yu JS. Identification of secretory gelsolin as a plasma biomarker associated with distant organ metastasis of colorectal cancer. J Mol Med (Berl). 2012;90(2):187–200.
Tsybin YO, Fornelli L, Kozhinov AN, Vorobyev A, Miladinovic SM. High-resolution and tandem mass spectrometry–the indispensable tools of the XXI century. Chimia (Aarau). 2011;65(9):641–5.
Turtoi A, Musmeci D, Wang Y, Dumont B, Somja J, Bevilacqua G, De Pauw E, Delvenne P, Castronovo V. Identification of novel accessible proteins bearing diagnostic and therapeutic potential in human pancreatic ductal adenocarcinoma. J Proteome Res. 2011;10(9):4302–13.
Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.

Vermaat JS, Gerritse FL, van der Veldt AA, Roessingh WM, Niers TM, Oosting SF, Sleijfer S, Roodhart JM, Beijnen JH, Schellens JH, Gietema JA, Boven E, Richel DJ, Haanen JB, Voest EE. Validation of serum amyloid α as an independent biomarker for progression-free and overall survival in metastatic renal cell cancer patients. Eur Urol. 2012;62(4):685–95. doi:10.1016/j.eururo.2012.01.020.

Vestal ML. The future of biological mass spectrometry. J Am Soc Mass Spectrom. 2011; 22(6):953–9.

Waller GR. Biochemical applications of mass spectrometry. New York: Wiley; 1972.

Wang H, Wong CH, Chin A, Kennedy J, Zhang Q, Hanash S. Quantitative serum proteomics using dual stable isotope coding and nano LC-MS/MSMS. J Proteome Res. 2009a;8:5412–22.

Wang P, Whiteaker JR, Paulovich AG. The evolving role of mass spectrometry in cancer biomarker discovery. Cancer Biol Ther. 2009b;8:1083–94.

Whiteaker JR, Zhang H, Eng JK, Fang R, Piening BD, Feng LC, Lorentzen TD, Schoenherr RM, Keane JF, Holzman T, Fitzgibbon M, Lin C, Zhang H, Cooke K, Liu T, Camp 2nd DG, Anderson L, Watts J, Smith RD, McIntosh MW, Paulovich AG. Head-to-head comparison of serum fractionation techniques. J Proteome Res. 2007;6:828–36.

Wiener MC, Sachs JR, Deyanova EG, Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal Chem. 2004;76:6085–96.

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez J-C, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser D. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and arnino acid analysis. Nat Biotechnol. 1996;14(1):61–5.

Williams R. Spectroscopy and the Fourier transform: an interactive tutorial. New York: VCH; 1995.

Wong J, Sibani S, Lokko NN, LaBaer J, Anderson KS. Rapid detection of antibodies in sera using multiplexed self-assembling bead arrays. J Immunol Methods. 2009;350:171–82.

Wright PC, Noirel J, Ow SY, Fazeli A. A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations. Theriogenology. 2012;77(4):738–765.e52.

Yamashita M, Fenn JB. Electrospray ion source. Another variation on the free-jet theme. J Phys Chem. 1984;88:4451–9.

Yang HY, Kwon J, Park HR, Kwon SO, Park YK, Kim HS, Chung YJ, Chang YJ, Choi HI, Chung KJ, Lee DS, Park BJ, Jeong SH, Lee TH. Comparative proteomic analysis for the insoluble fractions of colorectal cancer patients. J Proteomics. 2012;75(12):3639–53.

Zenobi R, Knochenmuss R. Ion formation in MALDI mass spectrometry. Mass Spectrom Rev. 1998;17:337–66.

Zhong Y, Hyung SJ, Ruotolo BT. Ion mobility-mass spectrometry for structural proteomics. Expert Rev Proteomics. 2012;9(1):47–58.

Zhou C, Simpson K, Lancashire LJ, Walker MJ, Dawson MJ, Unwin RD, Rembielak A, Price P, West C, Dive C, Whetton AD. Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery. J Proteome Res. 2012;11(4):2103–13.

Zhu X, Fan WG, Li DP, Kung H, Lin MC. Heme oxygenase-1 system and gastrointestinal inflammation: a short review. World J Gastroenterol. 2011;17(38):4283–8.

Zimmer JSD, Monroe ME, Qian WJ, Smith RD. Advances in proteomics data analysis and display using an accurate mass time tag approach. Mass Spectrom Rev. 2006;25(3):450–82.

Zoidakis J, Makridakis M, Zerefos PG, Bitsika V, Esteban S, Frantzi M, Stravodimos K, Anagnou NP, Roubelakis MG, Sanchez-Carbayo M, Vlahou A. Profilin 1 is a potential biomarker for bladder cancer aggressiveness. Mol Cell Proteomics. 2012;11(4):M111.009449.

Elena López Villar, Ph.D., Senior Scientist, Spain Dr. Elena López acquired an international patent (proteomics and clinical diagnoses) at CSIC (Consejo Superior de Investigaciones Científicas) during her master's thesis. Subsequently, she got doctor's degree at UCM (Universidad Complutense de Madrid) in collaboration with the Odense University of Denmark (Ph.D. Marie Curie) applying proteomics for microbiology and immunology studies. In addition, she was responsible for phosphoproteomic core at the CNIO (Centro Nacional de Investigaciones Oncológicas) proteomic unit.

Later, she obtained a postdoc fellowship in MICINN (Ministerio de Ciencia e Innovación) for proteomic clinical and cancer research at the Hospital Universitario 12 de Octubre. Currently, she is a researcher collaborating with the Hospital Universitario Niño Jesùs (clinical oncology proteomics research). During her education and experience, she published more than 20 international scientific articles, and she is, currently, an editorial member of the Board of the Clinical and Translational Medicine Journal.

William Chi-Shing Cho, Ph.D., RCMP, Chartered Scientist (UK), FHKIMLS, FHKSMDS, FIBMS (UK), Queen Elizabeth Hospital, Hong Kong Dr. William Chi-Shing Cho is a scientific officer at the Department of Clinical Oncology in Queen Elizabeth Hospital. His main research interests have been focusing on cancer studies utilizing high-throughput technologies to discover biomarkers for cancer diagnosis, treatment prediction, and prognostication. He is a Chartered Scientist granted by the Science Council (UK) and a fellow member of several institutes, including the Institute of Biomedical Science (UK), Hong Kong Institute of Biomedical Science, and Hong Kong Society for Molecular Diagnostic Sciences. Dr. Cho has published over 150 papers and plenty of books covering cancer biomarkers, proteomics, microRNAs, and Chinese medicine.

He serves as the editor in chief, editor, and associate editor of a number of international medical journals. Dr. Cho is also an international grant reviewer of the Hope Funds for Cancer Research (USA), Cancer Research UK, Science Foundation (Ireland), National Medical Research Council (Singapore), Academia Sinica Investigator Award (Taiwan), etc.

Chapter 5
Toward Development of Novel Peptide-Based Cancer Therapeutics: Computational Design and Experimental Evaluation

Elena Pirogova and Taghrid Istivan

Abstract Drug research and discovery are of critical importance in human health care. Computational approaches for drug lead discovery and optimization have proven successful in many recent research programs. These methods have grown in their effectiveness not only because of improved understanding of the basic science, the biological events and molecular interactions that define a target for therapeutic intervention, but also because of advances in algorithms, representations, and mathematical procedures for studying such processes. Advances in genomics and proteomics and development of new bioinformatics methods contribute greatly to the process of rational drug design, which can be a cost-effective solution to drug discovery. Peptides are emerging as a novel class of drugs for cancer therapy, and many efforts have been made to develop peptide-based pharmacologically active compounds. In this study, we present and discuss three novel bioactive peptide analogues, designed using the Resonant Recognition Model (RRM), and discuss their biological effects on normal and tumor cells from mouse and human origins.

Keywords Cancer • Drug design • Computational analysis • Cancer therapeutics • Peptides • Signal processing

E. Pirogova, Ph.D. (✉)
School of Electrical and Computer Engineering, RMIT University,
GPO Box 2476, Melbourne 3001, VIC, Australia

Health Innovations Research Institute, RMIT University,
Melbourne, VIC, Australia
e-mail: elena.pirogova@rmit.edu.au

T. Istivan, Ph.D.
Department of Biotechnology and Environmental Biology,
School of Applied Sciences, RMIT University,
Plenty Rd, Bundoora West, VIC, Australia

Health Innovations Research Institute, RMIT University,
Melbourne, VIC, Australia

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_5,

5.1 Introduction

Drug design is a time-consuming multistep experimental process that includes synthesis of a drug molecule followed by in vitro and in vivo evaluation of its pharmaceutical potency and efficacy for medical treatment. Often this intense discovery process could lead to a product with poor pharmacokinetic properties that can even produce toxic and adverse side effects in humans (Bidwell and Raucher 2009). Combinatorial chemistry is a powerful tool used by medicinal chemistry for the design of new drug candidates (Dolle 1998). Combinatorial methods allow generating a large number of compounds in a short period of time (compared to the traditional synthesis of a single compound). However, there is a need to select the products to be synthesized from the vast pool of those possible to be produced (Pickett et al. 2000). As the structures of more and more protein targets become available through crystallography and NMR analysis, computational methods can use the known structure of a protein target as a route to discover novel lead compounds. These *in silico* methods include de novo design, virtual screening, and fragment-based discovery.

The most powerful approach for designing drug-oriented peptides is hybridization of structure-based and combinatorial chemistry methodologies. This approach proposes to screen large quantities of drug candidates. The peptides are designed using a semi-rational process, the so-called evolutionary computation (Pickett et al. 2000).

The importance and broad functional role of peptides in life processes became apparent only in the 1950s and early 1960s, when the continuous development of increasingly sensitive analytical methods and techniques for isolation and purification started a new era in pharmacology (Porath and Flodin 1959; Peterson and Sober 1956; Craig and King 1958; Sewald and Jakubke 2002). Native peptides can be directly applied as pharmacologically active compounds only to a very limited extent. The major disadvantages of the application of a peptide in a biological system – for example, rapid degradation by proteases, hepatic clearance, undesired side effects by interaction of conformationally flexible peptides with different receptors, and low membrane permeability due to their hydrophilic character – prohibit the use of oral application in most cases because of their detrimental effects (Sewald and Jakubke 2002). However, peptide chemistry can contribute considerably to drug development. The interaction of a peptide or a protein epitope with a receptor or an enzyme is the initial event based on molecular recognition and generally elicits a biological response. Many efforts have been made to develop pharmacologically active peptide-based compounds, including peptide modification and the design of peptidomimetics (Sewald and Jakubke 2002). While modified peptides contain non-proteinogenic or modified amino acid building blocks, peptidomimetics are non-peptidic compounds that imitate the structure of a peptide in its receptor-bound conformation and – in the case of agonists – the biological mode of action on the receptor level.

According to the definition by Ripka and Rich (1998), three different types of peptidomimetics may be distinguished:

Type I: these are peptides modified by amide bond isosteres and secondary structure mimetics. These derivatives are usually designed to closely match the peptide backbone.

Type II: these are small non-peptide molecules that bind to a receptor or enzyme (functional mimetics). However, despite being often presumed to serve as structural analogues of native peptide ligands, these non-peptide antagonists often bind to a different receptor subsite and, hence, do not necessarily mimic the parent peptide.

Type III: these may be regarded as ideal mimetics, because they are non-peptide compounds and contain the functional groups necessary for the interaction of a native peptide with the corresponding protein (pharmacophoric groups) grafted onto a rigid scaffold. The design of all three types of peptidomimetics may be assisted by X-ray crystallographic or nuclear magnetic resonance (NMR) data, computational de novo design, and combinatorial chemistry.

In peptide-based drug design, it is most important to determine a specific peptide sequence with a high affinity binding to a particular protein surface. Solving a peptide binding problem involves finding a region on the protein surface suitable for peptide binding, finding the appropriate peptide for this region, and peptide refinement to enable stable binding which is required for inhibition. When the binding surface is known, a peptide can be designed de novo. In other cases, for a given peptide, the region on the protein's surface with the optimum binding conditions should be searched. Petsalaki et al. (2009) introduced a method based on a bioinformatics approach that could successfully find the binding sites for the peptides. A similar approach, the de novo molecular design computational tool ProLigand, was adopted by Frenkel et al. (1995). Known peptides were docked to unknown locations on given proteins by Hetenyi and Van der Spoel (2002) using AutoDock. There have been successful attempts for computational peptide design that use knowledge-based search strategies and diverse sets of statistical descriptors, different training databases, hydrophobicity scales, and motif regularities (Juretic et al. 2009). Automated peptide binding search techniques from known epitopes or protein libraries have been successfully used as bioinformatics tools (Mayrose et al. 2007; Moreau et al. 2008; Stein and Aloy 2008). There are different computational binding tools such as the sequence moment concept, artificial neural networks, fuzzy neural networks, and Hidden Markov Model for checking the suitability of inhibitory peptides for binding on MHC class II proteins (Unal et al. 2010). The suitability of a ligand as a drug was tested using Bayesian neural network analysis (Ajay et al. 1998). The application of genetic algorithms for the design of peptides has been an important line of research, examples of which are *in silico* peptide screening and the application of a genetic algorithm to determine an inhibitory peptide against the Parkinson's disease-related protein a-synuclein (Abe et al. 2007), peptides as thrombin inhibitors (Kamphausen et al. 2002; Riester et al. 2005), integer linear programming (Klepeis et al. 2004), design of hexapeptides against

stromelysin protein (Singh et al. 1996), and a peptide buildup approach in combination with a genetic algorithm (Budin et al. 2001a, b, c).

Currently, immunotherapy is at the forefront of experimental cancer therapies. This methodology utilizes the power of the immune system and its focused ability to destroy cancer cells (Bright and Franz 2002). Cancer vaccine development is becoming more complex and challenging with each advance in the field. It ranges from molecular characterization of candidate vaccine antigens or peptides to formulation of an optimal vaccine and to administration and monitoring of such a vaccine in appropriately designed clinical trials (Waldmann 2003). The idea of vaccinating to treat cancer, i.e., the administration of a therapeutic vaccine, is not new. For decades, researchers and clinicians have studied and debated the possibility of vaccinating against cancer (Old 1996). Vaccines consisting of peptides derived from the protein sequence of a candidate tumor-associated or specific antigens represent the tip of the anticancer vaccine spear (Bright and Franz 2002; Knutson et al. 2001). Additional essential signaling components in cancer cells are being discovered, and it has been shown that individual peptides can be derived to inhibit their function in a targeted fashion (Lev et al. 2004; McCarty 2004; Tortora et al. 2003). These peptides can be used for monotherapy or in combination with conventional chemotherapeutic agents. Since multiple pathways become dysfunctional when a tumor develops and cancer cells accumulate oncogenic mutations as they progress, the greatest advancement can be achieved by combining therapeutic agents, which address different hallmarks of cancer. This concept, called "multifocal signal modulation therapy" (MSMT), is a very promising approach, and researchers have demonstrated that combinations of signal modulators achieve dramatic suppression of tumor growth (Lev et al. 2004; McCarty 2004; Tortora et al. 2003).

Anticancer peptide therapy is an emerging field that uses bioactive therapeutic peptides to kill cancer cells. In the past 15–20 years, much effort has been directed to developing peptides capable of eliciting therapeutic responses in cells. Early work was pursued with the goal of using peptides as tools to probe the mechanisms and functional consequences of various protein–protein interactions, but it soon became apparent that peptides capable of mimicking or interfering with important intraprotein contacts could be useful as therapeutic molecules (Bidwell and Raucher 2009).

It is important to note that, as opposed to small molecule drugs, peptides are easily designed to target almost any protein of interest using "rational" methods. As the sequence, structure, and interaction partners of many oncogenic proteins are known, peptides can be designed to inhibit these interactions by using a sequence from the interaction domain. Peptides can be easily produced and their sequence easily modified using chemical synthesis or molecular biology techniques. However, the utilization of peptides for cancer therapy is limited at present by poor pharmacokinetic parameters and tumor deposition (Talmadge 1998; Lipka et al. 1996). When applied in vivo, peptides are rapidly degraded in circulation, and their relatively large size and often charged nature make them unable to penetrate cancer cell membranes. These limitations can be overcome through the use of nonnatural

amino acids or macromolecular carriers to enhance peptide stability and through the use of cell-penetrating peptides to increase membrane permeability (Bidwell and Raucher 2009).

Therapeutic peptides can be grouped into three classes: (1) peptides that interfere with proliferative signal transduction cascades, (2) peptides that arrest the cell cycle by modulating cyclin-dependent kinase activity, and (3) peptides that can directly induce apoptosis by modulating proteins that control apoptotic response (Bidwell and Raucher 2009).

In this study, the Resonant Recognition Model (RRM) was employed to design de novo the bioactive peptides mimicking the activity of three selected therapeutic proteins interleukin 12 (IL12), myxoma virus M-T5 protein (MV-T5), and tumor necrosis factor (TNF-α) known to play significant roles in cancer treatment. The cytotoxicity of the designed peptides was experimentally evaluated on mammalian tumor and normal cell lines, with the findings presented and discussed below.

5.2 Resonant Recognition Model

The Resonant Recognition Model (RRM) (Cosic 1995, 1997) essentially belongs to the field of protein and DNA sequence analysis. There are different approaches in this field that could be classified into four main areas:

- General statistical analysis (based on amino acid frequencies)
- Signal searches (searches intended to uncover potential structural elements of nucleotide or amino acid sequences)
- Similarity (homology) searches
- Structural predictions

However, the available methods can be characterized by a different predictive accuracy, especially in predicting biological function from sequence similarities. Similar sequences may appear in totally alien proteins as a result of chance or, occasionally, by convergent evolution of sequences with similar properties. There are a number of unexpected significant resemblances between functionally different proteins, and on the one hand, no sequence homology can be observed for functionally related peptides. The optimal alignment programs introduce certain improvements in sequence analysis intended to distinguish between analogous and homologous sequences, but they are still based on sequence similarities.

The RRM allows investigation of the periodicity of structural motifs with defined physicochemical characteristics, which determine biological properties of protein and DNA sequences. The RRM is based on signal processing methods and incorporates a physical property of nucleotides or amino acids (the electron–ion interaction potential) resulting in decoding of the sequence into a signal and transformation of the signal into a spectrum by Fourier or wavelet transform. Multiplication of such spectra extracts information that is found to correlate with the biological function of the analyzed molecules (Cosic 1994, 1995, 1997; Pirogova et al. 2002). The Fourier

transform is useful tool in sequence analysis as it allows easy manipulation with signals and extracting information out of a great number of sequences as well as prediction of the points in the sequence that could be the most important for the biological function of the analyzed molecule.

In the RRM (Cosic 1994, 1995, 1997; Pirogova et al. 2002), at first the protein primary structure, i.e., amino acid sequences, is presented as a numerical series by assigning to each amino acid a physical parameter value relevant to the protein's biological activity. A number of amino acid indices (437 have been published) have been found to correlate in some way with the biological activity of the whole protein. It was shown (Cosic 1994, 1995, 1997; Pirogova et al. 2002) that the highest correlation can be achieved with parameters related to the energy of delocalized electrons in each amino acid. This observation is explained by the fact that the electrons delocalized from the particular amino acid have the strongest impact on the electronic distribution of the whole protein (Cosic 1994, 1995, 1997; Pirogova et al. 2002). The energy of delocalized electrons, calculated as the electron–ion interaction pseudo-potential (EIIP) (Veljkovic and Slavic 1972) of each amino acid residue, is employed in the RRM for conversion of protein primary sequence into a numerical sequence. Thus, these numerical series represent the distribution of free electrons' energies along the protein.

At the second stage of the RRM calculation, the obtained numerical series are converted into a discrete Fourier spectrum, which carried the same information content about the arrangement of amino acids in the sequence as the original numerical sequence (Cosic 1994, 1995, 1997; Pirogova et al. 2002). A multiple cross-spectral function is defined and calculated to obtain the common frequency components from the spectra of a group of proteins. Peaks in such function denote common frequency components for all sequences analyzed (Cosic 1994, 1995, 1997; Pirogova et al. 2002). It was shown that one characteristic frequency characterizes one particular biological function or interaction (Cosic 1994, 1995, 1997; Pirogova et al. 2002). But this frequency can be regarded as characteristic feature of the particular biological function when the following criteria are met:

- Single peak only can be observed for a group of protein sequences with the same biological activity.
- If no significant or prominent peak in the spectrum can be seen, then these protein sequences are biologically unrelated.
- Different peak frequencies observed in cross spectrum correspond to different biological functions.

In our previous studies, it was shown that proteins with the same biological function have a common frequency in their numerical spectra. Hence, the calculated RRM frequency can be regarded as important feature for protein biological function or interaction. It was also proven that proteins and their targets (receptors, binding proteins, and inhibitors) display the same characteristic frequency in their interactions. Despite the fact that a protein and its target have different biological functions, they can participate in the same biological process, which is defined by the same RRM frequency (Cosic 1997). When this RRM characteristic

frequency is calculated, it is possible then to identify the individual amino acids in a protein sequence that contribute most to this specific characteristic frequency and, thus, to the observed protein's biological behavior. It is also possible to design *de novo* bioactive peptides having only the determined characteristic frequency and consequently the desired biological function (Cosic 1997; Pirogova et al. 2002).

5.3 Bioactive Peptide Design Using the RRM

It is possible to determine the RRM characteristic frequency from analysis of the power spectra of proteins. In addition, from the analysis of their phase spectra, we can identify the corresponding phase for a particular frequency. On the basis of determined RRM characteristic frequencies and phases for a particular group of protein sequences, we can design amino acid sequences (short peptides) having those specific characteristics related to a protein's biological function. It is expected that the designed peptide will exhibit the desired biological activity (Pirogova et al. 2002; Cosic and Pirogova 2007).

The strategy for design of such defined peptides is presented below:

1. The RRM characteristic frequency is determined from the multiple cross-spectral function for a group of protein sequences that share a common biological function (interaction).
2. The phases are calculated for the characteristic frequency or frequencies of a particular protein, which is selected as a parent for an agonist/antagonist.
3. The minimal length of the designed peptide is defined by the appropriate frequency resolution. An inverse Fourier transformation (IFT) is used to calculate a numerical sequence of different lengths, which exhibits the same prominent characteristic frequency as a parent protein.
4. To determine the amino acids that correspond to each element of the new numerical sequence defined above, the tabulated electron–ion interaction potential (EIIP) parameter values are used. The resulting new amino acid sequence represents the anticipated designed peptide (Cosic 1997; Pirogova et al. 2002; Cosic and Pirogova 2007).

In previous studies, the RRM approach was applied to structure–function analysis of basic fibroblast growth factor (bFGF) (Cosic et al. 1994). Property–pattern characteristics for biological activity and receptor recognition for a group of FGF-related proteins were defined and then used to aid the design of a set of peptides which can act as bFGF antagonists. Molecular modeling techniques were then employed to identify the peptide within this set with the greatest conformational similarity to the putative receptor domain of bFGF. The 16-mer peptide, which exhibits no sequence homology to bFGF, antagonized the stimulatory effect of bFGF on fibroblast thymidine incorporation and cell proliferation, but exerted no effect itself in these in vitro bioassays (Cosic et al. 1994).

The RRM was also successfully applied for the analysis of HIV envelope proteins. The interaction between HIV virus envelope proteins and CD4 cell surface antigen has a central role in the process of virus entry into the host cell. Thus, blocking the interaction between the envelope glycoproteins and CD4 surface antigen, known to be the HIV receptor, should inhibit infection (Krsmanovic et al. 1998). For this purpose, 6 peptides, each of 20 amino acids in length, were designed using the RRM methodology. To validate the RRM computational predictions, the activities of the designed peptides were evaluated experimentally. These investigations were performed initially by evaluating the reactivity and cross-reactivity of all designed peptides with their corresponding antibodies (Krsmanovic et al. 1998). The results obtained showed significant cross-reactivity to the polyclonal antibodies raised against peptides that share at least one characteristic frequency and phase at this frequency. The results provided an experimental confirmation of the concept that RRM frequency characteristics present important parameters associated with biomolecular recognition and in particular, the antibody–antigen recognition.

5.4 Design of Peptide Analogue with Anticancer Activity

Here we present the application of the RRM approach to structure–function analysis of 3 selected therapeutic proteins: interleukin 12, myxoma virus MT, and tumor necrosis factor, which are known to possess anticancer biological activities. The short linear bioactive peptides RRM-IL, RRM-MV, and RRM-TNF were designed to mimic murine IL12 (Pirogova et al. 2011), myxoma virus MT5 protein (Istivan et al. 2011; Almansour et al. 2012a), and murine TNF-α (unpublished), respectively.

5.4.1 Interleukin 12 (IL12)

IL12 has an essential role in the interaction between the innate and adaptive arms of immunity by regulating inflammatory responses, innate resistance to infection, and adaptive immunity. In experimental tumor models, recombinant IL12 treatment has a dramatic antitumor effect on transplantable tumors, on chemically induced tumors, and in tumors arising spontaneously in genetically modified mice. IL12 utilizes effect or mechanisms of both innate resistance and adaptive immunity to mediate antitumor resistance. The stimulating activity of IL12 on antigen-specific immunity relies mostly on its ability to determine or augment Th1 and cytotoxic T lymphocyte responses. Because of this ability, IL12 has a potent adjuvant activity in cancer and other vaccines (Colombo and Trinchieri 2002; Borden

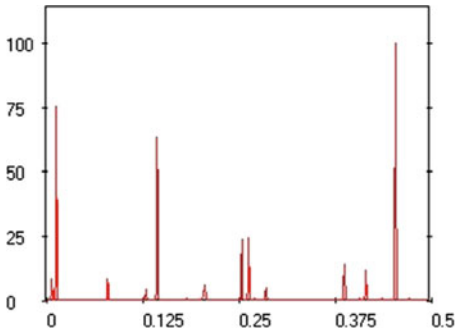


Fig. 5.1 Multiple cross-spectral function of 13 IL12 protein sequences

and Sondel 1990). IL12 was shown to have potent antitumor effects in murine models of melanoma, sarcoma, kidney cancer, lung cancer, colon cancer, and ovarian cancer (Kozar and Kaminski 2003; Den Otter and Jacobs 2008). As reported, systemic or peritumoral injection of IL12 can induce complete regression of established tumors, inhibit the formation of distant metastases, and substantially prolong the survival of tumor-bearing mice. These studies have identified doses of IL12 that can induce impressive tumor responses without causing overt toxicity (Kozar and Kaminski 2003).

Thirteen IL12 proteins from different origins were analyzed using the RRM. The characteristic RRM frequency (most prominent) was identified at $f_{RRM} = 0.4531$ as shown in Fig. 5.1. According to the RRM concepts, this prominent peak characterizes the common biological activity of analyzed IL12 proteins. Less prominent peaks observed in Fig. 5.1 confirm that these selected IL12 proteins can be involved in different biological processes (interact with other proteins). Mouse IL12 sequence (NP_032377, 215 aa, Entrez Protein Database) was selected as a parent protein to design a short bioactive peptide having IL12-like activity. The RRM-IL12 sequence, AREDLDERAQQKREDLDP, is an 18 amino acid linear peptide. The length of the designed peptide is defined using the ration $L = 1/f_{RRM}$. The synthetic peptide was designed with the frequency $f_{RRM} = 0.4531$ and phase $\varphi = 3.069$. The synthetic peptide sequence has the following characteristics: a molecular weight of 2.184 kDa, theoretical pI of 4.39, and an estimated half-life of 4.4 h in mammalian reticulocytes.

A similar procedure was used to design the negative control peptide analogue (RRM-C), which has a different "inactive" frequency and phase ($f_C = 0.2$, $\varphi_C = 1.5$) and does not express IL12 activity. The 22 amino acid linear peptide CVLQDCVLQDCVIQDCVLQDCV was designed as a negative control for biological cytotoxicity assays (molecular weight of 2.454 kDa, theoretical pI of 3.32, and estimated half-life of 1.2 h). The cytotoxic effects of the novel bioactive peptide RRM-IL12 and negative control peptide were experimentally validated on B16F0 mouse melanoma cell culture (Pirogova et al. 2011).

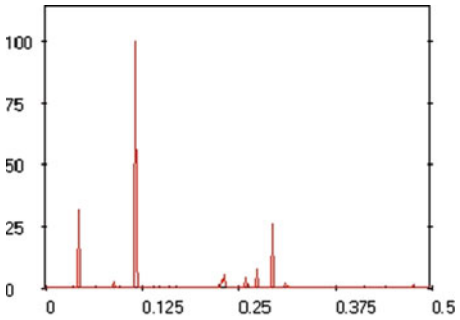


Fig. 5.2 Multiple cross-spectral function of 10 MV protein sequences

5.4.2 Myxoma Virus Proteins (MV)

The concept of using viruses to target and destroy cancerous cells is a novel form of cancer therapy. An optimal oncolytic virus candidate is required to possess a selective tropism to specifically infect and destroy the cancerous tissue without causing damage to normal tissue (Stanford et al. 2008). More recently, myxoma virus (MV) has been added to the oncolytic virus candidate list. MV is a poxvirus of the Leporipoxvirus genus and the Chordopoxvirinae subfamily of Poxviridae (Stanford et al. 2008; Fenner and Ross 1994; Wang et al. 2006; Werden and McFadden 2008; Sypula et al. 2004). It causes a lethal disease known as myxomatosis in the European rabbit, Oryctolagus cuniculus, but it is nonpathogenic in other vertebrates including humans (Stanford et al. 2008; Fenner and Ross 1994; Wang et al. 2006; Werden and McFadden 2008; Sypula et al. 2004). This virus has become of significant attention to researchers due to its oncolytic properties in human cancers, which have been reported in several studies both in vitro and in vivo (Stanford et al. 2008; Fenner and Ross 1994; Wang et al. 2006; Werden and McFadden 2008; Sypula et al. 2004).

In our recent study (Istivan et al. 2011), we used the RRM to design RRM-MV, a short linear peptide (18 amino acids, MW 2.345 kDa) with anticancer bioactive frequencies, as an analogue for the 49-kDa M-T5 protein molecule of myxoma virus. The 483 aa ankyrin-repeat protein NM-T5 protein (AAC55050) was selected as a parent protein for this analysis in addition to MV proteins: M-T1 (NP_051880), NM-T2 (NP_051879), T2 (AAA46632), T3C (CAA09973), MT3 (CAA0997), M-T4 (NP_051716), M-T6 (CAA09975), NM-T7 (AAA46631), and M-T8 (AAA46630) (Hetenyi and Van der Spoel 2002). Here we present the results of our computational analysis of myxoma virus (MV) proteins and the rational design of bioactive peptide analogue having the selected MV-T5-like activity. The RRM approach was used for analysis of selected ten MV proteins. Multiple cross-spectral function was obtained, and RRM characteristic frequency characterizing the common biological activity of the selected MV sequences was determined at $f_{MV} = 0.1152$ shown in Fig. 5.2.

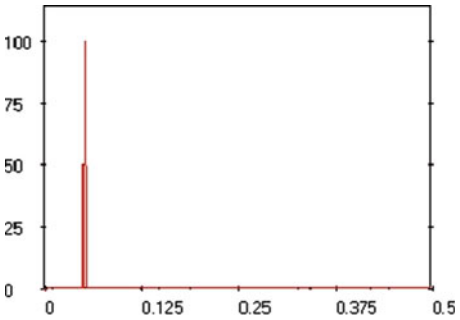


Fig. 5.3 Multiple cross-spectral function of 22 TNF-α protein sequences

On the basis of the characteristic frequencies and corresponding phases determined for selected MV proteins (the computational analysis is described in details in (Istivan et al. 2011)), we designed a short peptide that can exhibit the desired biological activity of MV protein. The 18 amino acid linear peptide sequence RRM-MV (MDDRWPLEYTDDTYEIPW) was designed with the frequency $f_{MV}=0.1152$ and phase $\varphi_{MV}=-0.457$. RRM-MV predicted theoretical pI, 3.66. Its estimated half-life in mammalian reticulocytes is 30 h (http://au.expasy.org/tools/protparam.html).

RRM-MV-C was also designed by the RRM with a frequency other than $f_{MV}=0.1152$ ($fc=0.08$, $\varphi_C=0.75$). It was used as a negative control peptide (Istivan et al. 2011; Almansour et al. 2012a).

5.4.3 Tumor Necrosis Factor (TNF-α)

Another therapeutic protein of our research interest was TNF-α which is a multifunctional cytokine playing significant roles in apoptosis, cell survival, inflammation, and immunity acting via two receptors. Current use of TNF-α in cancer therapy is in the regional treatment of locally advanced soft tissue sarcomas and metastatic melanomas. In this study, 22 mammalian TNF-α protein sequences have been analyzed, and their RRM characteristic frequency was identified (Fig. 5.3).

The original sequence TNF-2 (AAA40459 murine, 235 amino acids) was selected as a parent protein to *de novo* design a short bioactive peptide RRM-TNF that can express the same biological activity as the selected TNF-2 protein. The RRM characteristic frequency for the mouse TNF-α peptide analogue was designed on the basis of the frequency and phase determined within the RRM analysis: $f_{RRM}=0.0508$, $\varphi=-1.393$. According to the RRM concept, the RRM frequency characterizes the common biological activity of the selected TNF-α proteins. The 20 mer peptide analogue sequence (unpublished) was selected for our experimental work based on the characters of the protein sequence: molecular weight, 2576.7; theoretical pI, 4.75; estimated half-life, 1 h (mammalian reticulocytes, in vitro); and instability index, stable (ProtParam tool ExPASy).

5.5 Evaluation of Cytotoxic Effects of the RRM-Designed Peptides on Cancer and Normal Cells

Recently, our research has been focused on evaluating the suitability of the computationally designed peptides as melanoma therapeutics. We applied several qualitative and quantitative cellular cytotoxicity assays in this approach.

1. *Qualitative Evaluation*
   The cytotoxic effects of the bioactive peptides RRM-MV, RRM-IL, RRM-TNF, and RRM-C were investigated on several mammalian cell lines at different concentrations and different incubation times using confocal laser scanning microscopy (CLSM) (Istivan et al. 2011; Almansour et al. 2012a). Evaluation of in vitro assays via CLSM revealed that the bioactive peptide analogues (RRM-MV, RRM-IL, and RRM-TNF) induced the characteristic morphological changes associated with apoptosis and/or necrosis. The three peptides were cytotoxic to melanoma cells when compared with the non-treated cell culture and with the cell cultures treated with the negative control peptide RRM-C on mouse melanoma B16 cells (Fig. 5.4) and on human melanoma MM96L cells (Fig. 5.5). Similar cytotoxic effects were also detected on the human carcinoma cell line COLO-16 (Istivan et al. 2011; Almansour et al. 2012a).

   The cytotoxic effects on cancer cell lines were dose dependent as higher concentrations of the RRM peptide analogues had a stronger necrotic effect on the mouse melanoma B16 cells (Fig. 5.6) and on all the other tested cancer cell lines. The micrographs in (Fig. 5.6) indicate the abundant detachment of the confluent cell layer and the numbers of necrotic cells in cultures treated with highest concentration (1.6 μg/mL). In addition to being dose dependent, the cytotoxic effects were time dependent, as longer incubation periods with lower peptide concentrations caused more significant effect when incubated for up to 18 h. However, it was noticed that the B16 cells started to recover from the cytotoxic effects of RRM-MV after 16 h of incubation; therefore, a second dose of the same concentration of bioactive peptide at 16 h was found to enhance the cytotoxic effect (Istivan et al. 2011).

   On the other hand, no cytotoxic effects (apoptosis and necrosis) were detected when normal (noncancerous) cell lines were incubated with any of the RRM-designed peptides even at 1.6 μg/mL, a peptide concentration that can cause significant cytotoxic effects on cancer cell lines. The CLSM micrographs in Fig. 5.7 show the mouse macrophage J774 semi-adherent normal cell line after treatment with the RRM-designed peptides. Similarly, we have investigated the cytotoxic effects of RRM-MV on the mouse skin fibroblast primary cell culture (Istivan et al. 2011) and on two normal human skin cell lines, the epidermal melanocytes cell line (HEM) and the dermal fibroblast cell line (HDF) (Almansour et al. 2012a). No cytotoxic effects were detected on any of the normal cell lines when compared to the non-treated cell cultures incubated under similar conditions. Furthermore, toxic effects were not detected when the above-mentioned normal cell cultures were treated with RRM-IL or with RRM-TNF (unpublished).

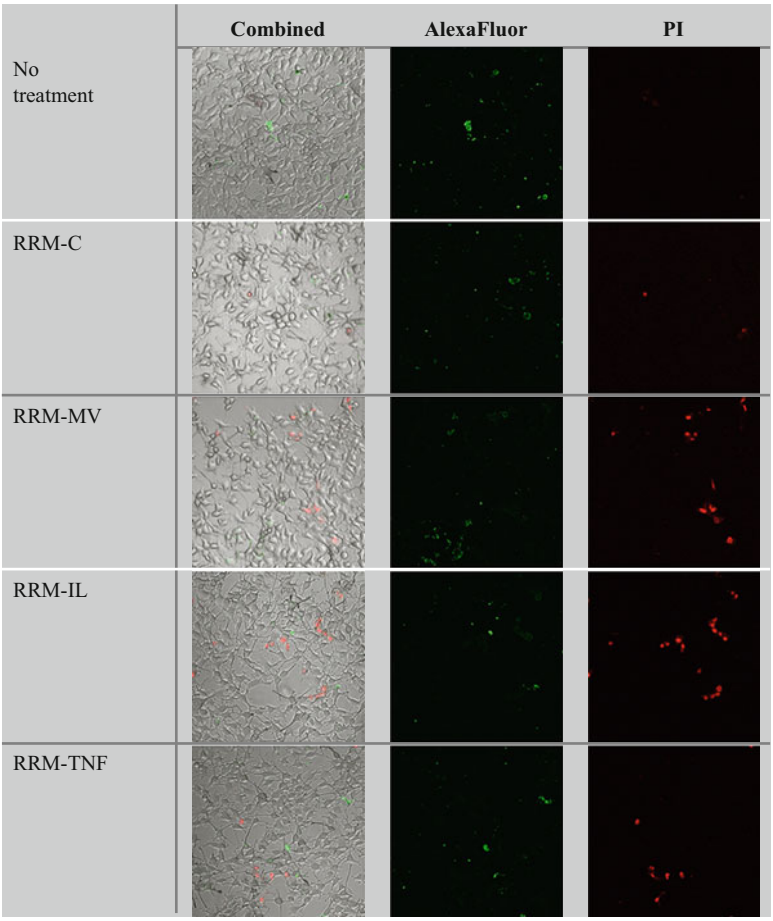


Fig. 5.4 CLSM micrographs (100× magnification) for the apoptosis/necrosis assay with annexin V-Alexa Fluor 488 and propidium iodide (*PI*) in the mouse melanoma cell line (B16F0) after 3-h incubation with 0.1 μg/mL of any the RRM-designed peptides. Cytotoxic changes including detachment of confluent layer, apoptotic cells (*green*), and necrotic cells (*red*) are obvious in cell lines treated with RRM-MV or RRM-IL or RRM-TNF

2. *Quantitative Evaluation*
   Assessment of cell cytotoxicity through LDH quantification is based on the detection of the release of lactate dehydrogenase upon the rupture of cellular membranes leading to cell death. We have applied this test to quantify the levels of LDH released from cells following treatment with the RRM-designed peptides. The LDH assay results supported our confocal microscopy data, as mouse melanoma B16 cells treated with RRM-IL, RRM-TNF, and RRM-MV produced significantly elevated LDH levels compared to the non-treated cells and to cells treated with the negative control peptide RRM-C (Fig. 5.8). Conversely, LDH levels were not significantly elevated when the J774 mouse macrophage cells were

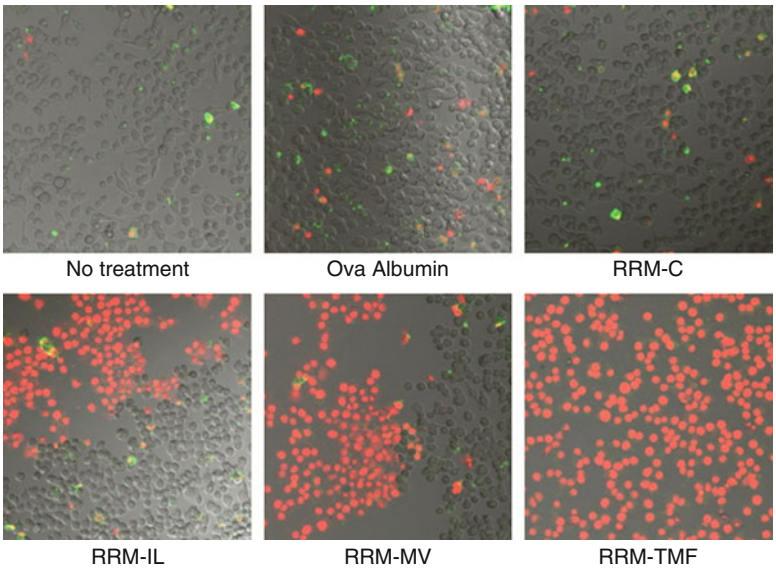


Fig. 5.5 CLSM micrographs (100× magnification) for the apoptosis/necrosis assay with annexin V-Alexa Fluor 488 (*green fluorescence*) and propidium iodide (*red fluorescence*) on the human melanoma cell line MM96L. Cell cultures were incubated with 200 ng/mL of each of *RRM-IL*, *RRM-MV*, *RRM-TNF*, and *RRM-C* for 3 h, and *ovalbumin* (200 ng/mL) was used as an additional negative control

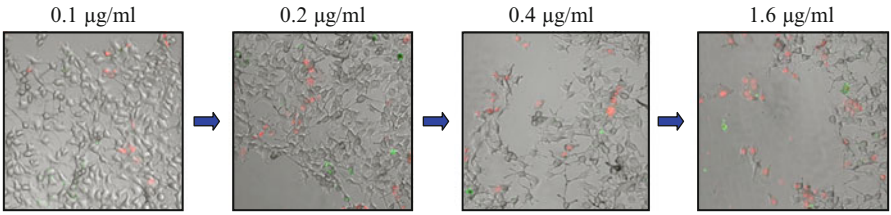


Fig. 5.6 CLSM micrographs (100× magnification) for the apoptosis/necrosis assay indicating the dose-dependent effect of RRM-MV on the mouse melanoma cell line (B16F0) after 3-h incubation with increasing concentrations of RRM-MV. Cytotoxic changes including detachment of confluent layer, apoptotic cells (*green*), and necrotic cells (*red*) are more obvious with higher concentrations of the peptide

treated with any of the RRM-designed peptides (Fig. 5.8), hence confirming the minimal cytotoxic effects in normal cells treated with the bioactive peptides which was detected by the CLSM qualitative technique (Figs. 5.4, 5.5, 5.6, and 5.7).

Levels of LDH released from damaged cells were also measured in cancer and normal mammalian cell lines following treatment with the bioactive peptide RRM-MV or with the negative control peptide RRM-C. Our statistical analysis data indicated the significant difference in LDH levels between cancer cells treated with RRM-MV when compared with non-treated cancer cells, with cancer cells treated with RRM-C, or with normal cells treated with any of the peptides (Fig. 5.9).

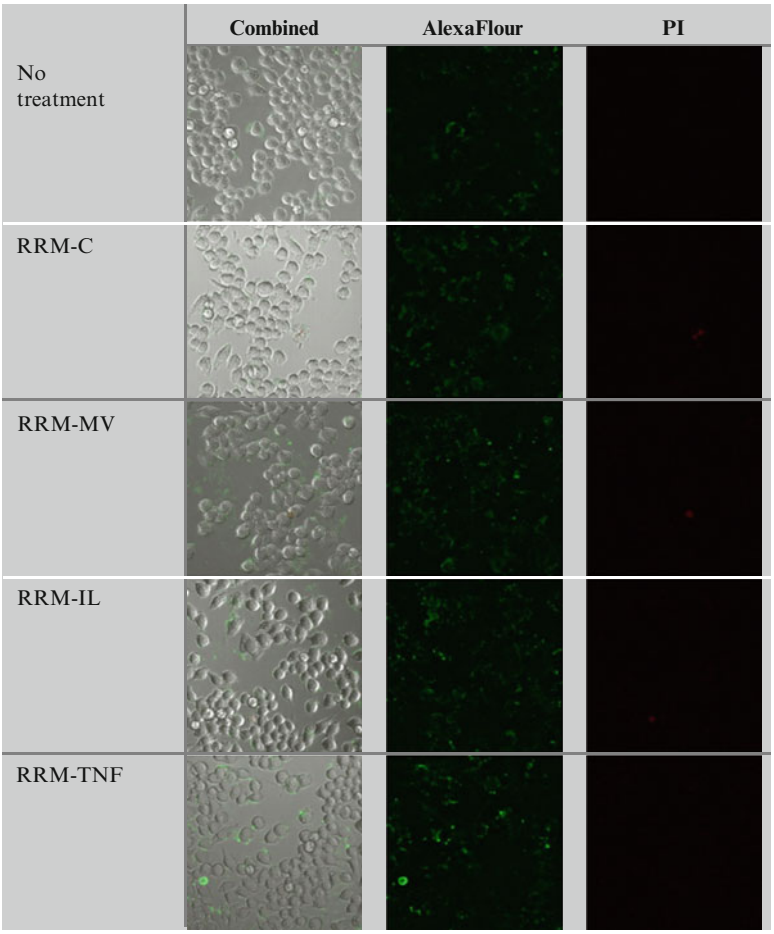


Fig. 5.7 CLSM micrographs (200× magnification) following the apoptosis/necrosis assay (with annexin V-Alexa Fluor 488 and propidium iodide) in the mouse macrophage normal cell line (J774) after 3-h incubation with 1.6 μg/mL of one of the RRM-designed peptides (RRM-C, RRM-MV; RRM-IL, RRM-TNF). No significant cytotoxic effects (apoptosis, necrosis, and cellular detachment) were detected in any of the treated cell cultures as compared with the non-treated cell culture similarly incubated in DMEM, indicating the minimal cytotoxic effect of the peptides on normal cells

5.5.1 Effect of RRM-Designed Peptides on Human Erythrocytes

To assess the suitability for the RRM-designed peptides as future cancer therapeutics, their toxicity on human erythrocytes was tested using the hemolysis assay. Human erythrocytes were incubated with various concentrations of the each of the following peptides: RRM-IL, RRM-MV, RRM-TNF, and RRM-C for 18 h. Results presented in (Fig. 5.10) indicated that the above-tested peptides did not induce erythrocyte lysis at any of the concentrations used when compared with the complete erythrocyte lysis positive control, thus concluding that the tested peptides did not harm human blood cells.

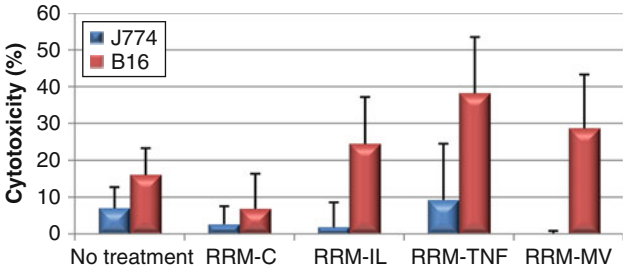


Fig. 5.8 LDH cytotoxicity assay for mouse melanoma (*B16*) and mouse macrophage (*J774*) cell lines after treatment with the RRM-designed peptides. Cells (3×10^5) were incubated for 3 h with *RRM-C, RRM-IL, RRM-TNF*, or *RRM-MV* at 1.6 μg/mL. Cells without treatment were similarly incubated for 3 h. Each bar represents mean ± standard errors of three separate experiments in triplicate

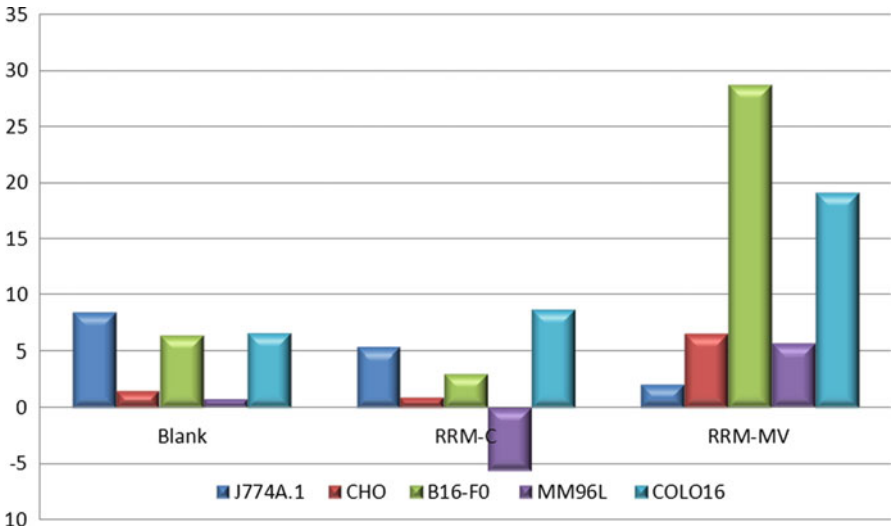


Fig. 5.9 Cytotoxic effect of RRM peptide analogues on normal and cancer cells measured by LDH assay. Cells (3×10^5) were incubated for 3 h with *RRM-C* or with *RRM-MV* at 400 ng/mL (for *COLO16*) and 1,600 ng/mL (for *CHO, J774A*, and *B16-F0*). Cells without treatment were similarly incubated for 3 h (blank). Each bar represents mean ± standard errors of three separate experiments in triplicate

5.5.2 *Evaluation of the Effects of Other Designed Negative Control Peptides Versus the Bioactive Peptide RRM-MV*

To assess the efficiency of the RRM to design bioactive peptide with a specific biological function, we have recently evaluated the biological effects of the bioactive peptide RRM-MV *versus* four negative control peptides on human skin cancer

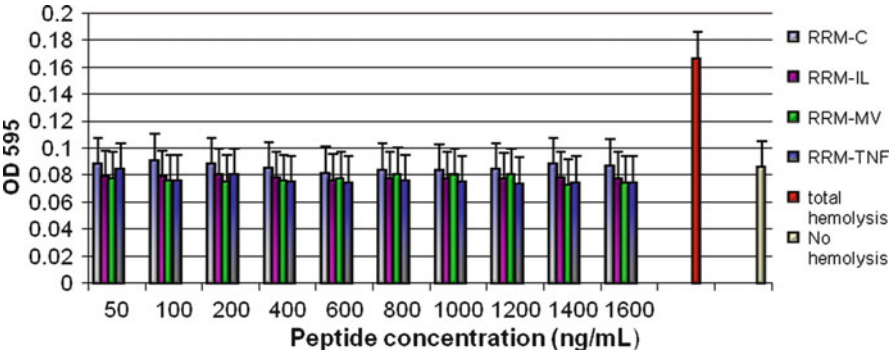


Fig. 5.10 The effect of the RRM-designed peptides on human erythrocytes. Washed human erythrocytes in PBS were incubated for 18 h with increasing concentrations of *RRM-C, RRM-IL, RRM-MV*, and *RRM-TNF*. Human red blood cells were relatively resistant to the RRM-designed peptides when incubated for up to 18 h, as cytolytic/hemolytic effects were not detected. Data values are the average of three independent experiments in triplicates

cells and on human normal skin cells in vitro (Almansour et al. 2012b). The negative control peptides were specifically designed *not* to possess the antitumor activity exhibited by RRM-MV on cancer cells. Hence, the negative control peptide RRM-MV-C (DDDCWHVLEKWTDDDRQA) was designed by RRM with a frequency other than $f_{MV} = 0.1152$ ($f_C = 0.08$, $\varphi_C = 0.75$). Furthermore, to investigate if any change in the order of the amino acid within the sequence of the bioactive peptide RRM-MV can influence its biological activity, the order of one amino acid in the original RRM-MV sequence was changed at different positions to create the following scrambled peptides:

MVC1-end (MDDRWPLEYTDDTYEI**WP**)
MVC2-mid (MDDRWP**EL**YTDDTYEIPW)
MVC3-side (MDDRWPLEYTD**TD**YEIPW)

The new controls were tested for their biological activity on cancer cells, in comparison with the cytotoxic activity of the RRM-MV. The qualitative microscopic data confirmed the lack of cytotoxic effects on the MM96L melanoma cell line when treated with the non-bioactive peptide controls (Fig. 5.11). None of the studied negative control peptides (RRM-MV-C, MVC1, MVC2, and MVC3) caused obvious apoptosis or necrosis on the tested cancer cell lines (MM96L and COLO-16) even after incubation for up to 18 h. Furthermore, the negative control peptides and the bioactive peptide RRM-MV did not have any cytotoxic effect on the normal human dermal fibroblast (HDF) (Almansour et al. 2012b). This can highlight the fact that the lack of the computationally determined RRM frequency in the control peptides RRM-C and RRM-MVC diminished their cytotoxic effect on cancer cells. In addition, any change in the unique amino acid sequence of the RRM-designed bioactive peptide (RRM-MV) will devalue its desired biological function.

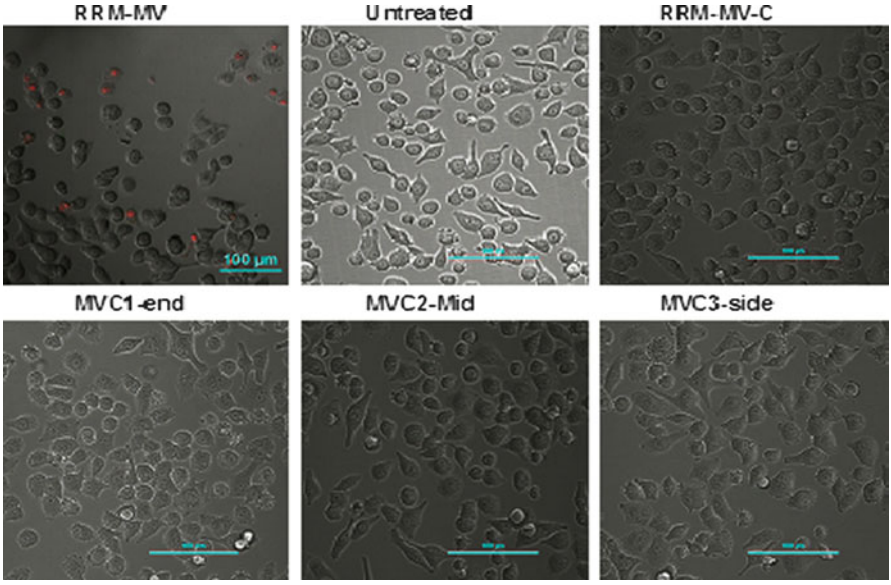


Fig. 5.11 CLSM combined micrographs for MM96L cell line following treatment with 200 ng/mL of the control peptides using Vybrant Apoptosis Assay kit, showing necrotic cells in the *RRM-MV*-treated cell culture only

5.5.3 Investigating Possible Targets for the Bioactive RRM-Designed Peptides in Cancer Cells

It was noticed that the bioactive peptides can only induce cytotoxic effects on cancer cells when the treated cell cultures are incubated at the body temperature 37°C, while there were no indications on cellular cytotoxicity when the cancer cell cultures were incubated at 4°C instead of 37°C after the RRM peptide treatment (Fig. 5.12). This is an indication to the possibility of the involvement of cellular apoptosis pathways targeted by the peptide treatment in these cells. To understand the mechanism of action of the bioactive RRM-designed peptides, we have investigated the effect of RRM-MV on some cellular pathways involved in apoptosis in cancer cells such as phospho-Akt (Istivan et al. 2011; Almansour et al. 2012a), cleaved casapase-3, and p-53 (unpublished).

Akt, or protein kinase B (PKB), is a serine/threonine kinase that plays a crucial role in the regulation of many cellular processes, including apoptosis or programmed cell death; therefore, we have investigated the effect of the bioactive peptide RRM-MV and the negative control RRM-C on Akt cell signaling pathway in mammalian cancer and normal cell lines. The effect of RRM-MV or RRM-C

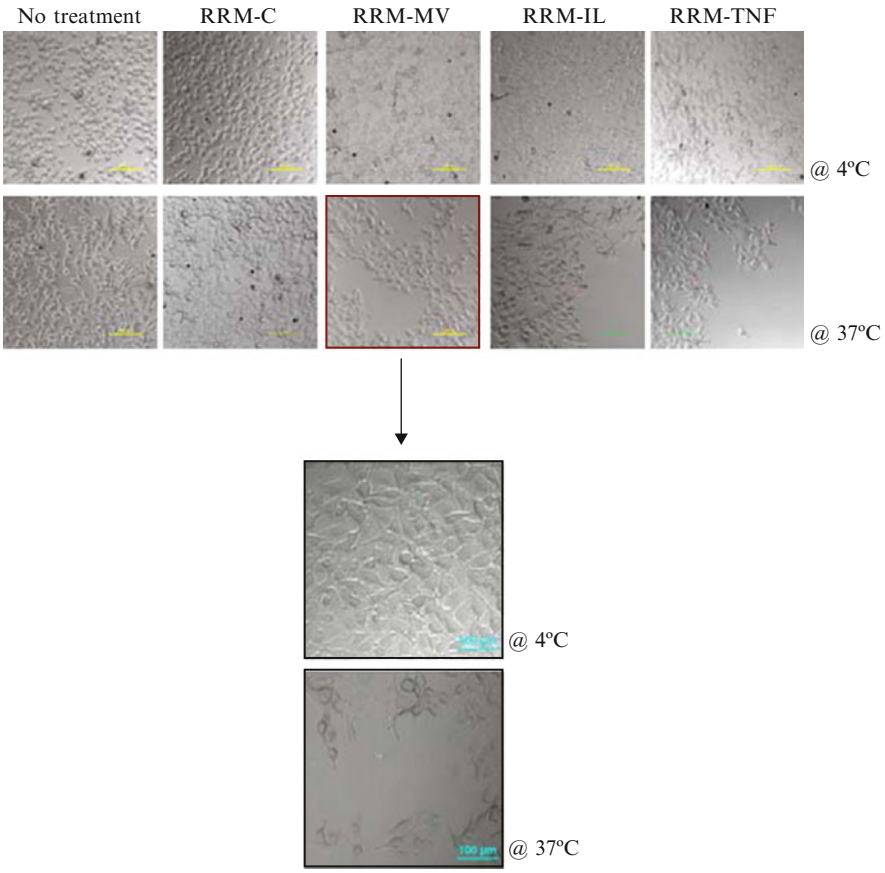


Fig. 5.12 CLSM micrographs for the apoptosis/necrosis assay in the mouse melanoma cell line (B16F0) after 18-h incubation with 0.1 μg/mL of the RRM-designed peptide. Cytotoxic changes including detachment of the confluent layer and necrotic cells (*red*) are only obvious in cancer cell cultures incubated at 37°C following treatment with *RRM-MV, RRM-IL*, or *RRM-TNF*. Cell cultures incubated at 4°C did not show any sign of cytotoxicity in comparison to cultures with similar treatments incubated at 37°C

treatment on Akt signaling pathways in B16F0, MM96L, COLO-16, and HDF cell lines was investigated using immunoblotting (Istivan et al. 2011; Almansour et al. 2012a). Our results indicated that the Akt signaling pathway did not seem to be affected by RRM-MV treatment in the mouse melanoma cell line (B16F0) (Istivan et al. 2011). On the contrary, the levels of p-Akt expression at Ser-473 and Thr-308 were slightly increased in the RRM-MV-treated human melanoma and carcinoma cells MM96L and COLO-16 (Almansour et al. 2012a). Levels of p-Akt

and total Akt were unaffected by RRM-C treatment in all tested cancer cell lines. Likewise, levels of p-Akt and total Akt expression were unaffected after RRM-MV or RRM-C treatment in the HDF normal human skin cell line (Almansour et al. 2012a).

Because p53 and cleaved casapase-3 proteins play a significant role in cell death and are the potential targets for RRM-MV treatment, the cellular expression of p53 and cleaved casapase-3 proteins in melanoma and carcinoma cells were assessed. We found that the levels of these proteins increased after the RRM-MV treatment only in cancer cells, while there was no noticeable increase in the levels of P53 and cleaved casapase-3 in normal cells indicating that RRM-MV can significantly affect cancer progression leading to apoptosis in melanoma and carcinoma cells without harming normal cells (unpublished).

5.6 Conclusion

In this chapter, a brief review of current advances and novel approaches in experimental and computational drug discovery, design, and development has been presented. As knowledge grows about the proteins involved in tumor cell development, peptides will be the first available inhibitors for therapy of the newly discovered target proteins. Therapeutic peptides can be developed for the inhibition or reactivation of a huge variety of important signaling molecules. Furthermore, these peptides can be very specific with regard to their target proteins and in some cases can also be specific with regard to the cancerous cell types. Therefore, owing to their ease of design and production and wide spectrum of potential targets, therapeutic peptides have a promising future in cancer therapy. The rational design of therapeutic peptides to inhibit interactions of interest is relatively easy and certainly much easier than designing small molecules to inhibit the same interactions. This gives drug developers access to inhibitors of many important protein–protein interactions to which small molecule inhibitors are not available. Furthermore, because therapeutic peptides can be very specific, this can reduce the likelihood of "off-target" effects. However, so far, the use of peptides for cancer treatment has been limited due to their poor performance pharmacologically. Limitations of stability in plasma, bioavailability, and tumor cell penetration have prevented the advance of peptides beyond preclinical testing. Therefore, the key issue for the development of this new class of drugs is not finding new therapeutic peptides, but finding new ways for their delivery to tumor sites for in vivo validation.

In this manuscript, we also discussed the use of de novo designed peptides for cancer therapy using the example of the IL12-like, MV-T5-like, and TNF-α-like short peptides, which were designed computationally using the RRM approach. The findings obtained within the experimental evaluation indicate the suitability of the RRM approach to design bioactive peptides with the desired biological functions.

Acknowledgements The authors would like to thank Emily Gan and Nahlah Almansour for their role in performing the biological experiments presented in this chapter. We would also like to thank Peter Coloe and Irena Cosic for their moral and financial support for this research. The experimental investigation of the synthetic peptides, reported in this manuscript, was supported by the RMIT Health Innovations Research Institute.

References

Abe K, Kobayashi N, Sode K, Ikebukuro K. Peptide ligand screening of asynuclein aggregation modulators by in silico panning. BMC Bioinformatics. 2007;8:451.

Ajay A, Walters WP, Murcko MA. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? J Med Chem. 1998;41:3314–24.

Almansour NM, Pirogova E, Coloe PJ, Cosic I, Istivan TS. A bioactive peptide analogue for myxoma virus protein with a targeted cytotoxicity for human skin cancer in vitro. J Biomed Sci. 2012a;19:65.

Almansour NM, Pirogova E, Coloe PJ, Cosic I, Istivan TS. Investigation of cytotoxicity of negative control peptides versus bioactive peptides on skin cancer and normal cells: a comparative study. Future Med Chem. 2012b;4(12):1553–65.

Bidwell III GL, Raucher D. Therapeutic peptides for cancer therapy. Part I – peptide inhibitors of signal transduction cascades. Expert Opin Drug Deliv. 2009;6(10):1033–47.

Borden EC, Sondel PM. Lymphokines and cytokines as cancer treatment immunotherapy realized. Cancer. 1990;65(3):800–14.

Bright RK, Franz RW. Book review. Peptide-based cancer vaccines. Leukemia. 2002;16:970–1. doi:10.1038/sj/leu/2402436.

Budin N, Ahmed S, Majeux N, Caflisch A. An evolutionary approach for structure-based design of natural and nonnatural peptidic ligands. Comb Chem High Throughput Screen. 2001a;4:661–73.

Budin N, Majeux N, Caflisch A. Fragment-based flexible ligand docking by evolutionary optimization. Biol Chem. 2001b;382:1365–72.

Budin N, Majeux N, Tenette-Souaille C, Caflisch A. Structure based ligand design by a build-up approach and genetic algorithm search in conformational space. J Comput Chem. 2001c;22:1956–70.

Colombo MP, Trinchieri G. Interleukin-12 in anti-tumor immunity and immunotherapy. Cytokine Growth Factor Rev. 2002;13(2):155–68.

Cosic I. Macromolecular bioactivity: is it resonant interaction between molecules? – Theory and applications. IEEE Trans Biomed Eng. 1994;41:1101–14.

Cosic I. Spectroscopy for fun and profit. Nat Biotechnol. 1995;13:236–8.

Cosic I. The resonant recognition model of macromolecular bioactivity: theory and applications. Basel: Birkhauser Verlag; 1997.

Cosic I, Pirogova E. Bioactive peptide design using the resonant recognition model. Nonlinear Biomed Phys. 2007;1:7.

Cosic I, Drummond AE, Underwood JR, Hearn MTW. In vitro inhibition of the actions of basic FGF by a novel 16 amino acid peptide. Mol Cell Biochem. 1994;130:1–9.

Craig LC, King TP. Design and use of a 1000-tube countercurrent distribution apparatus. Fed Proc. 1958;17:1126–34.

Den Otter W, Jacobs JJL. Local therapy of cancer with free IL12. Cancer Immunol Immunother. 2008;57(7):931–50.

Dolle RE. Comprehensive survey of chemical libraries yielding enzyme inhibitors, receptor agonists and antagonists, and other biologically active agents: 1992 through 1997. Mol Divers. 1998;3:199–233.

Fenner F, Ross J. Myxomatosis. In: Thompson GV, King CM, editors. The European rabbit, the history and biology of a successful colonizer. Oxford: Oxford University Press; 1994. p. 205–39.

Frenkel D, Clark DE, Li J, Murray CW, Robson B. Pro-ligand – an approach to de-novo molecular design 4. Application to the design of peptides. J Comput Aided Mol Des. 1995;9:213–25.
Hetenyi C, Van der Spoel D. Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci. 2002;11:1729–37.
Istivan T, Pirogova E, Gan E, Almansour NM, Coloe PJ, Cosic I. Biological effects of a de novo designed myxoma virus peptide analogue: evaluation of cytotoxicity on tumor cells. PLoS One. 2011;6(9):e24809. Published online, 19th September, 2011.
Juretic D, Vukicevic D, Ilic N, Antcheva N, Tossi A. Computational design of highly selective antimicrobial peptides. J Chem Inf Model. 2009;49:2873–82.
Kamphausen S, Holtge N, Wirsching F, Morys-Wortmann C, Riester D, et al. Genetic algorithm for the design of molecules with desired properties. J Comput Aided Mol Des. 2002;16:551–67.
Klepeis JL, Floudas CA, Morikis D, Tsokos CG, Lambris JD. Design of peptide analogues with improves activity using a novel de novo protein design approach. Ind Eng Chem Res. 2004;43:3817–26.
Knutson KL, Schiffman K, Disis ML. Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T cell immunity in cancer patients. J Clin Invest. 2001;107:477–84.
Kozar K, Kaminski R. Interleukin 12-based immunotherapy improves the antitumor effectiveness of a low-dose 5-aza-2′-deoxycitidine treatment in L1210 leukemia and B16F10 melanoma models in mice. Clin Cancer Res. 2003;9:3124–33.
Krsmanovic V, Biquard JM, Sikorska-Walker M, Cosic I, Desgranges C, Trabaud MA, et al. Investigation into the cross-reactivity of rabbit antibodies raised against nonhomologous pairs of synthetic peptides derived from HIV-1 gp120 proteins. J Pept Res. 1998;52(5):410–4120.
Lev DC, Kim LS, Melnikova V, Ruiz M, Ananthaswamy HN, Price JE. Dual blockade of EGFR and ERK1/2 phosphorylation potentiates growth inhibition of breast cancer cells. Br J Cancer. 2004;91:795–802.
Lipka E, Crison J, Amidon GL. Transmembrane transport of peptide type compounds: prospects for oral delivery. J Control Release. 1996;39(2–3):121–9.
Mayrose I, Penn O, Erez E, Rubinstein ND, Shlomi T. Pepitope: epitope mapping from affinity-selected peptides. Bioinformatics. 2007;23:3244–6.
McCarty MF. Targeting multiple signaling pathways as a strategy for managing prostate cancer: multifocal signal modulation therapy. Integr Cancer Ther. 2004;3:349–80.
Moreau V, Fleury C, Piquer D, Nguyen C, Novali N. PEPOP: computational design of immunogenic peptides. BMC Bioinformatics. 2008;9:71.
Old LJ. Immunotherapy for cancer. Sci Am. 1996;275:136–43.
Peterson EA, Sober HA. Chromatography of proteins. I. Cellulose ion-exchange adsorbents. J Am Chem Soc. 1956;78:751–5.
Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB. Accurate prediction of peptide binding sites on protein surfaces. PloS Comput Biol. 2009;5:e1000335.
Pickett SD, McLay IM, Clark DE. Enhancing the hit-to-lead properties of lead optimization libraries. J Chem Inf Comput Sci. 2000;40:263–72.
Pirogova E, Fang Q, Akay M, Cosic I. Investigation of the structural and functional relationships of oncogene proteins. Proc IEEE. 2002;90(12):1859–68.
Pirogova E, Istivan T, Gan E, Cosic I. Advances in methods for therapeutic peptide discovery, design and development. Curr Pharm Biotechnol. 2011;12:1117–27.
Porath J, Flodin P. Gel filtration: a method for desalting and group separation. Nature. 1959;183:1657–9.
Riester D, Wirsching F, Salinas G, Keller M, Gebinoga M. Thrombin inhibitors identified by computer-assisted multiparameter design. Proc Natl Acad Sci U S A. 2005;102:8597–602.
Ripka AS, Rich DH. Peptidomimetic design. Curr Opin Chem Biol. 1998;2:441–52.
Sewald N, Jakubke H-D. Peptides: chemistry and biology. Weinheim: Wiley-VCH Verlag GmbH & Co; 2002.
Singh J, Ator MA, Jaeger EP, Allen MP, Whipple DA. Application of genetic algorithms to combinatorial synthesis: a computational approach to lead identification and lead optimization. J Am Chem Soc. 1996;118:1669–76.

Stanford MM, Shaban M, Barret JW, Werden SJ, Gilbert P, et al. Myxoma virus oncolysis of primary and metastatic B16F10 mouse tumors in vivo. Mol Ther. 2008;16:52–9.
Stein A, Aloy P. A molecular interpretation of genetic interactions in yeast. FEBS Lett. 2008;582:1245–50.
Sypula J, Wang F, Ma Y, Bell J, McFadden G. Myxoma virus tropism in human tumor cells. Gene Ther Mol Biol. 2004;8:103–14.
Talmadge JE. Pharmacodynamic aspects of peptide administration biological response modifiers. Adv Drug Deliv Rev. 1998;33(3):241–52.
Tortora G, Caputo R, Damiano V, Melisi D, et al. Combination of a selective cyclooxygenase-2 inhibitor with epidermal growth factor receptor tyrosine kinase inhibitor ZD1839 and protein kinase A antisense causes cooperative antitumor and antiangiogenic effect. Clin Cancer Res. 2003;9:1566–72.
Unal EB, Gursoy A, Erman B. VitAL: Viterbi Algorithm for de novo peptide design. PLoS One. 2010;5(6):e10926.
Veljkovic V, Slavic M. General model of pseudopotentials. Phys Rev Lett. 1972;29:105–8.
Waldmann TA. Immunotherapy: past, present and future. Nat Med. 2003;9:269–77. doi:10.1038/nm0303-269.
Wang G, Barret J, Stanford M, Werden S, Johnston J, et al. Infection of human cancer cells with myxoma virus requires Akt activation via interaction with a viral ankyrin-repeat host range factor. Proc Natl Acad Sci. 2006;103:4640–5.
Werden SJ, McFadden G. The role of cell signalling in poxvirus tropism: the case of the M-T5 host range protein of myxoma virus. Biochim Biophys Acta. 2008;1784:228–37.

Elena Pirogova PhD, Australia Dr. Pirogova is a senior lecturer at School of Electrical and Computer Engineering, RMIT University, Australia. She graduated in 1991 from National Technical University of Ukraine (NTUU) with a BEng (Hons) in Chemical Engineering in the field of physical chemistry and electrochemistry. In 2002, she completed a PhD degree in biomedical engineering at the Department of Electrical and Computer Systems Engineering, Monash University, Australia. Elena has produced 57 journals and refereed conference papers and 4 book chapters.

Dr. Pirogova is program director, Bachelor of Biomedical Engineering at RMIT University. She is research group leader of health and biological effects of electromagnetic radiation, Biophysics and Bioengineering Program, Health Innovations Research Institute, RMIT University.

Dr. Pirogova's major research focus is investigation of biological and health effects of electromagnetic radiation on tissue, cells, and proteins; protein modeling; and design of novel semiconducting biomaterials. In particular, she is interested in therapeutic applications of electromagnetic energy. She is currently working on development of a novel methodology for cancer therapy using de novo designed peptides and infrared radiation. Dr. Pirogova also leads another collaborative project, which is focused on design and development of induction system, based on the low-frequency electromagnetic device, for promotion of wound healing. The outcomes of her research were presented at the prestigious international biomedical engineering conferences and published in the PLoS One, Current Pharmaceutical Biotechnology, Proceedings of the IEEE, IEEE Transactions on Nanobioscience, IEEE Transaction on Information Technology in Biomedicine, Molecular Simulation, etc.

Taghrid Istivan is a lecturer at RMIT University, Australia. She was awarded PhD qualification in biotechnology from RMIT University in 2005. Her research interests are in developing novel therapeutics and in microbial pathogenesis. She also holds MSc. in Microbial Genetics.

Chapter 6
Advances in Proteomic Methods

Xianyin Lai

Abstract It has been almost 20 years since the word proteome was coined in 1994. Proteomic techniques have experienced three generations: 2-DE MALDI-TOF, bottom-up, and top-down, although top-down has not yet reached its full potential. Though numerous techniques have been developed to assist the achievement of accurate results in large scale with high-throughput capability, the field of proteomics is still driven by the development of new technologies for sample processing, protein identification, quantification, structure, and function. Advances in sample preparation mainly focus on sample fractionation, enrichment, and separation. Current trends aim to obtain certain types of peptides and proteins or specific organelles. Advances in protein identification are aimed to detect and identify more proteins accurately; while advances in protein quantification are focused on improvement of accuracy, precision, and reproducibility. Structural and functional proteomics have begun, but require continued development to reach their full potential. In summary, every technique is developed to improve performance and/or reduce cost.

Keywords Proteomics • Advance • Sample preparation • Identification • Quantification • Structure • Function

X. Lai, Ph.D. (✉)
Department of Cellular and Integrative Physiology,
Indiana University School of Medicine, 1345 West 16th Street,
Room 306, Indianapolis, IN 46202, USA
e-mail: xlai@iupui.edu

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_6,

6.1 Introduction

6.1.1 What Is Proteomics

Originally, proteomics was defined as "the study of proteins, how they're modified, when and where they're expressed, how they're involved in metabolic pathways and how they interact with one another" in 1996 by Wilkins, who coined the term proteome in 1994 as "all proteins expressed by a genome, cell or tissue" (Wilkins et al. 1996; Huber 2003). Along with the development of proteomics, its concept has been extended to all aspects of proteins: to define and characterize the identities, quantities, structures, and functions of proteins in different cellular contexts (Phizicky et al. 2003). Proteomics can be defined simply as a large-scale approach to analyze the entire set of proteins expressed in a biosample (Yates 2000; Corthals and Nelson 2001). The words comprehensive and global have been used very often to substitute for large scale (Anderson and Anderson 1998; Ziegler 2001).

6.1.2 Why Study Proteins

In the steps leading from a gene to a fully functional protein, a gene is transcribed to mRNA, mRNA is translated into a protein, and the protein is modified to a fully functional form. Although significant achievements have been made in genomics, mRNA measurements cannot substitute for protein measurements for several reasons: (1) proteins, rather than nucleic acids, mediate most of the physiologic functions within the cell, and (2) proteins are regulated in a variety of ways, many of which do not involve changes in mRNA levels, such as translational regulation, posttranslational modification, and protein degradation. The protein and mRNA responses are only partially correlated (Anderson and Anderson 1998; Knepper 2002). Therefore, the study of proteomics is essential and complementary to the area of functional genomics.

6.1.3 Current Proteomic Methodologies

It has been almost 20 years since the word proteome was coined in 1994. With terms proteome, proteomic, or proteomics in titles or abstracts searched in the PubMed database, 40,455 articles have been found between January 1, 1995, and July 31, 2012. In the same period, 250,093 articles have been published. The data from 1995 to 2011 of proteomics and genomics publications have been compared in Fig. 6.1. Genomics publications have a consistent rate of increase, at an average of 9% per year from 1995 to 2012. Proteomics publications have experienced a dramatic increase with an average rate of 164% per year before 2002, a rapid increase with

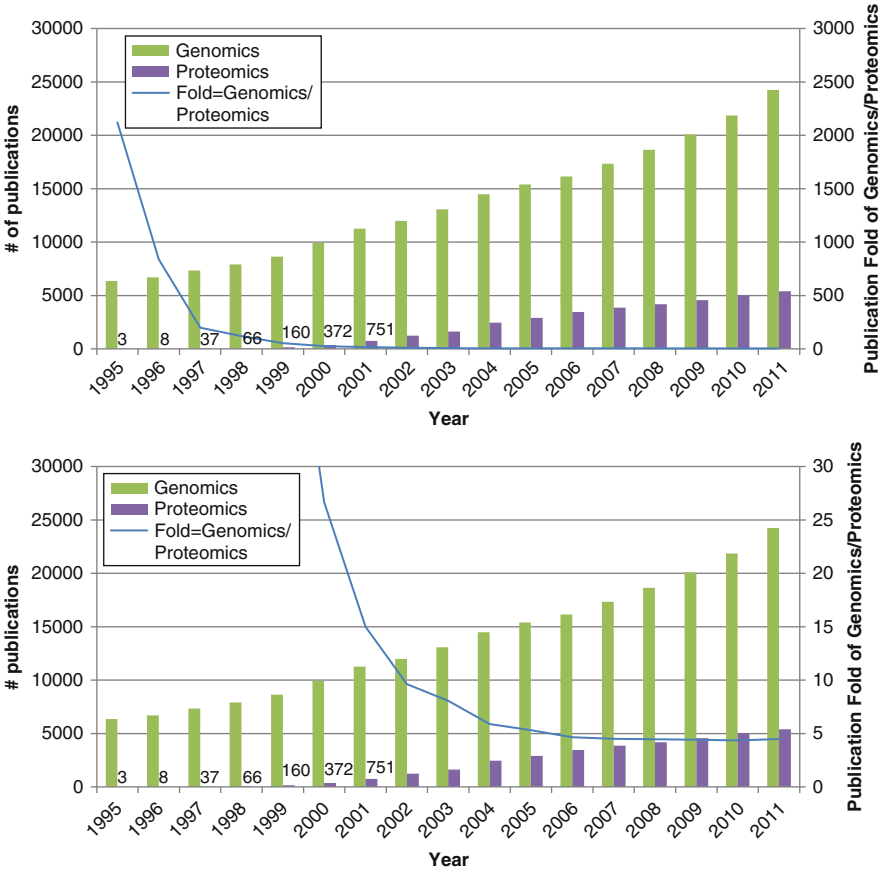


Fig. 6.1 The numbers and comparison of proteomics and genomics publications from 1995 to 2011. More and more proteomics and genomics papers were published every year during that period. Proteomics publications had a significantly greater rate of increase than genomics publications before 2007. Since 2006, the ratio of genomics and proteomics publications has remained at a stable level

an average rate of 37% per year between 2002 and 2006, and a consistent increase rate at an average of 9% per year from 2007 to 2012. Since 2006, the ratio of genomics and proteomics publications has remained around 4.5. Overall, genomics has been popular for nearly the last 20 years, while proteomics has become more and more popular until it reached a stable stage compared with genomics in recent years.

Historically, proteomics originated with the combination of 2D gel electrophoresis (2-DE) and mass spectrometry (MS). Since then, proteomics has experienced numerous technical developments. The field of proteomics is still driven by the development of new technologies for sample processing, protein identification, quantification, structure, and function.

In proteomics, an ideal method should have the ability to accurately analyze all proteins in a sample and to complete the analysis in seconds. Accuracy is at the core of proteomics. Global, along with comprehensive and large scale, is the most common word used to describe the analysis of all proteins in a sample. High-throughput is a phrase generally applied to describe the completion of a sample in seconds. Currently, such an ideal method does not exist. However, any advance in each proteomic step aims to improve accuracy, scale, or throughput capabilities of present methods.

In the development of proteomics, too many terms or acronyms have been coined, causing communication problems for researchers who are not in proteomics fields. For example, SELDI-TOF was coined for surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. Actually, SELDI is a MALDI (matrix-assisted laser desorption/ionization) technique and not a new ionization method. The difference is that the SELDI surface is used for protein fractionation (Hutchens and Yip 1993). Many other terms have been coined to express a combination of multiple techniques, such as MudPIT for multidimensional protein identification technology (Washburn 2004). This chapter's composition is based on experimental processes and principles that summarize the most important advances of proteomic methods for biomedical research in the past 20 years with representative techniques, not focusing on introducing every technique that has been published. Techniques that are simply improvements of existing techniques without significant innovation or techniques that have limited value in biomedical applications have been ignored.

6.2 Sample Preparation

Generally, a biological sample may be an entire organism, an organ, a specific tissue, a type of cells, particular organelles, certain proteins, unique peptides, or a special body fluid (Fig. 6.2). Due to the sensitivity limitation of proteomic methods, low-abundance proteins at each organizational level are very difficult to analyze. Therefore, the fractionation of a sample at a sublevel enriches the low-abundance proteins, enabling the analysis of more proteins and leading to a global analysis. Other than laser capture microdissection (LCM) to isolate cells of interest from a tissue (Cheng et al. 2012), techniques for fractionation of organs, tissues, cells, and body fluids are simple and relatively easy to carry out. On the contrary, organelle, protein, and peptide fractionations gain continuous attention, because they significantly increase the number of proteins that can be analyzed. Fractionation, enrichment, and depletion are the most common words used to describe different processes and techniques to achieve this increase, but their aim is the same: to enable the detection of more proteins, especially for those in low abundance. Meanwhile, protein and peptide separations greatly increase the number of proteins that can be analyzed. Therefore, besides LCM, this section will focus on the techniques of organelle, protein, and peptide fractionation and the techniques of protein and peptide separation (Fig. 6.3), which improve the global capability of current proteomic methods.

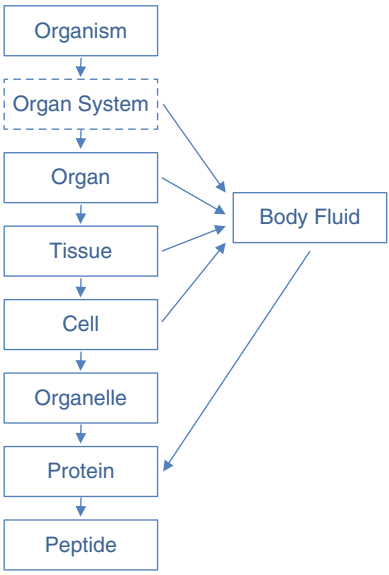


Fig. 6.2 Samples for proteomic analysis at different levels. Generally, a biological sample may be an entire organism, an organ, a specific tissue, a type of cells, particular organelles, certain proteins, unique peptides, or a special body fluid

Since most experiments are performed at the peptide level (bottom-up proteomics), many techniques have been developed for the generation of peptides from proteins. A summarization of those techniques also is included in this section.

6.2.1 Cell Isolation

Laser Capture Microdissection (LCM): LCM has been extensively applied to isolate tumor cells from their surrounding tissues, such as breast cancer (Liu et al. 2012), lung cancer (Xu et al. 2012a), and kidney cancer (Kurban et al. 2012). Because tumor cells are always surrounded by non-tumor cells and the percentage of tumor cells varies dramatically between individual tumor samples, LCM allows enrichment of tumor cells and reduces sample bias. Its selectivity for specific cell populations is able to greatly improve the diagnosis. The development, platforms, principle, applications, advantages, limitations, and future of LCM have been reviewed recently (Cheng et al. 2012).

6.2.2 Organelle Fractionation

An organelle is a specialized subunit of a cell and has a specific function. The most common subunits of an animal cell are plasma membrane, cytosol, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum, vacuole, and cytoskeleton. Minor

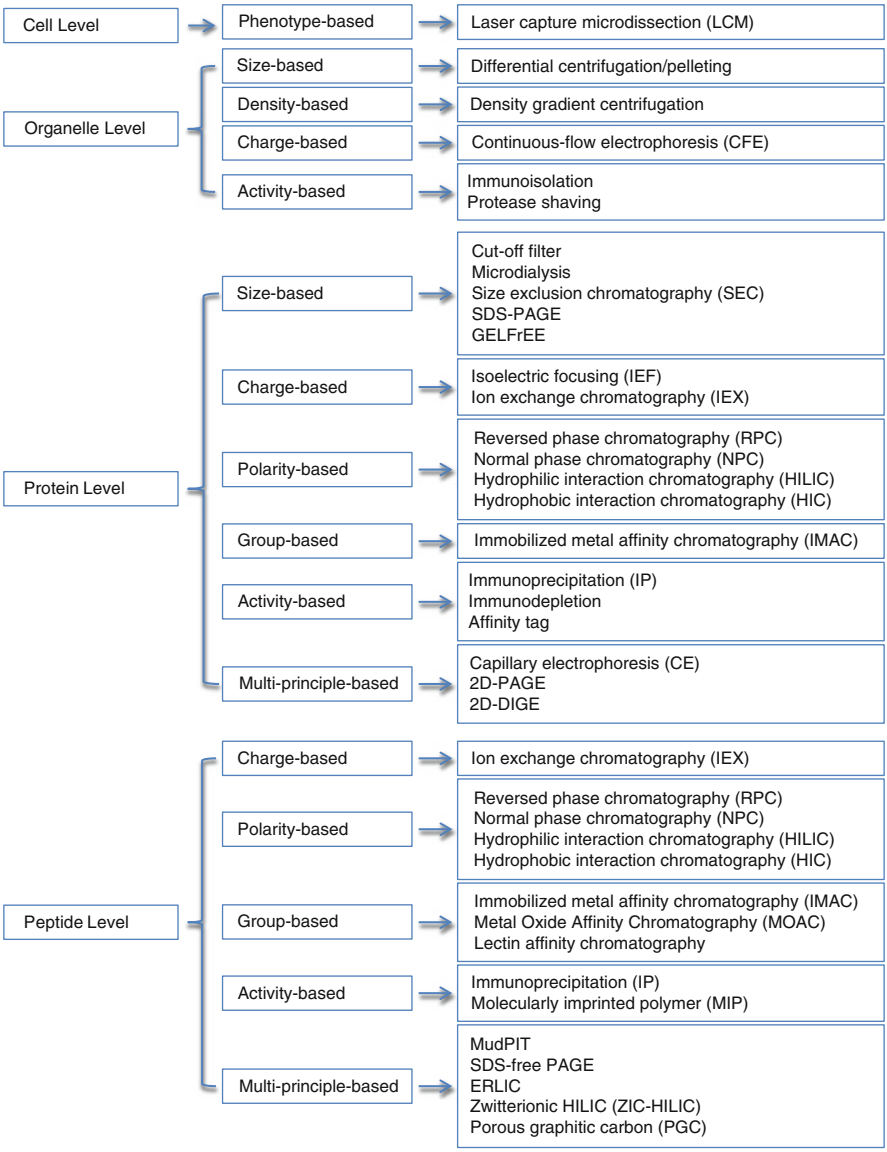


Fig. 6.3 Representative methods of fractionation, enrichment, and separation. According to the mechanisms of fractionation, enrichment, and separation, representative techniques are summarized for different levels, including cell, organelle, protein, and peptide

cell components include cilium, nucleolus, vesicle, centriole, myofibril, ribosome, proteasome, peroxisome, melanosome, lysosome, autophagosome, and acrosome. In addition to the components mentioned above, many more subunits have been defined, such as cell surface proteins, lipid raft-enriched membranes, exosome, and endosome.

Based on the focus of a specific research project, a new subunit may be determined and its name coined.

Various cell organelles differ in size, density, charge, and characteristic protein constituents. Fractionation methods have been developed based on these properties. Some organelles overlap in size, density, and charge. Combination of multiple methods may be required when a specific organelle is to be enriched.

6.2.2.1 Size-Based

Differential Centrifugation/Pelleting: Differential centrifugation or pelleting is an old-fashioned approach to separate organelles principally on the basis of size difference, although differences in density can also be involved in a minor way (Roodyn 1965). Due to the size overlap of different organelles, it is difficult to obtain a highly pure sample in a single step. However, the method is simple, cheap, and rapid. A purer separation of the organelle can be obtained by resuspending and recentrifuging them. This centrifugation/pelleting method has been extensively applied in sample preparation.

6.2.2.2 Density-Based

Density Gradient Centrifugation: The most common density-based separation method involves the formation of a density gradient solute, overlaying samples on the top, and centrifugation until different organelles have reached and stay in a certain gradient solute equal to their own density (de Araujo et al. 2008). For instance, to enrich exosomes in cerebrospinal fluid (CSF), the sample is loaded on the top fraction of a step gradient comprising layers of 2, 1.3, 1.16, 0.8, 0.5, and 0.25 M sucrose. The gradients are centrifuged for 2.5 h at 100,000×*g*. Confirmed by exosome's characteristic proteins, exosomes are enriched in a fraction of a density of 1.10–1.14 g/ml (Street et al. 2012).

6.2.2.3 Charge-Based

Continuous-Flow Electrophoresis (CFE): Due to the presence of acidic and basic groups on an organelle's surface, organelles are charged at neutral pH, allowing migration in an electric field. Though this method provides some distinct advantages over centrifugation, it has not gained wide popularity for organelle fractionation, because the equipment is expensive and has been used for relatively specialized tasks, not for routine isolations. It has been applied in the isolation of lysosomes and endosomes (Graham 1993).

6.2.2.4 Activity-Based

Immunoisolation: In this method, the organelle-specific protein is bound to an antibody, and the antibody is attached to a solid support, such as magnetic beads. When

the magnetic beads are isolated, the attached organelle is enriched. For example, to isolate the neuronal porosome complex from synaptosomes and brain tissue, SNAP-25-specific antibody conjugated to protein A Sepharose® has been applied (Lee et al. 2012a). To enrich synaptic vesicles, the monoclonal SV2 antibody conjugated to the magnetic beads has been used (Burre et al. 2007).

Protease Shaving: A successful strategy for identification of surface proteins is "shaving" intact living cells with proteases. Trypsin is added to live cells suspended in PBS containing 30% sucrose (pH 7.4), and the digestion is carried out at 37 °C. The digestion mixture is centrifuged, and the supernatant is collected using 0.22-μm pore-size filters. The collection contains the proteins on the cell surface, called "surfaceome" or "surfomes" (Olaya-Abril et al. 2012).

6.2.3 Protein Fractionation, Enrichment, and Separation

6.2.3.1 Size-Based

Cutoff Filter: Filtration devices with a certain nominal molecular weight cutoff (NMWC) have been commercially available for many years and are used extensively. The principle is very simple: proteins with molecular weights greater than the NMWC are retained, and proteins with molecular weights lower than the NMWC pass through the filter membrane. Four commercially available filter membranes, Microcon (Millipore), Centrisart (Sartorius), Amicon Ultra (Millipore), and Vivaspin (Sartorius), have been evaluated to enrich low molecular weight proteins from human plasma (Greening and Simpson 2010).

Microdialysis: To exchange and desalt various media with high yield and low loss at very small sample volumes, a microdialysis tool, MicroDialyzer, has been designed to sample volumes from 10 to 100 μl. Equipped with low molecular weight cutoff (MWCO) membranes, MicroDialyzer can sufficiently reduce salt, urea, and detergent content within a short time to enrich proteins (Maischak et al. 2012).

Size-Exclusion Chromatography (SEC): SEC is a chromatographic method in which proteins are separated by their size. Small proteins penetrate every corner of the pore system of the stationary phase, have more volume to traverse, and elute later. On the contrary, large proteins cannot penetrate the pore system, have less volume to traverse, and elute sooner. In a proteomic study of bronchoalveolar lavage, an SEC column from Tosoh Bioscience LLC, TSK GEL G3000 SW, was used to fractionate bronchoalveolar lavage samples. Six fractions per sample were collected and each fraction analyzed by LC-MS/MS (Kosanam et al. 2012).

SDS-PAGE: SDS (sodium dodecyl sulfate) is a detergent with a negative charge (anionic). When proteins are solubilized with SDS, they will bind SDS in amounts proportional to their relative molecular masses, enabling all proteins to

acquire the same charge-to-mass ratio and migrate toward the positive pole when placed in an electric field. Polyacrylamide gel electrophoresis (PAGE) is a process in which proteins are pulled through the polyacrylamide gel by an electromotive force. Polyacrylamide consists of a labyrinth of tunnels through a meshwork of cross-linked fibers composed of acrylamide monomers, restraining larger molecules from migrating as fast as smaller molecules. The separation of proteins by SDS-PAGE is dependent almost entirely on the differences in their size that is determined by the relative molecular mass of protein. Normally, a sample is loaded in gel lanes in an SDS-PAGE, the lanes are cut into multiple slices containing separated protein bands, and each slice is processed for LC-MS/MS analysis. A term, geLC-MS/MS, has been coined for this procedure (Vasilj et al. 2012).

Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE): In a GELFrEE device, a polyacrylamide gel column is used to separate proteins according to their intrinsic molecular weights. Unlike conventional analytical gels, GELFrEE enables proteins that elute from the gel column to be recovered in solution. As a preparative gel electrophoresis device, it affords rapid and broad mass range proteome fractionation (Tran and Doucette 2008).

6.2.3.2 Charge-Based

Isoelectric Focusing (IEF): Proteins contain both acidic and basic amino acids. Thus, the overall charge on a protein is determined by pH of its surrounding environment. At a certain pH, a protein carries no charge, because its constituent negative and positive charges are equal. That pH is the isoelectric point (pI) of the protein. Proteins are positively charged at pH values below their pI and negatively at higher pH values. Introduced by Kolin (1954), IEF is an electrophoretic separation based on the pIs of proteins. IEF has experienced three phases: gel rods, immobilized pH gradients (IPGs), and ZoomIEF (Issaq and Veenstra 2008; Zuo and Speicher 2000). The polyacrylamide gel rods are formed in glass or plastic tubes and contain carrier ampholytes that form a fluid pH gradient in an electric field; in IPGs, the ampholytes (immobilines) are attached to acrylamide molecules and cast into the gels to form a fixed pH gradient; using narrow-range IPG strips, ZoomIEF allows a larger number of proteins to separate, because proteins are spread out over a greater physical distance.

Ion-Exchange Chromatography (IEX): IEX is the most popular method for protein separation, including anion-exchange chromatography (AX) and cation-exchange chromatography (CX). In AX, negatively charged proteins are attracted to a positively charged solid support. Conversely, in CX, positively charged proteins are attracted to a negatively charged solid support. Proteins are separated based on differences between their overall charges. For example, a novel anion-exchange resin suitable for both discovery research and clinical manufacturing purposes has been reported (Porat et al. 2012). The effects of resin ligand density on cation-exchange

chromatography performance in preparative monoclonal antibody purification processes have been evaluated (Fogle et al. 2012).

6.2.3.3 Polarity-Based

Reversed-Phase Chromatography (RPC): RPC has a nonpolar stationary phase and an aqueous and moderately polar mobile phase. Proteins with a larger hydrophobic surface area are retained longer, and proteins with higher polar surface area are retained to a lesser extent. RPC is the most extensively applied and published technology in protein separation (Sheng et al. 2011).

Normal-Phase Chromatography (NPC): In contrast to RPC, NRP has a polar stationary surface and a nonpolar, nonaqueous mobile phase. In this method, proteins are separated based on their adsorption to the polar stationary surface. Adsorption strengths increase with increased protein polarity. For instance, NPC has been applied in peptide separation for the identification and quantification of glycoproteins (Ding et al. 2009).

Hydrophilic Interaction Chromatography (HILIC): The mechanisms of HILIC and NPC are different. In NPC, proteins are in equilibrium between a state of adsorption to the polar stationary phase and dissolution in the nonpolar mobile phase, whereas in HILIC, proteins partition between a water-enriched layer, immobilized on the hydrophilic stationary phase, and the less polar mobile phase. Retention and selectivity of stationary phases for hydrophilic interaction chromatography and its application in proteomics have been reviewed (Guo and Gaiki 2011; Boersema et al. 2008).

Hydrophobic Interaction Chromatography (HIC): HIC is very similar to RPC because both are based on hydrophobic interactions between proteins and the hydrophobic stationary phase. However, the surface of an RPC medium is usually more hydrophobic than that of an HIC medium. Several major differences exist in HIC and RPC selectivity. Research has found that some hydrophilic and hydrophobic proteins are retained differently on HIC and RPC columns (Fausnaugh et al. 1984).

6.2.3.4 Group-Based

Proteins with posttranslational modifications, such as phosphorylation and glycosylation, are not readily enriched according to their modification type using size-, charge-, and polarity-based methods. However, their special chemical groups permit chemical group-based approaches to enrich proteins of interest.

Immobilized Metal Affinity Chromatography (IMAC): The principle of IMAC is based on the interaction between transition metal cations (Co^{2+}, Cu^{2+}, Ni^{2+}, Zn^{2+}) and certain superficial protein residues (histidine, cysteines, and tryptophan) through the formation of chelates. The development of IMAC and its most important applications have been reviewed (Block et al. 2009).

6.2.3.5 Activity-Based

Activity-based protein enrichment is a functional proteomic technology that uses specially designed antibodies or other chemical probes that react with mechanistically related specific proteins, protein complexes, or classes of enzymes. Activity-based protein profiling (ABPP) has been used to characterize enzyme function directly in native biological systems on a global scale (Cravatt et al. 2008).

Immunoprecipitation (IP): IP is an activity-based protein enrichment approach that is applied extensively in biomedical research. An antibody specifically binds to a particular protein and then is precipitated, isolating the protein from the solution. This process easily enriches a particular protein or protein complex from a sample containing thousands of different proteins (Layton et al. 2012). However, the majority of proteins usually identified in IP experiments are nonspecific binders, because the solid matrices, e.g., agarose, sepharose, and magnetic beads, are the main contributors to nonspecific binding. Researchers using this approach should design and optimize an IP experiment to identify and overcome possible pitfalls (ten Have et al. 2011).

Immunodepletion: Immunodepletion is a method for removing target proteins from a mixture. In serum, 98% of the protein mass is composed of approximately 22 proteins. Most depletion kits are primarily immobilized antibodies against these high-abundance proteins. This strategy is similar with IP, but the enrichment lies in the unbound, low-abundance proteins. As in the IP method, nonspecific binding is an issue that can lead to the concomitant removal of some nontargeted proteins of potential interest (Bellei et al. 2011).

Affinity Tag: Affinity tag technology was first developed by Beckwith's group in 1980 (Shuman et al. 1980). In this approach, the gene of the protein of interest is cloned into vectors that contain a tag DNA. Expression of the cloned gene results in a tag fusion protein. In the labeling reaction, the tag substrate of choice is covalently attached to the tag. Based on the labeling of the tag substrate, various applications have been reported, such as protein pulldown. For instance, to identify plasma membrane proteins that undergo retrograde transport to the trans-Golgi network (TGN), benzylguanine-tagged plasma membrane proteins that are subsequently targeted to the retrograde route are covalently captured by a TGN-localized SNAP-tagged fusion protein. This protein enrichment method is novel for the study of retrograde protein trafficking (Shi et al. 2012).

6.2.3.6 Multiprinciple-Based

Capillary Electrophoresis (CE): CE is designed to separate ions by their size to charge ratio in the interior of a small capillary filled with an electrolyte. If two ions have the same size, the one with a greater charge will move faster in an electric field. If two ions have the same charge, the smaller ion has a greater migration

rate. Six separation modes of capillary electrophoresis are available, including capillary zone electrophoresis (CZE, normally referred to as CE), capillary gel electrophoresis (CGE), capillary electrochromatography (CEC), micellar electrokinetic capillary chromatography (MEKC), capillary isoelectric focusing (cIEF), and capillary isotachophoresis (CITP). Advances in interface development of CE-MS and its application in proteomics have been reviewed lately (Ramautar et al. 2012).

Two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE): As describe above, IEF is charge-based and SDS-PAGE is size-based. Therefore, the combination of both principles applied in two-dimensional gel electrophoresis is charge-size-based. The history of 2D-PAGE has been introduced by Issaq and Veenstra (2008). 2D-PAGE was a foundational tool in proteomic research in the past decades. However, conventional 2D-PAGE methods suffer from a few shortcomings involving accuracy, scale, and throughput. Proteins of similar pI and denatured molecular weight are resolved at the same physical location in the gel, called co-migration. This makes it impossible to accurately determine the relative abundance of an individual protein within a mixed spot. Reproducibility also has been an issue with 2D-PAGE, affecting the comparison of two different protein mixtures. Lower-abundance proteins, hydrophobic proteins, and proteins with extreme molecular weight and/or pI cannot be resolved easily by 2D-PAGE. Also, 2D-PAGE suffers from a narrow dynamic detection range and limited sensitivity. 2D-PAGE methods are time-consuming and labor-intensive for protein separation, resulting in low-throughput analysis. Consequently, many methods have been developed to overcome these limitations.

Two-Dimensional Differential In-Gel Electrophoresis (2D-DIGE): To improve the reproducibility, dynamic range, and sensitivity, 2D-DIGE was developed in 1997 (Unlu et al. 1997). Typically, samples for comparison are labeled with different cyanine fluorescent dyes and loaded on one gel to eliminate any error related to gel misalignment and improve the reproducibility. Because fluorescent dye is used, the dynamic range and sensitivity also increases. However, 2D-DIGE does not overcome the accuracy issues mentioned in 2D-PAGE, is costly, and is not compatible with subsequent steps of sample preparation for MS analysis.

Other Advances: A fully automated electrophoresis system for rapid protein analysis has been developed. Using this system, complete 2D separation can be achieved within 1.5 h (Hiratsuka et al. 2007). However, only 662 protein spots were detected from a mouse liver sample using this system. The high-throughput value of this system is diminished by the low number of proteins detected. Microfluidic 2D-PAGE and 2D-DIGE separation systems have been reported for separations of protein complexes (Yang et al. 2009; Emrich et al. 2007). Many other methods have been developed and announced, such as nonequilibrium pH gel electrophoresis (NEPHGE) (Lopez 1999) and ZoomIEF (Zuo and Speicher 2000). However, they have not been applied extensively in proteomics labs. Even though improvements have been achieved by those methods, convenience and cost hamper the application of new techniques.

Summary: In addition to separating proteins, 2-DE enables the relative quantification of the proteins by calculating the intensity of protein spots, which will be reviewed in protein quantification section. Although developments have been achieved to overcome the previously mentioned limitations of 2D-PAGE, co-migration and low-throughput remain as issues limiting the routine application of 2D-PAGE in proteomics. The current trend in proteomics is that less and less 2-DE is applied in global proteomic projects but remains useful for identification of new proteins, mutated proteins, and proteins with posttranslational modifications.

6.2.4 Peptide Fractionation, Enrichment, and Separation

6.2.4.1 Charge-Based

Ion-Exchange Chromatography (IEX): IEX is the most popular method for protein separation as described above. However, IEX has been applied more extensively in peptide separation, because most proteomic analyses are carried out at the peptide level. The separation principle is the same for both protein and peptide. Usually, IEX is used as the first dimension of the two-dimensional liquid chromatography-tandem mass spectrometry workflow (2D LC-MS/MS).

6.2.4.2 Polarity-Based

Reversed-Phase Chromatography (RPC): RPC can be used in protein separation as described above, but it is the most popular method for peptide separation. RPC is used as the second dimension of the two-dimensional liquid chromatography-tandem mass spectrometry workflow (2D LC-MS/MS). Recently, a reversed-phase column of small superficially porous particles has been successfully applied in the separation of peptides. The particles are very highly efficient because of their extremely narrow particle size distribution and higher density, which are able to form efficient and uniformly packed beds (Schuster et al. 2012).

Besides RPC, normal-phase chromatography (NPC) (Ding et al. 2007), hydrophilic interaction chromatography (HILIC) (Gilar et al. 2011), and hydrophobic interaction chromatography (HIC) (Lienqueo et al. 2007) have been used in peptide separation as well. According to the popularity of their application in peptide separation, their rank should be RPC, HILIC, NPC, and HIC.

6.2.4.3 Group-Based

Immobilized Metal Affinity Chromatography (IMAC): IMAC is the most widely used method for the enrichment of phosphopeptides and has been applied to a broad range of samples. Numerous research studies and reviews regarding this technique have been published (Kagedal 2011; Block et al. 2009).

Metal Oxide Affinity Chromatography (MOAC): Historically, IMAC was the first widely used technique for phosphopeptide enrichment. MOAC, essentially with TiO_2 as support, was developed later as an alternative method. The sequential use of the two methods has been proposed to increase the number of phosphopeptides purified. However, a comparison of IMAC and MOAC for phosphopeptide enrichment by column chromatography highly recommends the application of only disposable IMAC columns to avoid storing Fe^{3+}-activated IMAC over a long period (Negroni et al. 2012).

Lectin Affinity Chromatography: Among the glycopeptide enrichment techniques, lectin affinity chromatography is the most widely used. Various lectins can be used specifically to enrich different glycopeptides according to their glycan structures. Concanavalin A (Con A) lectin has frequently been used for the enrichment of N-glycopeptides with mannose. Wheat germ agglutinin (WGA) lectin specifically binds glycopeptides with N-acetyl glucosamine (GlcNAc) and sialic acid (Xu et al. 2009; Ozohanics et al. 2012).

Porous graphitized carbon and hydrophilic affinity isolation have been reported to capture glycopeptides. However, non-glycosylated peptide binding has been observed when they are applied (Ito et al. 2009).

6.2.4.4 Activity-Based

Immunoprecipitation (IP): IP can be performed at both protein and peptide levels. An antibody can be generated against peptides with a specific sequence and then is applied to enrich those peptides from complex protein lysates. An example where this can be applied successfully is in the immunoaffinity enrichment of K-GG peptides to identify ubiquitination sites. Ubiquitin is a small regulatory protein, consisting of 76 amino acids. Ubiquitination is a posttranslational modification, coupling the C-terminus (-GG) of ubiquitin to lysine (K) on substrate proteins, forming K-GG signature peptides where the C-terminal diglycine of ubiquitin remains covalently attached to the substrate. An antibody against peptides with the sequence $CXXXXXXK^{GG}XXXXXX$ is created, and the K-GG peptides are captured with the antibody (Bustos et al. 2012). IP method has been applied in enrichment of tyrosine phosphorylated peptides also (Rush et al. 2005).

Molecularly Imprinted Polymer (MIP): In IP, the antigen-antibody reaction is a lock-and-key combination. The antigen, a protein or peptide, is a key that fits into a lock, the antibody. In fact, the process of MIP is to make an artificial tiny lock for a specific molecule that serves as a miniature key. The same lock-and-key combinations are formed in IP and MIP. A new MIP designed to bind the side chain of phosphotyrosine has been used as artificial receptors for enrichment of phosphorylated peptides (Helling et al. 2011). In comparison with general

enrichment methods for phosphorylated peptides such as TiO_2-based methods, the pTyr-imprinted polymers offered high selectivity for pTyr-containing peptides down to the low fmol level.

6.2.4.5 Multiprinciple-Based

Multidimensional Protein Identification Technology (MudPIT): MudPIT is the most well-known peptide separation technique in proteomics. The original MudPIT includes a strong cation-exchange (SCX) column as the first-dimensional separation and a reversed-phase (RP) column as the second-dimensional separation (Washburn 2004). In principle, it is a charge-polarity-based technique. Numerous applications and reviews have been published regarding MudPIT (Di Palma et al. 2012; Wu et al. 2012a; Xu et al. 2012b). The idea is simple and straightforward. Based on the peptide separation principle, any two, three, or more principles can be bound together to form a new concept, and a new term can be coined, such as SCX-RPLC (Lee et al. 2012b).

SDS-Free PAGE: SDS-PAGE has been applied extensively in protein separation. In this approach, all proteins are negatively charged and their migration rate toward the anode is based on their size. SDS-PAGE is rarely applied in peptide separation. However, in SDS-free PAGE, proteins and peptides migrate based on their electrophoretic mobility, which is proportional to their charge and inversely proportional to their size. Since SDS-PAGE is the most popular tool for protein separation, very few applications of SDS-free PAGE in protein separation have been reported. SDS-free PAGE is mainly used in peptide fractionation (Ramos et al. 2012).

Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC): ERLIC is also called eHILIC and ion-pair normal phase, providing convenient separations of highly charged peptides that cannot readily be resolved by other means. In ERLIC, basic residues are likely oriented away from the stationary phase, enhancing selectivity for neutral and acidic residues (Alpert 2008). The selectivity for peptides is based on the combination of charge and hydrophilic interactions.

Zwitterionic HILIC (ZIC-HILIC): The ZIC-HILIC column is a zwitterionic type of hydrophilic interaction chromatography column with sulfobetaine functional groups. The mechanism of separation is based on the combination of charge and hydrophilic interactions. It has been applied to the separation of glycopeptides, demonstrating that the ZIC-HILIC separation column has selectivity for sialylated N-glycopeptides (Takegawa et al. 2006).

Porous Graphitic Carbon (PGC): PGC separates peptides in a mixed mode combining both polarity-based and charge-based interactions. PGC withstands extremes of pH and higher temperatures than traditional stationary phases and does not require high levels of nonvolatile salts that must be removed prior to LC-MS analysis. In a comprehensive 2D LC-MS/MS analysis, 40% more peptides have been identified using off-line PGC fractionation compared to SCX (Griffiths et al. 2012).

6.2.5 Protein Extraction

Pressure-Assisted Protein Extraction: High hydrostatic pressure promotes water penetration into the inner core of proteins, reduces the size of protein aggregates, causes denaturation, and greatly improves protein solubilization. A method using heat combined with high hydrostatic pressure (40,000 psi) has been applied for the recovery of intact proteins from formaldehyde-fixed, paraffin-embedded (FFPE) mouse liver, resulting in a fourfold increase in protein extraction efficiency, a threefold increase in the extraction of intact proteins, and up to a 30-fold increase in the number of proteins identified, compared to matched tissue extracted with heat alone (Fowler et al. 2012). Additionally, the number of proteins identified in the FFPE tissue was nearly identical to that of matched fresh-frozen tissue.

6.2.6 Protein Solubilization

Microwave-Assisted Protein Solubilization (MAPS): In MAPS, proteins suspended in a solubilization reagent are subjected to microwave irradiation for 30 s, followed by cooling the sample on ice to room temperature and then intermittent homogenization by vortex. With this method, proteins are dissolved in reagents with high efficiency and become more susceptible to trypsin digestion, compared to other conventional protein solubilization techniques (Ye and Li 2012).

6.2.7 Protein Reduction/Alkylation

Volatile Reduction/Alkylation: Most protein identification methods require protein digestion (limited proteolysis). To increase the efficiency of digestion, protein disulfide bonds must be disrupted by reduction. But free sulfhydryl groups are highly reactive and will spontaneously oxidize with other sulfhydryl groups. Thus, sulfhydryls are routinely blocked by alkylation. A one-step procedure for the reduction and alkylation of cysteine residues using volatile reagents triethylphosphine and iodoethanol has been developed (Hale et al. 2004). The greatest advantage is that the excess reagent can be removed by evaporation in a relatively short time frame, reducing analytical variability compared with other procedures using desalting approaches.

6.2.8 Protein Digestion

Pressure-Assisted Proteolysis Using a Syringe: In this approach, the solution mixture of protein and trypsin is transferred to the cylindrical tube of a 3-mL syringe. The syringe is loaded onto a syringe pump, applying a pressure of approximately 6 atm to the sample solution. Using this method, greater numbers of peptides have been observed in 30 min

of pressure-assisted digestion than in overnight atmospheric pressure digestion (Yang et al. 2010). At atmospheric pressure 138 MPa, proteins are digested, obtaining a 20-fold increased protein recovery and improved reproducibility (Freeman and Ivanov 2011).

Infrared-Assisted Proteolysis: In this case, the protein and trypsin mixture is digested under an IR lamp at 250 W in an IR-assisted proteolysis system. The digestion time is reduced significantly to only 5 min. The sequence coverage is better than that obtained by conventional in-solution tryptic digestion (Bao et al. 2009).

Microwave-Assisted Proteolysis: Denatured proteins are aspirated into an enzyme tip and placed into a microwave oven with exhaust module for 2 min at 70 W. High sequence coverage has been obtained using this method. The major advantages of in-tip digestion are to easily handle a small amount of sample and allow high-throughput analysis (Hahn et al. 2009).

Immobilized Enzyme Reactor (IMER): Compared to in-solution digestion, IMER has a number of advantages: low degree of autodigestion even at high enzyme concentration, easy isolation and removal from the digested fragments prior to MS, and rapid enzymatic reaction. Various solid supports have been used for protease-immobilization, such as monolith (Spross and Sinz 2010), glass (Fan and Chen 2007), magnetic nanoparticles (Li et al. 2007), polymethyl methacrylate (PMMA) (Liu et al. 2010), polydimethylsiloxane (PDMS), and polytetrafluoroethylene (PTFE) microtubes (Yamaguchi et al. 2009). IMER has been described as a promising tool for high-throughput proteome profiling. However, it is designed to handle a small amount of sample, which limits the number of proteins identified at a large scale. Only 541 proteins were identified from 5-μg rat liver extract using a metal-ion chelate immobilized enzyme reactor (IMER) supported on organic-inorganic hybrid silica monolith (Ma et al. 2011).

Ultrasonic Wave: Two different ultrasonic energy sources, the sonoreactor and the ultrasonic probe, have been used for enzymatic digestion of proteins for protein identification by MALDI-TOF MS. Both ultrasonic energy sources have successfully digested proteins (Rial-Otero et al. 2007).

Laser-Assisted IMER: Lasers with an emitting wavelength have high penetration ability and photothermal effects, exciting overtone or harmonic vibrations of chemical bonds within the organic tissue components to generate more protein cleavage sites exposed to trypsin. For example, an 808-nm laser has been employed to successfully enhance the proteolysis efficiency of monolithic immobilized enzyme microreactor (Zhang et al. 2011).

Outer Membrane Protease T (OmpT): While most traditional enzymatic approaches generate predominantly small peptides less than 2 kDa, OmpT, which cleaves between less common dibasic sites, produces a distribution with a greater number of peptides more than 3 kDa (>6.3 kDa on average) (Wu et al. 2012b). Large OmpT peptides are able to differentiate closely related protein isoforms and detect many posttranslational modifications. A proteomic experiment using this enzymatic digestion is called “middle-down” proteomics.

Multiple Enzymes: Due to its cleavage specificity and the occurrence of lysine and arginine in a protein sequence, trypsin alone is not able to achieve high sequence coverage for most proteins. To obtain high sequence coverage, multiple enzymes, such as trypsin, GluC, and AspN, are applied in the digestion step (Hoelz et al. 2006).

6.2.9 Integration of the Sample Processing Workflow

Filter-Aided Sample Preparation (FASP): Before the term FASP was coined in 2009 in Nature Methods (Wisniewski et al. 2009), an almost identical method had been reported in 2005 in proteomics (Liebler and Ham 2009). In FASP, a sample solubilized in 4% SDS is concentrated into microliter volumes by an ultrafiltration device. The filter acts as a "proteomic reactor" for detergent removal, buffer exchange, chemical modification, and protein digestion (Wisniewski et al. 2009). The key feature of the method is the ability of the filter membrane to enrich and purify proteins first, then peptides. This removes those substances that would otherwise interfere with subsequent peptide separation and detection. However, nonspecific binding and protein or peptide losses make this method subject to a substantial decline in the number of identifications at low sample loads. Even at higher sample loads, digestion efficiencies and peptide recoveries are variable (Liebler and Ham 2009).

Precipitation/On-Pellet Digestion: In this strategy, a strong, detergent-containing buffer is employed to extract proteins. The detergent, protease inhibitors, and nonprotein matrix components are removed by precipitation of the proteins using acetone. The protein pellet is dissolved by a 4-h tryptic digestion under vigorous agitation and then reduced, alkylated, and digested into its full complement of tryptic peptides with additional trypsin (Duan et al. 2009). Using precipitation, protein recovery and reproducibility must be considered before application of this method.

Single-Tube Sample Preparation: Membrane proteins are solubilized with the volatile surfactant perfluorooctanoic acid (PFOA) that solubilizes membrane proteins as effectively as sodium dodecyl sulfate (SDS) and does not significantly inhibit trypsin activity at a concentration up to 0.5% (w/v). Protein reduction and alkylation are carried out with volatile reagents, so they can be easily removed by evaporation, enabling a single-tube shotgun proteomics method. In this approach, no proteins/peptides are lost in any experimental steps (Kadiyala et al. 2010).

6.3 Protein Identification

Before tandem mass spectrometry was applied, Edman degradation and peptide mass fingerprinting (PMF) were used to sequence amino acids in a peptide. However, Edman degradation is a low-throughput and time-consuming approach, because it is characterized by short peptide reads (limited to about 30 AA) (Bandeira et al. 2008). Using peptide mass fingerprinting, identification becomes more difficult in cases of

complex mixtures and contaminated samples (Fremout et al. 2012). Currently, MS/MS dominates protein identification technology due to its high-throughput capability and fewer limitations. Therefore, only tandem MS-based identification techniques are discussed in this section.

6.3.1 Ion Source

Matrix-Assisted Laser Desorption/Ionization (MALDI): MALDI was the first ionization technique applied in proteomics. Matrix absorbs energy from the laser and is ionized, a proton is transferred to the analyte molecules from excited matrix molecules, and ionization of analyte molecules occurs (Fenselau 1997). Because MALDI is not able to couple HPLC analysis online, it generally has been replaced by electrospray ionization (ESI).

Electrospray Ionization (ESI): Currently, ESI is the most extensively applied ionization technique in proteomics, due to its ability to produce gaseous ions directly from an aqueous or aqueous/organic solvent system by creating a fine spray of highly charged droplets in the presence of a strong electric field (Siuzdak 1994). HPLC and tandem mass spectrometry, LC-MS/MS, is the most widely used analytical approach in proteomics.

Surface-Activated Chemical Ionization (SACI): SACI first published in 2003 (Cristoni et al. 2003) is an ionization method that significantly improves ionization efficiency and sensitivity. It has been applied in quantitative phosphorylation analysis (Cirulli et al. 2012), human serum samples (Sogno et al. 2012), and protein extraction from leaf samples (Finiguerra et al. 2010). However, more experimental comparisons between SACI and electrospray ionization (ESI) are needed to determine whether SACI or ESI is more capable in proteomic analysis.

In-Spray Supercharging: Because multiple charging assists the observation of large molecular ions with narrow mass-to-charge ratio (m/z) range mass analyzers, it is possible to analyze intact protein ions using low-to-medium resolution ion trap mass spectrometers (IT MS). A novel approach, in-spray supercharging, has been developed to increase the average charge state distribution of peptides and proteins by introducing supercharged reagents directly into the electrospray's Taylor cone (Miladinovic et al. 2012).

6.3.2 Mass Analyzer

6.3.2.1 m/z-Based

Time-of-Flight (TOF): TOF is a method of mass spectrometry in which an ion's mass-to-charge ratio is determined via a time measurement. MALDI-TOF was the original and most popular platform for protein identification before the prevalence of LC-MS/MS.

Quadrupole Mass Analyzer: A typical quadrupole mass analyzer includes four circular rods set parallel to each other. The quadrupole filters ions based on their *m*/*z*. In practice, a linear series of three quadrupoles are used: the first (Q1) and third (Q3) quadrupoles act as mass filters, and the middle (q2) quadrupole is employed as a collision cell. Quadrupole instruments are often reasonably priced, but they cannot perform multiple-stage mass spectrometry (MSn) like the ion trap (Wong and Cooks 1997).

Linear Ion Trap: The linear ion trap is a quadrupole ion trap (Hager 2002). This analyzer is the core component in one of the most prominent mass spectrometers in proteomics, the LTQ from Thermo Scientific. Advantages of the linear trap include increased ion storage capacity, faster scan times, and simplicity of construction.

Fourier Transform Ion Cyclotron Resonance (FT-ICR): FT-ICR provides the highest resolution and greatest accuracy among all mass analyzers (Marshall et al. 1998). However, the FT-ICR mass spectrometer is the most expensive instrument, enabling fewer laboratories to afford it. Therefore, it has not been used extensively in proteomics.

Orbitrap: In terms of resolution and mass accuracy, the Orbitrap analyzer is second only to the FT-ICR mass spectrometer and more affordable (Makarov 2000). Currently, the Orbitrap mass spectrometer is one of the most popular instruments in proteomics.

6.3.2.2 Mobility-Based

Ion Mobility (IM): IM is a technique used to separate ionized molecules in the gas phase based on their mobility in a carrier buffer gas. In an LC-ESI-IM-CID-MS approach, mixtures of proteins are separated by reversed-phase LC, eluted proteins are electrosprayed into the gas phase, and the gas-phase ions are separated based on differences in their gas-phase mobilities. CID-MS is performed after IM separation. This particular usage of IM couples with LC and MS to generate fragments of the whole protein (Sowell et al. 2004). It enables the detection of all the possible sequence variations and modifications in the protein of interest thereby increasing the primary sequence coverage and confidence of sequence assignments (Zinnel et al. 2012). This technology has been integrated in AB Sciex, Thermo Scientific, Agilent Technologies, and Waters mass spectrometers.

6.3.3 Fragmentation

Collision-Induced Dissociation (CID): CID, also called collisionally activated dissociation (CAD), is mainstream in proteomics. The molecular ions collide with neutral molecules, such as helium, to generate mainly b and y fragment or product ions. However, CID often selectively cleaves certain interresidue bonds, leaving amino acid sequence gaps. Labile covalent modifications often have the propensity to undergo elimination prior to peptide backbone fragmentation, making it difficult to

determine the site(s) of covalent modification, such as posttranslational modifications (Bakhtiar and Guan 2006).

Higher-Energy C-Trap Dissociation (HCD): In HCD, fragmentation is performed in a collision cell at the far end of the C-trap. The collision cell is supplied with a radio frequency voltage and pressurized with nitrogen gas. The advantages of HCD fragmentation include generation of low molecular weight reporter ions, which is useful in label-based quantification (Olsen et al. 2007).

Electron Capture Dissociation (ECD): In ECD, a beam of electrons is fired into the trapped cloud of sample ions, forming unstable radical ions that then fragment to produce predominantly c and z product ions from peptide precursor ions. ECD produces richer fragment spectra from highly charged peptides and leaves posttranslational modifications (PTMs) intact on the peptide backbone. ECD is almost exclusively performed in FT-ICR instruments (Chalkley 2010).

Electron-Transfer Dissociation (ETD): ETD uses anions, most commonly fluoranthene ions, to transfer electrons to the analyte, forming radical ions that then fragment similarly to ECD. ETD can be performed in quadrupole ion traps, making the technique much more sensitive and affordable than ECD in an FT-ICR instrument (Chalkley 2010). ECD/ETD is complementary to traditional CID. This complementarity enables increased sequence coverage by choosing either CID or ETD, when both fragmentation methods are available on a single instrument.

Infrared Multiphoton Dissociation (IRMPD): IRMPD is able to efficiently perform at low rf amplitudes in quadrupole linear ion traps (QLT), does not induce ion scattering, and require only a single rf device. The use of static QLT rf amplitude and dynamic IRMPD irradiation time is an effective strategy for both identification and quantification of isobaric-tagged peptides in complex mixtures (Ledvina et al. 2012).

Laser-Induced Dissociation (LID): Because only those molecules absorbing energy at the appropriate wavelength will fragment, LID is much more specific than CID and has significantly expanded during the last decade for biomolecular analysis. The UV-Vis laser provides an alternative excitation method that may be distinct from CID and offers complementary structural information (Enjalbert et al. 2011).

Negative Electron-Transfer Dissociation (NETD): NETD fragments peptides and proteins along the backbone at the Cα-C bond, generating a- and x-type product ion fragments. The advantages of NETD include more predictable peptide ion dissociation and compatibility with PTMs (Rumachik et al. 2012).

6.3.4 Mass Spectrometer

A mass spectrometer consists of three major components: an ion source, a mass analyzer, and a detector. The ion source and detector have undergone few significant advances in the past 20 years. As described earlier, the mass analyzer is the “heart” of the mass spectrometer. Many mass spectrometers are named after their analyzers,

such as MALDI-TOF, QTOF, LTQ, and LTQ Orbitrap. Advances in mass spectrometric instrumentation actually are the result of combining various multiple analyzers in a single instrument, called hybrid instrument, thereby providing higher resolution, mass accuracy, sensitivity, and dynamic range.

Orbitrap Elite: Currently, the most powerful instrument used in proteomics is the Orbitrap Elite. It incorporates improvements in three main areas. First, an ion path has been applied in the ion transfer optics to block the line of sight, to achieve more robust operation. Second, the tandem MS acquisition speed of the dual-cell linear ion trap exceeds 12 Hz. Last but most importantly, the Orbitrap analyzer has been redesigned, and its resolution has been increased twofold for the same transient length by employing a compact, high-field Orbitrap analyzer, achieving resolving power up to 240,000 at m/z 400 for a 768-ms transient (Michalski et al. 2012).

6.3.5 Bottom-Up, Middle-Down, and Top-Down

Three strategies have been applied in proteomics research based on the mass spectrometry instrumentation available in a laboratory.

Bottom-Up: In this approach, proteins are digested with a proteolytic enzyme, generating peptides that have much smaller molecular weights. The peptides are ionized and analyzed by the mass spectrometer. Currently, this is the most popular approach to identify and quantify proteins in complex samples, because a mass spectrometer that analyzes peptides in large-scale generally is affordable by most laboratories.

Top-Down: In the top-down approach, intact proteins, rather than peptides, are ionized and analyzed by the mass spectrometer. This approach has been applied and improved since the early period of proteomics. Numerous experiments of intact protein analysis have succeeded, but its success declines in the high-mass region (Wu et al. 2012b). At present, the application of top-down approaches remains rare in routine, large-scale proteomic analysis of complex samples.

Middle-Down: In this approach, developed after bottom-up and top-down approaches, proteins are digested specifically into a few large fragments, instead of small peptides. The large fragments are ionized and analyzed by the mass spectrometer. This strategy has been developed later than bottom-up and top-down. Presently, only a few applications have been published using this approach (Meyer et al. 2011).

Summary: In a complex sample, such as a cellular proteome, proteins may exist in multiple isoforms, diverse modifications, and numerous fragments generated by endogenous protein cleavages. To accurately identify a protein in its fully functional form, intact proteins should be analyzed directly using the mass spectrometer. The top-down approach holds the best prospects for intact protein analysis. However, it currently suffers from two major limitations: (1) high-mass proteins are difficult to

analyze using currently available mass spectrometers, and (2) proteins are multiply charged, generating very complex spectra and limiting the approach to isolated proteins. Before novel ionization methods and analyzers are developed, a routine, large-scale proteomic analysis of complex samples using top-down LC-MS/MS method is unlikely. The top-down approach should be the ultimate goal of proteomics, to accurately and comprehensively detect and identify protein isoforms, modifications, and fragments that occur in a living cell.

The bottom-up approach has been extensively applied due to the availability of affordable instruments. However, in large-scale proteomic analyses, most proteins are identified by only a few peptides, which are minor fractions of the total protein sequence. Consequently, this approach is rarely able to distinguish protein isoforms, fully characterize modifications, and differentiate protein fragments from an intact protein. Bottom-up will gradually be replaced by middle-down when high-end mass spectrometers, such as the Orbitrap Elite, are affordable for more and more proteomic laboratories. It will eventually be replaced by top-down when novel mass spectrometers are available and affordable.

Middle-down is a hybrid approach based on 2–20-kDa peptides and can be performed readily on currently available high-end mass spectrometers (Wu et al. 2012b). Middle-down generates larger peptide fragments than peptides generated by bottom-up, increasing its ability to distinguish protein isoforms, characterize modifications, and differentiate protein fragments from an intact protein. Currently, mass spectrometers able to analyze large peptides are available, but not affordable for many proteomic laboratories. Once such instruments are affordable, middle-down will replace bottom-up. Once instruments for top-down are available and affordable, middle-down will be preplaced by top-down. Currently, it is likely that the application of middle-down approaches will increase until novel instruments are created for the top-down approach.

6.4 Protein Quantification

6.4.1 Array-Based

Protein array analysis is based on binding between a bait molecule and an analyte, which are subsequently detected by a probe. The bait molecule can be an antibody, protein, peptide, drug, nucleic acid, cell, phage, etc. The analyte is a protein. The probe is a molecule with a signal-generating moiety, such as a labeled antibody. The intensity of the signal is proportional to the quantity of an analyte bound to the bait molecule. An image of the spot pattern is captured, analyzed, and interpreted (Liotta et al. 2003).

According to whether the analyte is captured from the solution phase or bound to the solid phase, protein microarrays include two major classes: forward-phase arrays (FPA) and reverse-phase arrays (RPA). In FPA, the analyte is captured from

the solution phase, and the bait molecule, such as an antibody, is immobilized onto the solid support. In contrast, in the RPA format, the analyte is bound to the solid phase and detected by the probe (Liotta et al. 2003).

6.4.1.1 Forward-Phase Array (FPA)

Antibody-Based: An antibody microarray containing 224 antibodies in 32 subarrays is available from Sigma-Aldrich. In each subarray, a single spot with non-labeled bovine serum albumin (BSA) as a negative control, a single spot with Cy3- and Cy5-conjugated BSA as a positive control, and duplicate spots of 7 antibodies are included. This array is able to detect 63 cell signaling proteins, 31 proteins controlling cell cycle, 18 nuclear proteins including some transcription factors, 35 cytoskeleton proteins, 29 apoptosis-related proteins, and 34 neurospecific proteins (Uzdensky et al. 2012).

To analyze the cellular proteomes of 24 pancreatic cancer cell lines and two controls, an antibody microarray made of 810 antibodies that permits the analysis of the expression levels of 741 distinct proteins has been used. In the set of 810 antibodies, 668 of these are affinity-purified, peptide-specific, and polyclonal antibodies produced by Eurogentec from rabbit; 142 of these are antibodies purchased from different commercial providers or obtained from collaborating partners (Alhamdani et al. 2012).

Peptide Array: Arrays of bait peptides are synthesized onto PEG-based membrane supports. The membrane is incubated with cell lysate and subjected to 350–365-nm light to cross-link with interacting proteins. The indirect and nonspecific interactors are removed by high-stringency, denaturing washes (Okada et al. 2012).

6.4.1.2 Reverse-Phase Array (RPA)

Antibody-Based: In reverse-phase protein array (RPPA), sample lysates are spotted onto an array. The array is then hybridized with a specific antibody to recognize the protein of interest. The protein signal is amplified with a biotinylated secondary antibody that binds to the primary antibody. The array is scanned, and the resulting image is quantified and analyzed by an array software (Okada et al. 2012).

6.4.2 2-DE-Based

2-DE-based protein quantification relies on staining. The most popular staining methods include the colorimetric methods, such as Coomassie blue and silver staining, or fluorescent stain methods. Specific staining reagents and methods are available for protein posttranslational modifications such as phosphorylation and glycosylation. A detailed review on protein gel stain methods has been published (Steinberg 2009). In addition to the traditional methods, new staining methods are continually emerging.

Native Protein Fluorescence: Although Coomassie brilliant blue has been widely used in 2D gel visualization with moderate detection sensitivity, drawbacks and limitations frequently arise from the significant background they may generate in subsequent mass spectrometric analysis. A stain-free gel electrophoretic detection approach based on native protein fluorescence has been developed to be a fast gel imaging system (Susnea et al. 2012).

6.4.3 MS-Based

In antibody-based and 2-DE-based quantification methods, protein abundances are determined directly at the protein level. However, in common MS-based methods (bottom-up), protein measurements are completed at the peptide level and then combined to calculate a summarized value for the protein from which they come.

Quantitation of analytes using MS is not new and can be traced back to 1972 (Bertilsson et al. 1972). In that paper, the fragment ion of 5-hydroxyindole-3-acetic acid (5-HIAA) diheptafluorobutyryl methyl ester derivative, *m/z* 538, was chosen to determine the concentration of 5-HIAA. Dideuterium-labeled 5-HIAA was synthesized and used as an internal standard to integrate a standard curve. The developed method required less than 2 min and enabled an accurate determination of 5-HIAA in the range of 2–50 ng/ml of human cerebrospinal fluid (CSF). In this pioneering study, the peak height of 5-HIAA derivative fragment ion was applied to quantify the concentration of 5-HIAA. At the present time, this approach (fragment ion + isotopic-labeled internal standard) continues to be applied extensively in quantification of small-molecule compounds. However, peak area rather than peak height is used for quantification (Tsikas et al. 2012; Zanchetti et al. 2012; Heinig et al. 2011). Developments in the last 50 years indicate that peak area is the most reliable measurement for quantification. The approach for small-molecule compounds also has been adopted for proteomics (Schmidt and Urlaub 2012). Because peptides are small-molecule compounds, measuring peak area directly should be considered a reliable method for peptide and protein quantification.

However, in addition to measuring the peptide peak area directly, numerous methods and techniques have been developed for protein quantification. While every approach claims high accuracy and precision (reproducibility), only when its cost is low, it is convenient, and it has fewer limitations can one claim it is more advanced than other approaches. Various publications have reviewed the principle, method, application, and pros and cons (Xie et al. 2011; Brewis and Brennan 2010).

In timeline of proteomic technological development, relative quantification appeared before absolute quantification and label-based methods were developed before label-free methods. Therefore, the introduction of each method will follow that sequence in this section.

6.4.3.1 Relative Quantification

Label-Based

In the early stages of proteomic technology, a few notable laboratories developed stable isotope-labeling methods and coined terms for their methods. Following that theme, many new terms have arisen. According to the developmental sequence of labeling methods, some of these are summarized in Table 6.1, including their name, stage of proteomic sample preparation workflow where labeling occurs, and pertinent characteristics.

Metabolic Labeling

^{15}N Labeling: The first labeling method for protein quantification was published in June 1999 (Oda et al. 1999). Wild-type and mutant cells are grown in medium containing natural isotopes of nitrogen and the same medium enriched in ^{15}N, respectively. After an appropriate growing period, the cells are combined, and the proteins are extracted and separated by 2-DE. Gel spots of interest were excised and are subjected to digestion, and the resulting peptide fragments were extracted and analyzed by MS. The intensity of pairs of peaks from unlabeled and ^{15}N-labeled peptides is used to determine relative protein abundance in the two cell types. This procedure can be applicable to any cell system that can be grown in isotopically enriched media. Other stable isotopes, ^{13}C, ^{18}O, and ^{2}H, could be used as well.

Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC): SILAC was introduced in 2002 (Ong et al. 2002). In this method, cell cultures are adapted to normal leucine or Leu-d3 media at the start of the experiment, and proteins are mixed directly after lysis and subjected to protein identification and quantification procedures. In contrast to the method used in ^{15}N-labeling described above, the labeling of peptides using stable isotope-labeled amino acids, rather than elements, is specific to its sequence, and the mass differential between two states can be specified more directly.

SILAC has been considered to be able to facilitate accurate and reliable quantitative proteomics by the introduction of the stable isotope-labeling amino acids in cell culture, combined with high-resolution mass spectrometry. However, several major sources of quantification errors exist, i.e., incomplete incorporation of isotopic amino acids, arginine-to-proline conversion, and experimental errors in final sample mixing. A label-swap replication of SILAC experiments has been developed to effectively correct experimental errors (Park et al. 2012).

Chemical Labeling

Isotope-Coded Affinity Tags (ICAT): To overcome the limitations of the ^{15}N-labeling approach, ICAT was developed in October 1999 (Gygi et al. 1999). ICAT is a combination of labeling and enrichment. The side chains of cysteinyl residues in two

Table 6.1 The comparison and summary of labeling methods according to their development sequence

		Labeling stage[a]				
Timeline	Name	ML	CL	EL	PL	Reagent, group, and amino acid
Jun 1999	^{15}N labeling	X				^{15}N-labeled (>96%) rich cell growth media
						n/a[b]
						n/a
Oct 1999	ICAT		X			ICAT reagent
						Thiol
						The side chains of cysteine residues
Mar 2001	GIST				X	*d*3-Methanol
						Carboxylic acid
						C-termini and the side chains of aspartic and glutamic acid
Sep 2001	^{18}O labeling			X		$H_2{}^{18}O$
						n/a
						Cysteine or lysine when trypsin is applied
2002	SILAC	X				Heavy isotope-labeled amino acids
						n/a
						The responding amino acids
2003	TMT				X	TMT reagent
						Amine
						N-termini and the side chains of lysine residues
2004	iTRAQ				X	iTRAQ reagent
						Amine
						N-termini and the side chains of lysine residues
2005	ICPL		X			*d*4-Nicotinoyloxy-succinimide
						Amine
						The side chains of lysine residues
2008	mTRAQ				X	mTRAQ reagent
						Amine
						N-termini and the side chains of lysine residues

[a]Before a peptide is analyzed by MS, several steps are included, such as protein synthesis, extraction, and digestion. Labeling could be carried out at any stage. Metabolic labeling (*ML*) happens during protein synthesis, chemical labeling (*CL*) occurs after protein extraction and before protein digestion, enzymatic labeling (*EL*) takes place during protein digestion, and post-digestion labeling (*PL*) happens after protein digestion

[b]n/a not applicable

reduced protein samples are derivatized with the isotopically light and heavy form of the ICAT reagent, respectively. The two samples are mixed and enzymatically cleaved to generate peptides. Due to the tags on the labeled peptides, the cysteine-containing peptides are enriched by avidin affinity chromatography. Finally, the peptides are identified and quantified by LC-MS/MS.

Isotope-Coded Protein Label (ICPL): ICPL was first published in 2005 (Schmidt et al. 2005). In ICPL, stable isotope labeling occurs at free amino groups of intact proteins. Thus, it is applicable to any protein sample, including extracts from tissues or body fluids, and compatible with all separation methods employed in proteomic studies.

Enzymatic Labeling

^{18}O Labeling: While metabolic labeling using ^{15}N requires labeling to occur in vivo and chemical labeling requires a protein to contain a certain amino acid such as cysteine or lysine, enzymatic labeling does not require in vivo metabolic labeling or a protein to contain a certain amino acid. Therefore, it is broadly applicable to all protein separation methods. The first ^{18}O-labeling method was published in September 2001 (Munchbach et al. 2000). ^{18}O isotopic labeling is carried out when the proteins are digested in a solution containing 50% $H_2{}^{18}O$.

Post-digestion Labeling

Global Internal Standard Technology (GIST): GIST was developed in March 2001 and so-named in 2002 (Chakraborty and Regnier 2002; Goodlett et al. 2001), enabling isotopic tagging of all peptides obtained after enzymatic cleavage of protein samples, because labeling of peptides converts carboxylic acids, which are present on the side chains of aspartic and glutamic acid as well as the carboxyl terminus, to their corresponding methyl esters.

Tandem Mass Spectrometry Tag (TMT): TMT is the first isobaric labeling method and was published in 2003 (Thompson et al. 2003). The tag reagent of TMT consists of three components: a unique mass reporter (tag) that is used for quantification, a mass normalizer that balances the mass of the tag to make all the tags equivalent in total mass, and a reactive moiety that cross-links to peptides. Due to the identical mass of the tags, the same peptide in all samples has an identical mass and LC retention time, simplifying analysis and potentially increasing analytical accuracy and precision.

A new TMT technique called cysTMT has been developed and is used to enrich TMT-labeled cysteine-containing peptides (Murray et al. 2012).

Isobaric Tags for Relative and Absolute Quantitation (iTRAQ): The principle of iTRAQ is the same as TMT, while the tag is a little different. iTRAQ reagents react with amino groups at the N-termini of peptides and with lysine side chains. The chemical structure of the iTRAQ reagent includes a charged reporter group (*m*/*z* 114–117) that is unique to each of the four reagents, a mass balance group (28–31 Da) to have an identical total mass of 145 Da for the four isobaric tags, and a peptide reactive group. The first iTRAQ method was published in 2004 and the term coined in 2005 (DeSouza et al. 2005; Ross et al. 2004). The original iTRAQ had four tags, allowing for the simultaneous quantification of up to four samples. Currently, eight tags are commercially available, permitting multiplexing of up to eight samples in a single experiment.

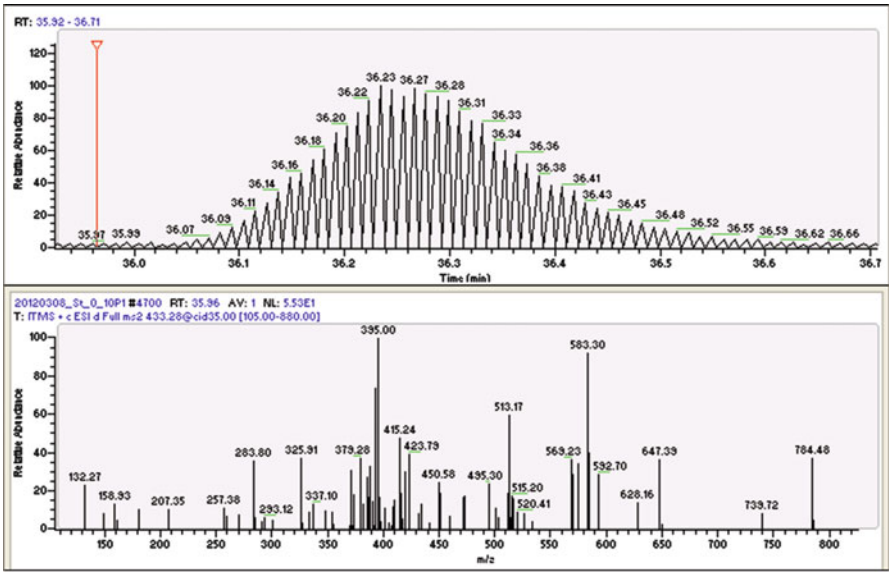


Fig. 6.4 An example of a peptide detected by LC-MS/MS. A peptide elution on an LC column and the peptide ion MS/MS spectrum provides its intensity at each time point, the MS/MS spectra, and the product ions and their intensity in the MS/MS spectrum. All the information has been used individually or conjointly in numerous label-free methods, such as spectral counting, ion intensity, MS/MS fragment ion intensity, and a combination of spectral counting and ion intensity measurements

Similar to cysTMT, cysTRAQ has been developed to enable iTRAQ reagent quantitation of peptides fractionated based on presence of a cysteine (Tambor et al. 2012). The weaknesses and strengths of iTRAQ for relative quantification and other technical issues have been discussed and published (Evans et al. 2012).

MRM-Compatible Tags for Relative and Absolute Quantitation (mTRAQ): mTRAQ reagents are MRM-compatible versions of reagents. Unlike the iTRAQ labels, the mTRAQ labels are designed to be nonisobaric to maximize possible differences in the multiple reaction monitoring (MRM) transitions. The first mTRAQ application was reported in 2008 (DeSouza et al. 2008). The mTRAQ methodology relies on MRM to target tryptic peptides from the protein of interest.

Label-Free

As shown in Fig. 6.4, a peptide ion provides useful information, including its intensity at each time point, the MS/MS spectra, and the ions and their intensity in the MS/MS spectrum. Using this information, different label-free methods have been developed, including spectral counting (SC), ion intensity, MS/MS fragment ion intensity, and a combination of spectral counting and ion intensity measurements.

Spectral Counting

The principle of spectral counting is very simple: the number of mass spectra identified for a protein is used as a measure of the protein's abundance (Hendrickson et al. 2006). Although spectral counting has been applied in many different biological complexes, protein quantification using spectral counting is challenging, because (1) using dynamic exclusion of ions in data acquisition to obtain more MS/MS fragments of low-abundance peptides dramatically affects spectral acquisition, and (2) co-eluted peptides in liquid chromatography competing for MS/MS analysis influence the spectral acquisition (Lai et al. 2011).

Ion Intensity

As mentioned above, the application of peak area that is based on ion intensity has a long history in the quantification of small-molecule compounds. There are no reasons to exclude the application of this method in peptide and protein quantification. On the contrary, it should be the first and best choice, because it is well-established and has been used extensively.

In this approach it is important to emphasize that protein quantification is not carried out directly but based on peptide quantification. Unfortunately, in comparative studies, individual peptide fold changes often exhibit a different fold change than other peptides from the same protein (Lai et al. 2011). Several possible explanations for this phenomenon are as follows: (1) some peptides have greater variation than others under the same chromatographic conditions, (2) PTM variably affects the relative abundance of unmodified peptides, (3) peptide sharing among diverse proteins causes inconsistent effects on some peptides, and (4) differential regulation of isoforms, misidentification, and misquantification also may occur (Lai et al. 2011; Erhard and Zimmer 2012). To detect outlier peptides, Lai et al. (2011) used a peptide intensity correlation calculation, and Erhard and Zimmer (2012) applied a complicated algorithm. Both studies identified the problem and developed a method to remove outlier peptides to improve quantitation.

MS/MS Fragment Ion Intensity

Two formats of MS/MS fragment ion intensity, spectral index (SI) and summed MS/MS TIC (SMT), replace spectral counting in the normalized spectral abundance factor (NSAF) formula, resulting in two algorithms, abbreviated as NSI and NSMT, respectively. Both NSI and NSMT improve overall accuracy over NSAF for the relative quantification of proteomes (Wu et al. 2012c).

Combination of Multiple Strategies

A label-free quantitative algorithm by combining measurements of spectral counting, ion intensity, and peak area on 1D PAGE-based proteomics has been reported (Gao et al. 2008). However, this method is developed for 1D PAGE-based LC-MS/MS analysis.

In protein quantification, many methods attempt to improve accuracy by complicating their algorithm or using normalization. Researchers must realize that the improvement is limited when the quality of raw data is poor.

Selected Reaction Monitoring (SRM)/Multiple Reaction Monitoring (MRM)

Traditional label-free quantification methods quantify hundreds to thousands of proteins in a mixture. SRM, also known as MRM, is a targeted protein quantification method. SRM/MRM is not a new mass spectrometry technique, but its application in proteomics is emerging as a complement to untargeted shotgun methods and is particularly useful in absolute quantification. The principle, application, advance, and future of SRM have been reviewed (Picotti and Aebersold 2012; Maiolica et al. 2012).

6.4.3.2 Absolute Quantification

As described above, SRM/MRM is able to relatively quantify proteins in a complex sample. When isotopically labeled, recombinant proteins or synthesized peptides are used as internal standards, and SRM/MRM is able to absolutely quantify proteins. The history, principle, and workflow of absolute quantification have been introduced and reviewed (Schmidt and Urlaub 2012; Bronsema et al. 2012).

Absolute Quantification (AQUA): In AQUA, peptides are synthesized with incorporated stable isotopes as internal standards, having identical physicochemical properties but are distinguished by a mass shift compared to target peptides generated by proteolysis. The ratio of the area under the curve (AUC) is used to determine concentration of a protein in the cell lysate (Gerber et al. 2003).

Protein Standard Absolute Quantification (PSAQ™): PSAQ™ was developed in 2007 to perform absolute quantification of target proteins in MS experiments using stable isotope-labeled full-length proteins as internal standards (Adrait et al. 2012). Quantification is carried out by comparing the protein MS signal to that of the isotope-labeled standard. Because PSAQ standards and their target proteins share the same biochemical properties, PSAQ standards can be added into the samples at the earliest stages of analysis to enhance protein quantification accuracy due to losses that may occur during sample prefractionation and/or incomplete proteolysis.

6.5 Protein Structure

Structural proteomics is a high-throughput endeavor for solving three-dimensional (3D) structures of proteins, protein complexes, and small-molecule-protein complexes. Structural proteomics is used to determine the detailed structure of the interfaces between proteins that may be important drug targets and the interactions

between proteins and ligands. Relatively few studies in structural proteomics have been published.

Biomolecular NMR plays a critical role in structural proteomics. Recently, a new strategy using 2D HSQC-type experiments has been developed to establish the complete backbone (^{1}H, ^{15}N, $^{13}C^{\alpha}$, and $^{13}C'$) assignment of small well-folded proteins in less than a day for high-throughput structural proteomics (Kumar et al. 2012).

Mass spectrometry is becoming an important tool for studying protein structure. In mass spectrometry-based structural proteomics, proteins are modified or labeled and subsequently analyzed with mass spectrometric techniques to characterize protein structures. A brief overview of structural proteomics methodologies has been published, including the common techniques used in structural proteomics, such as cross-linking, photoaffinity labeling, limited proteolysis, chemical protein modification, and hydrogen/deuterium exchange (Serpa et al. 2012). Importantly, none of these methods alone are able to provide complete structural information, but a "combination" of these complementary approaches provides vital information.

6.6 Protein Function

Functional proteomics is actually interaction proteomics, focusing on protein interactions with a "bait" of interest, such as a protein, DNA, RNA, or small molecule. Two main types of methods have been applied in interaction proteomics, including genetic-based methods and MS-based methods.

6.6.1 Genetic-Based

Yeast Two-Hybrid (Y2H): The Y2H system is a genetic method that detects protein-protein interactions. In this method, two parts of a transcription factor are separated: DNA-binding domain linked to a known "bait" protein X and the activation domain linked to an unknown "fish" protein Y. Only when X and Y proteins are bound, a complete transcription factor reconstitutes, allowing the corresponding reporter gene expression, which is readily detected. One application is to detect new interactions of a protein with a known function by library screening (Lecrenier et al. 1998).

6.6.2 MS-Based

Affinity Purification Coupled to MS: MS-based methods for protein interaction analysis are identical to protein identification and quantification in the MS analysis but differ with respect to sample preparation. Immunoprecipitation (IP)

described earlier in protein fractionation, enrichment, and separation is the most common method used in MS-based protein interaction studies. An enriched protein complex is identified and quantified with MS, thereby revealing the interactions.

6.7 Conclusions

Generally, proteomic techniques can be classified into three generations. In the first generation, proteins are quantified by 2-DE and identified by MALDI-TOF MS. In the second generation, also called bottom-up proteomics, proteins are identified and quantified by LC-MS/MS at the peptide level. Currently, proteomics is in the second generation. In the next generation of proteomics, also called top-down proteomics, proteins are identified and quantified by LC-MS/MS at the protein level. Top-down proteomics has been successful in the analysis of simple samples, but not in the routine, large-scale analysis of complex samples. Before a new generation of mass spectrometers is available for large-scale top-down proteomics, an alternative approach, called middle-down, is a great choice to improve protein identification and quantification.

As far as protein absolute quantification is performed, a reliable and reproducible method should be validated. According to the "Guidance for Industry: Bioanalytical Method Validation" from the FDA, a particular method used for quantitative measurement of analytes in a given biological matrix has to be validated reliable and reproducible for the intended use. Accuracy, precision, selectivity, sensitivity, reproducibility, and stability are the fundamental parameters. These requirements are of the utmost importance in assessing the performance of the particular method. Some of them should be considered when any method is used in sample preparation, protein identification, relative protein quantification, protein structure, and protein function. Cost, as in time, instrumentation, and operation, is another major concern in method development. In reality, every technique should be developed to improve performance and/or reduce cost.

Advances in sample preparation mainly focus on sample fractionation, enrichment, and separation. Current trends aim to obtain a certain type of peptide and protein or a specific organelle. Other advances in sample preparation involve improved protein extraction, solubilization, reduction/alkylation, and digestion. Advances in protein identification are the result of improvements in ionization, mass analyzers, and fragmentation techniques. Advances in protein quantification are aimed at improving accuracy, precision, selectivity, sensitivity, reproducibility, and stability. Structural and functional proteomics have begun but require continued development.

In summary, advances in each step of proteomic technology are directed at building a method that is able to generate accurate and global results with high-throughput capability.

Acknowledgments The author is grateful to Dr. Frank A. Witzmann (Indiana University School of Medicine, Indianapolis, IN) for his support and critical reading of the manuscript. This work was supported in large part by NIEHS RC2ES018810 and NIGMS R01GM085218 (FAW).

References

Adrait A, Lebert D, Trauchessec M, Dupuis A, Louwagie M, Masselon C, Jaquinod M, Chevalier B, Vandenesch F, Garin J, et al. Development of a protein standard absolute quantification (PSAQ) assay for the quantification of *Staphylococcus aureus* enterotoxin A in serum. J Proteomics. 2012;75:3041–9.

Alhamdani MS, Youns M, Buchholz M, Gress TM, Beckers MC, Marechal D, Bauer A, Schroder C, Hoheisel JD. Immunoassay-based proteome profiling of 24 pancreatic cancer cell lines. J Proteomics. 2012;75:3747–59.

Alpert AJ. Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. Anal Chem. 2008;80:62–76.

Anderson NL, Anderson NG. Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis. 1998;19:1853–61.

Bakhtiar R, Guan Z. Electron capture dissociation mass spectrometry in characterization of peptides and proteins. Biotechnol Lett. 2006;28:1047–59.

Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR. Automated de novo protein sequencing of monoclonal antibodies. Nat Biotechnol. 2008;26:1336–8.

Bao H, Lui T, Zhang L, Chen G. Infrared-assisted proteolysis using trypsin-immobilized silica microspheres for peptide mapping. Proteomics. 2009;9:1114–17.

Bellei E, Bergamini S, Monari E, Fantoni LI, Cuoghi A, Ozben T, Tomasi A. High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. Amino Acids. 2011;40:145–56.

Bertilsson L, Atkinson Jr AJ, Althaus JR, Harfast A, Lindgren JE, Holmstedt B. Quantitative determination of 5-hydroxyindole-3-acetic acid in cerebrospinal fluid by gas chromatography-mass spectrometry. Anal Chem. 1972;44:1434–8.

Block H, Maertens B, Spriestersbach A, Brinker N, Kubicek J, Fabis R, Labahn J, Schafer F. Immobilized-metal affinity chromatography (IMAC): a review. Methods Enzymol. 2009;463:439–73.

Boersema PJ, Mohammed S, Heck AJ. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. Anal Bioanal Chem. 2008;391:151–9.

Brewis IA, Brennan P. Proteomics technologies for the global identification and quantification of proteins. Adv Protein Chem Struct Biol. 2010;80:1–44.

Bronsema KJ, Bischoff R, van de Merbel NC. Internal standards in the quantitative determination of protein biopharmaceuticals using liquid chromatography coupled to mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;893–894:1–14.

Burre J, Zimmermann H, Volknandt W. Immunoisolation and subfractionation of synaptic vesicle proteins. Anal Biochem. 2007;362:172–81.

Bustos D, Bakalarski CE, Yang Y, Peng J, Kirkpatrick DS. Characterizing ubiquitination sites by peptide based immunoaffinity enrichment. Mol Cell Proteomics. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22729469.

Chakraborty A, Regnier FE. Global internal standard technology for comparative proteomics. J Chromatogr A. 2002;949:173–84.

Chalkley R. Instrumentation for LC-MS/MS in proteomics. Methods Mol Biol. 2010;658:47–60.

Cheng L, Zhang S, Maclennan GT, Williamson SR, Davidson DD, Wang M, Jones TD, Lopez-Beltran A, Montironi R. Laser-assisted microdissection in translational research: theory, technical considerations, and future applications. Appl Immunohistochem Mol Morphol. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22495368.

Cirulli C, Coccetti P, Alberghina L, Tripodi F. A surface-activated chemical ionization approach allows quantitative phosphorylation analysis of the cyclin-dependent kinase inhibitor Sic1 phosphorylated on Ser201. Rapid Commun Mass Spectrom. 2012;26:1527–32.

Corthals GL, Nelson PS. Large-scale proteomics and its future impact on medicine. Pharmacogenomics J. 2001;1:15–9.

Cravatt BF, Wright AT, Kozarich JW. Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. Annu Rev Biochem. 2008;77:383–414.

Cristoni S, Bernardi LR, Biunno I, Tubaro M, Guidugli F. Surface-activated no-discharge atmospheric pressure chemical ionization. Rapid Commun Mass Spectrom. 2003;17:1973–81.

de Araujo ME, Huber LA, Stasyk T. Isolation of endocitic organelles by density gradient centrifugation. Methods Mol Biol. 2008;424:317–31.

DeSouza L, Diehl G, Rodrigues MJ, Guo J, Romaschin AD, Colgan TJ, Siu KW. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. J Proteome Res. 2005;4:377–86.

DeSouza LV, Taylor AM, Li W, Minkoff MS, Romaschin AD, Colgan TJ, Siu KW. Multiple reaction monitoring of mTRAQ-labeled peptides enables absolute quantification of endogenous levels of a potential cancer marker in cancerous and normal endometrial tissues. J Proteome Res. 2008;7:3525–34.

Di Palma S, Hennrich ML, Heck AJ, Mohammed S. Recent advances in peptide separation by multidimensional liquid chromatography for proteome analysis. J Proteomics. 2012;75:3791–813.

Ding W, Hill JJ, Kelly J. Selective enrichment of glycopeptides from glycoprotein digests using ion-pairing normal-phase liquid chromatography. Anal Chem. 2007;79:8891–9.

Ding W, Nothaft H, Szymanski CM, Kelly J. Identification and quantification of glycoproteins using ion-pairing normal-phase liquid chromatography and mass spectrometry. Mol Cell Proteomics. 2009;8:2170–85.

Duan X, Young R, Straubinger RM, Page B, Cao J, Wang H, Yu H, Canty JM, Qu J. A straightforward and highly efficient precipitation/on-pellet digestion procedure coupled with a long gradient nano-LC separation and orbitrap mass spectrometry for label-free expression profiling of the swine heart mitochondrial proteome. J Proteome Res. 2009;8:2838–50.

Emrich CA, Medintz IL, Chu WK, Mathies RA. Microfabricated two-dimensional electrophoresis device for differential protein expression profiling. Anal Chem. 2007;79:7360–6.

Enjalbert Q, Simon R, Salvador A, Antoine R, Redon S, Ayhan MM, Darbour F, Chambert S, Bretonniere Y, Dugourd P, Lemoine J. Photo-SRM: laser-induced dissociation improves detection selectivity of selected reaction monitoring mode. Rapid Commun Mass Spectrom. 2011;25:3375–81.

Erhard F, Zimmer R. Detecting outlier peptides in quantitative high-throughput mass spectrometry data. J Proteomics. 2012;75:3230–9.

Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, Pandhal J, Smith D, Pham TK, Karunakaran E, et al. An insight into iTRAQ: where do we stand now? Anal Bioanal Chem. 2012;404(4):1011–27.

Fan H, Chen G. Fiber-packed channel bioreactor for microfluidic protein digestion. Proteomics. 2007;7:3445–9.

Fausnaugh JL, Kennedy LA, Regnier FE. Comparison of hydrophobic-interaction and reversed-phase chromatography of proteins. J Chromatogr. 1984;317:141–55.

Fenselau C. MALDI MS and strategies for protein analysis. Anal Chem. 1997;69:661A–5.

Finiguerra A, Spadafora A, Filadoro D, Mazzuca S. Surface-activated chemical ionization time-of-flight mass spectrometry and labeling-free approach: two powerful tools for the analysis of complex plant functional proteome profiles. Rapid Commun Mass Spectrom. 2010;24:1155–60.

Fogle J, Mohan N, Cheung E, Persson J. Effects of resin ligand density on yield and impurity clearance in preparative cation exchange chromatography. I. Mechanistic evaluation. J Chromatogr A. 2012;1225:62–9.

Fowler CB, Waybright TJ, Veenstra TD, O'Leary TJ, Mason JT. Pressure-assisted protein extraction: a novel method for recovering proteins from archival tissue for proteomic analysis. J Proteome Res. 2012;11:2602–8.

Freeman E, Ivanov AR. Proteomics under pressure: development of essential sample preparation techniques in proteomics using ultrahigh hydrostatic pressure. J Proteome Res. 2011;10:5536–46.
Fremout W, Dhaenens M, Saverwyns S, Sanyova J, Vandenabeele P, Deforce D, Moens L. Development of a dedicated peptide tandem mass spectral library for conservation science. Anal Chim Acta. 2012;728:39–48.
Gao BB, Stuart L, Feener EP. Label-free quantitative analysis of one-dimensional PAGE LC/MS/MS proteome: application on angiotensin II-stimulated smooth muscle cells secretome. Mol Cell Proteomics. 2008;7:2399–409.
Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003;100:6940–5.
Gilar M, Yu YQ, Ahn J, Xie H, Han H, Ying W, Qian X. Characterization of glycoprotein digests with hydrophilic interaction chromatography and mass spectrometry. Anal Biochem. 2011;417:80–8.
Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R, Kolker E. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. Rapid Commun Mass Spectrom. 2001;15:1214–21.
Graham JM. Continuous-flow electrophoresis. Application to the isolation of lysosomes and endosomes. Methods Mol Biol. 1993;19:41–9.
Greening DW, Simpson RJ. A centrifugal ultrafiltration strategy for isolating the low-molecular weight (<or=25 K) component of human plasma proteome. J Proteomics. 2010;73:637–48.
Griffiths JR, Perkins S, Connolly Y, Zhang L, Holland M, Barattini V, Pereira L, Edge A, Ritchie H, Smith DL. The utility of porous graphitic carbon as a stationary phase in proteomics workflows: two-dimensional chromatography of complex peptide samples. J Chromatogr A. 2012;1232:276–80.
Guo Y, Gaiki S. Retention and selectivity of stationary phases for hydrophilic interaction chromatography. J Chromatogr A. 2011;1218:5920–38.
Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999;17:994–9.
Hager JW. A new linear ion trap mass spectrometer. Rapid Commun Mass Spectrom. 2002;16:512–26.
Hahn HW, Rainer M, Ringer T, Huck CW, Bonn GK. Ultrafast microwave-assisted in-tip digestion of proteins. J Proteome Res. 2009;8:4225–30.
Hale JE, Butler JP, Gelfanova V, You JS, Knierman MD. A simplified procedure for the reduction and alkylation of cysteine residues in proteins prior to proteolytic digestion and mass spectral analysis. Anal Biochem. 2004;333:174–81.
Heinig K, Wirz T, Bucheli F, Monin V, Gloge A. Sensitive determination of a pharmaceutical compound and its metabolites in human plasma by ultra-high performance liquid chromatography-tandem mass spectrometry with on-line solid-phase extraction. J Pharm Biomed Anal. 2011;54:742–9.
Helling S, Shinde S, Brosseron F, Schnabel A, Muller T, Meyer HE, Marcus K, Sellergren B. Ultratrace enrichment of tyrosine phosphorylated peptides on an imprinted polymer. Anal Chem. 2011;83(5):1862–5.
Hendrickson EL, Xia Q, Wang T, Leigh JA, Hackett M. Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. Analyst. 2006;131:1335–41.
Hiratsuka A, Kinoshita H, Maruo Y, Takahashi K, Akutsu S, Hayashida C, Sakairi K, Usui K, Shiseki K, Inamochi H, et al. Fully automated two-dimensional electrophoresis system for high-throughput protein analysis. Anal Chem. 2007;79:5730–9.
Hoelz DJ, Arnold RJ, Dobrolecki LE, Abdel-Aziz W, Loehrer AP, Novotny MV, Schnaper L, Hickey RJ, Malkas LH. The discovery of labile methyl esters on proliferating cell nuclear antigen by MS/MS. Proteomics. 2006;6:4808–16.
Huber LA. Is proteomics heading in the wrong direction? Nat Rev Mol Cell Biol. 2003;4:74–80.
Hutchens TW, Yip TT. New desorption strategies for the mass-spectrometric analysis of macromolecules. Rapid Commun Mass Spectrom. 1993;7:576–80.

Issaq H, Veenstra T. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE): advances and perspectives. Biotechniques. 2008;44:697–8, 700.
Ito S, Hayama K, Hirabayashi J. Enrichment strategies for glycopeptides. Methods Mol Biol. 2009;534:195–203.
Kadiyala CS, Tomechko SE, Miyagi M. Perfluorooctanoic acid for shotgun proteomics. PLoS One. 2010;5:e15332.
Kagedal L. Immobilized metal ion affinity chromatography. Methods Biochem Anal. 2011;54:183–201.
Knepper MA. Proteomics and the kidney. J Am Soc Nephrol. 2002;13:1398–408.
Kolin A. Separation and concentration of proteins in a pH field combined with an electric field. J Chem Phys. 1954;22:1628–9.
Kosanam H, Sato M, Batruch I, Smith C, Keshavjee S, Liu M, Diamandis EP. Differential proteomic analysis of bronchoalveolar lavage fluid from lung transplant patients with and without chronic graft dysfunction. Clin Biochem. 2012;45:223–30.
Kumar D, Borkar A, Hosur RV. Facile backbone (1H, 15N, 13Ca, and 13C′) assignment of 13C/15N-labeled proteins using orthogonal projection planes of HNN and HN(C)N experiments and its automation. Magn Reson Chem. 2012;50:357–63.
Kurban G, Gallie BL, Leveridge M, Evans A, Rushlow D, Matevski D, Gupta R, Finelli A, Jewett MA. Needle core biopsies provide ample material for genomic and proteomic studies of kidney cancer: observations on DNA, RNA, protein extractions and VHL mutation detection. Pathol Res Pract. 2012;208:22–31.
Lai X, Wang L, Tang H, Witzmann FA. A novel alignment method and multiple filters for exclusion of unqualified peptides to enhance label-free quantification using peptide intensity in LC-MS/MS. J Proteome Res. 2011;10:4799–812.
Layton MJ, Faux MC, Church NL, Catimel B, Kershaw NJ, Kapp EA, Nowell C, Coates JL, Burgess AW, Simpson RJ. Identification of a Wnt-induced protein complex by affinity proteomics using an antibody that recognizes a sub-population of beta-catenin. Biochim Biophys Acta. 2012;1824:925–37.
Lecrenier N, Foury F, Goffeau A. Two-hybrid systematic screening of the yeast proteome. BioEssays News Rev Mol Cellular Dev Biol. 1998;20:1–5.
Ledvina AR, Lee MV, McAlister GC, Westphall MS, Coon JJ. Infrared multiphoton dissociation for quantitative shotgun proteomics. Anal Chem. 2012;84:4513–19.
Lee JS, Jeremic A, Shin L, Cho WJ, Chen X, Jena BP. Neuronal porosome proteome: molecular dynamics and architecture. J Proteomics. 2012a;75:3952–62.
Lee JH, Hyung SW, Mun DG, Jung HJ, Kim H, Lee H, Kim SJ, Park KS, Moore RJ, Smith RD, Lee SW. A fully automated multi-functional ultrahigh pressure liquid chromatography system for advanced proteome analyses. J Proteome Res. 2012b;11(8):4373–81.
Li Y, Xu X, Deng C, Yang P, Zhang X. Immobilization of trypsin on superparamagnetic nanoparticles for rapid and effective proteolysis. J Proteome Res. 2007;6:3849–55.
Liebler DC, Ham AJ. Spin filter-based sample preparation for shotgun proteomics. Nat Methods. 2009;6:785; author reply 785-786.
Lienqueo ME, Salazar O, Henriquez K, Calado CR, Fonseca LP, Cabral JM. Prediction of retention time of cutinases tagged with hydrophobic peptides in hydrophobic interaction chromatography. J Chromatogr A. 2007;1154:460–3.
Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J, Petricoin 3rd EF. Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell. 2003;3:317–25.
Liu T, Bao H, Chen G. Inflation bulb-driven microfluidic reactor for infrared-assisted proteolysis. Electrophoresis. 2010;31:3070–3.
Liu NQ, Braakman RB, Stingl C, Luider TM, Martens JW, Foekens JA, Umar A. Proteomics pipeline for biomarker discovery of laser capture microdissected breast cancer tissue. J Mammary Gland Biol Neoplasia. 2012;17(2):155–64.
Lopez MF. Nonequilibrium pH gel electrophoresis (NEPHGE). Methods Mol Biol. 1999;112:129–31.

Ma J, Hou C, Liang Y, Wang T, Liang Z, Zhang L, Zhang Y. Efficient proteolysis using a regenerable metal-ion chelate immobilized enzyme reactor supported on organic-inorganic hybrid silica monolith. Proteomics. 2011;11:991–5.

Maiolica A, Junger MA, Ezkurdia I, Aebersold R. Targeted proteome investigation via selected reaction monitoring mass spectrometry. J Proteomics. 2012;75:3495–513.

Maischak H, Tautkus B, Kreusch S, Rhode H. Proteomic sample preparation by microdialysis: easy, speedy, and nonselective. Anal Biochem. 2012;424:184–6.

Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem. 2000;72:1156–62.

Marshall AG, Hendrickson CL, Jackson GS. Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev. 1998;17:1–35.

Meyer B, Papasotiriou DG, Karas M. 100% protein sequence coverage: a modern form of surrealism in proteomics. Amino Acids. 2011;41:291–310.

Michalski A, Damoc E, Lange O, Denisov E, Nolting D, Muller M, Viner R, Schwartz J, Remes P, Belford M, et al. Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol Cell Proteomics. 2012;11:O111 013698.

Miladinovic SM, Fornelli L, Lu Y, Piech KM, Girault HH, Tsybin YO. In-spray supercharging of peptides and proteins in electrospray ionization mass spectrometry. Anal Chem. 2012;84(11):4647–51.

Munchbach M, Quadroni M, Miotto G, James P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. Anal Chem. 2000;72:4047–57.

Murray CI, Uhrigshardt H, O'Meally RN, Cole RN, Van Eyk JE. Identification and quantification of S-nitrosylation by cysteine reactive tandem mass tag switch assay. Mol Cell Proteomics. 2012;11:M111 013441.

Negroni L, Claverol S, Rosenbaum J, Chevet E, Bonneu M, Schmitter JM. Comparison of IMAC and MOAC for phosphopeptide enrichment by column chromatography. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;891–892:109–12.

Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci U S A. 1999;96:6591–6.

Okada H, Uezu A, Soderblom EJ, Moseley 3rd MA, Gertler FB, Soderling SH. Peptide array X-linking (PAX): a new peptide-protein identification approach. PLoS One. 2012;7:e37035.

Olaya-Abril A, Gomez-Gascon L, Jimenez-Munguia I, Obando I, Rodriguez-Ortega MJ. Another turn of the screw in shaving gram-positive bacteria: optimization of proteomics surface protein identification in *Streptococcus pneumoniae*. J Proteomics. 2012;75:3733–46.

Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods. 2007;4:709–12.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1:376–86.

Ozohanics O, Turiak L, Drahos L, Vekey K. Comparison of glycopeptide/glycoprotein enrichment techniques. Rapid Commun Mass Spectrom. 2012;26:215–17.

Park SS, Wu WW, Zhou Y, Shen RF, Martin B, Maudsley S. Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC). J Proteomics. 2012;75:3720–32.

Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S. Protein analysis on a proteomic scale. Nature. 2003;422:208–15.

Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012;9:555–66.

Porat A, Winters D, Cai L, Smith S, Abroson F, Tam LT, Shen Z, Hecht R. A novel anion-exchange resin suitable for both discovery research and clinical manufacturing purposes. Prep Biochem Biotechnol. 2012;42:304–21.

Ramautar R, Heemskerk AA, Hensbergen PJ, Deelder AM, Busnel JM, Mayboroda OA. CE-MS for proteomics: advances in interface development and application. J Proteomics. 2012;75:3814–28.

Ramos Y, Besada V, Castellanos-Serra L. Peptide fractionation by SDS-free polyacrylamide gel electrophoresis for proteomic analysis via DF-PAGE. Methods Mol Biol. 2012; 869:197–204.

Rial-Otero R, Carreira RJ, Cordeiro FM, Moro AJ, Santos HM, Vale G, Moura I, Capelo JL. Ultrasonic assisted protein enzymatic digestion for fast protein identification by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Sonoreactor versus ultrasonic probe. J Chromatogr A. 2007;1166:101–7.

Roodyn DB. The classification and partial tabulation of enzyme studies on subcellular fractions isolated by differential centrifuging. Int Rev Cytol. 1965;18:99–190.

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004;3:1154–69.

Rumachik NG, McAlister GC, Russell JD, Bailey DJ, Wenger CD, Coon JJ. Characterizing peptide neutral losses induced by negative electron-transfer dissociation (NETD). J Am Soc Mass Spectrom. 2012;23:718–27.

Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, Zhang H, Zha XM, Polakiewicz RD, Comb MJ. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat Biotechnol. 2005;23:94–101.

Schmidt C, Urlaub H. Absolute quantification of proteins using standard peptides and multiple reaction monitoring. Methods Mol Biol. 2012;893:249–65.

Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics. 2005;5:4–15.

Schuster SA, Boyes BE, Wagner BM, Kirkland JJ. Fast high performance liquid chromatography separations for proteomic applications using fused-core(R) silica particles. J Chromatogr A. 2012;1228:232–41.

Serpa JJ, Parker CE, Petrotchenko EV, Han J, Pan J, Borchers CH. Mass spectrometry-based structural proteomics. Eur J Mass Spectrom. 2012;18:251–67.

Sheng S, Skalnikova H, Meng A, Tra J, Fu Q, Everett A, Van Eyk J. Intact protein separation by one- and two-dimensional liquid chromatography for the comparative proteomic separation of partitioned serum or plasma. Methods Mol Biol. 2011;728:29–46.

Shi G, Azoulay M, Dingli F, Lamaze C, Loew D, Florent JC, Johannes L. SNAP-tag based proteomics approach for the study of the retrograde route. Traffic. 2012;13:914–25.

Shuman HA, Silhavy TJ, Beckwith JR. Labeling of proteins with beta-galactosidase by gene fusion. Identification of a cytoplasmic membrane component of the *Escherichia coli* maltose transport system. J Biol Chem. 1980;255:168–74.

Siuzdak G. The emergence of mass spectrometry in biochemical research. Proc Natl Acad Sci U S A. 1994;91:11290–7.

Sogno I, Conti M, Consonni P, Noonan DM, Albini A. Surface-activated chemical ionization-electrospray ionization source improves biomarker discovery with mass spectrometry. Rapid Commun Mass Spectrom. 2012;26:1213–18.

Sowell RA, Koeniger SL, Valentine SJ, Moon MH, Clemmer DE. Nanoflow LC/IMS-MS and LC/IMS-CID/MS of protein mixtures. J Am Soc Mass Spectrom. 2004;15:1341–53.

Spross J, Sinz A. A capillary monolithic trypsin reactor for efficient protein digestion in online and offline coupling to ESI and MALDI mass spectrometry. Anal Chem. 2010; 82:1434–43.

Steinberg TH. Protein gel staining methods: an introduction and overview. Methods Enzymol. 2009;463:541–63.

Street JM, Barran PE, Mackay CL, Weidt S, Balmforth C, Walsh TS, Chalmers RT, Webb DJ, Dear JW. Identification and proteomic profiling of exosomes in human cerebrospinal fluid. J Transl Med. 2012;10:5.

Susnea I, Bernevic B, Wicke M, Ma L, Liu S, Schellander K, Przybylski M. Application of MALDI-TOF-mass spectrometry to proteome analysis using stain-free gel electrophoresis. Top Curr Chem. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22547356.

Takegawa Y, Deguchi K, Ito H, Keira T, Nakagawa H, Nishimura S. Simple separation of isomeric sialylated N-glycopeptides by a zwitterionic type of hydrophilic interaction chromatography. J Sep Sci. 2006;29:2533–40.

Tambor V, Hunter CL, Seymour SL, Kacerovsky M, Stulik J, Lenco J. CysTRAQ – a combination of iTRAQ and enrichment of cysteinyl peptides for uncovering and quantifying hidden proteomes. J Proteomics. 2012;75:857–67.

ten Have S, Boulon S, Ahmad Y, Lamond AI. Mass spectrometry-based immuno-precipitation proteomics – the user's guide. Proteomics. 2011;11:1153–9.

Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003;75:1895–904.

Tran JC, Doucette AA. Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. Anal Chem. 2008;80:1568–73.

Tsikas D, Suchy MT, Mitschke A, Beckmann B, Gutzki FM. Measurement of nitrite in urine by gas chromatography-mass spectrometry. Methods Mol Biol. 2012;844:277–93.

Unlu M, Morgan ME, Minden JS. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis. 1997;18:2071–7.

Uzdensky A, Kristiansen B, Moan J, Juzeniene A. Dynamics of signaling, cytoskeleton and cell cycle regulation proteins in glioblastoma cells after sub-lethal photodynamic treatment: antibody microarray study. Biochim Biophys Acta. 2012;1820:795–803.

Vasilj A, Gentzel M, Ueberham E, Gebhardt R, Shevchenko A. Tissue proteomics by one-dimensional gel electrophoresis combined with label-free protein quantification. J Proteome Res. 2012;11:3680–9.

Washburn MP. Utilisation of proteomics datasets generated via multidimensional protein identification technology (MudPIT). Brief Funct Genomic Proteomic. 2004;3:280–6.

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology. 1996;14:61–5.

Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009;6:359–62.

Wong PSH, Cooks RG. Ion trap mass spectrometry. Curr Sep. 1997;16:85–92.

Wu Q, Yuan H, Zhang L, Zhang Y. Recent advances on multidimensional liquid chromatography-mass spectrometry for proteomics: from qualitative to quantitative analysis – a review. Anal Chim Acta. 2012a;731:1–10.

Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, Ahlf DR, Early BP, Thomas PM, Sweedler JV, Kelleher NL. A protease for 'middle-down' proteomics. Nat Methods. 2012b;9(8):822–4.

Wu Q, Zhao Q, Liang Z, Qu Y, Zhang L, Zhang Y. NSI and NSMT: usages of MS/MS fragment ion intensity for sensitive differential proteome detection and accurate protein fold change calculation in relative label-free proteome quantification. Analyst. 2012c;137:3146–53.

Xie F, Liu T, Qian WJ, Petyuk VA, Smith RD. Liquid chromatography-mass spectrometry-based quantitative proteomics. J Biol Chem. 2011;286:25443–9.

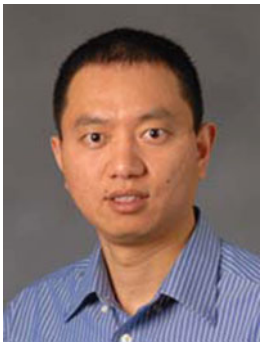
Xu Y, Wu Z, Zhang L, Lu H, Yang P, Webley PA, Zhao D. Highly specific enrichment of glycopeptides using boronic acid-functionalized mesoporous silica. Anal Chem. 2009;81:503–8.

Xu Y, Cao LQ, Jin LY, Chen ZC, Zeng GQ, Tang CE, Li GQ, Duan CJ, Peng F, Xiao ZQ, Li C. Quantitative proteomic study of human lung squamous carcinoma and normal bronchial epithelial acquired by laser capture microdissection. J Biomed Biotechnol. 2012a;2012:510418.

Xu X, Liu K, Fan ZH. Microscale 2D separation systems for proteomic analysis. Expert Rev Proteomics. 2012b;9:135–47.

Yamaguchi H, Miyazaki M, Honda T, Briones-Nagata MP, Arima K, Maeda H. Rapid and efficient proteolysis for proteomic analysis by protease-immobilized microreactor. Electrophoresis. 2009;30:3257–64.

Yang S, Liu J, Lee CS, Devoe DL. Microfluidic 2-D PAGE using multifunctional in situ polyacrylamide gels and discontinuous buffers. Lab Chip. 2009;9:592–9.
Yang HJ, Hong J, Lee S, Shin S, Kim J. Pressure-assisted tryptic digestion using a syringe. Rapid Commun Mass Spectrom. 2010;24:901–8.
Yates 3rd JR. Mass spectrometry. From genomics to proteomics. Trends Genet. 2000;16:5–8.
Ye X, Li L. Microwave-assisted protein solubilization for mass spectrometry-based shotgun proteome analysis. Anal Chem. 2012;84(14):6181–91.
Zanchetti G, Floris I, Piccinotti A, Tameni S, Polettini A. Rapid and robust confirmation and quantification of 11-nor-Delta9-tetrahydrocannabinol-9-carboxylic acid (THC-COOH) in urine by column switching LC-MS-MS analysis. J Mass Spectrom. 2012;47:124–30.
Zhang P, Gao M, Zhu S, Lei J, Zhang X. Rapid and efficient proteolysis through laser-assisted immobilized enzyme reactors. J Chromatogr A. 2011;1218:8567–71.
Ziegler Z. Quantitative proteomics goes global. Anal Chem. 2001;73:251A.
Zinnel NF, Pai PJ, Russell DH. Ion mobility-mass spectrometry (IM-MS) for top-down proteomics: increased dynamic range affords increased sequence coverage. Anal Chem. 2012;84:3390–7.
Zuo X, Speicher DW. A method for global analysis of complex proteomes using sample prefractionation by solution isoelectrofocusing prior to two-dimensional electrophoresis. Anal Biochem. 2000;284:266–78.



Xianyin Lai, Ph.D., Assistant Research Professor, USA Xianyin Lai is an assistant research professor in Department of Cellular & Integrative Physiology, Indiana University School of Medicine. He received his Ph.D. from Peking University, China. His research involves the application of proteomic techniques to investigate "molecular biomarkers" and elucidate "molecular mechanisms" in biomedical research. His experience relates to areas in proteomics, biomarker, biochemistry, separation science, bioanalysis, HPLC, mass spectrometry, and analytical chemistry.

In the past several years, he has been involved in several research projects to uncover the mechanism of the growth and maintenance of renal and liver cysts, assessing the effects of unrefined carbon nanomaterials on barrier epithelial cell protein expression, and performing biomarker research of alcohol abuse using cutting-edge proteomic techniques. Dr. Lai also has developed a label-free quantitative mass spectrometric software platform, IdentiQuantXL™.

Chapter 7
Clinical and Biomedical Mass Spectrometry: New Frontiers in Drug Developments and Diagnosis

Ákos Végvári, Melinda Rezeli, David Erlinge, and György Marko-Varga

Abstract Healthcare systems today are undergoing major restructuring. From the patient's perspective, expectations focusing on high quality treatments for most common diseases – such as cancer, cardiovascular diseases, neurodegenerative diseases, diabetes, and others – have gone unmet in most countries throughout the world. Today, a number of protein expression and analysis platforms is available, which can generate large-scale maps of proteins related to healthy and diseased states. These mass spectrometry-based technologies are used on a daily basis by thousands of research laboratories around the world. The major interest is focused on discovery and validation of novel biomarkers in various diseases, as well as on targeted proteomics where quantification of multiple protein biomarkers is achieved. We present these technological developments in relation to disease diagnosis and treatment and provide two examples where significant progress has been made.

Keywords Mass spectrometry • Prostate cancer • Prostate-specific antigen • Biomarkers • MRM

Á. Végvári, Ph.D. • M. Rezeli, Ph.D.
Division of Clinical Protein Science and Imaging,
Department of Measurement Technology and Industrial Electrical Engineering,
Lund University, Biomedical Center C13, SE-211 84 Lund, Sweden

D. Erlinge, Ph.D.
Department of Cardiology, Lund University, Biomedical Center D12,
SE-211 84 Lund, Sweden

G. Marko-Varga, Ph.D. (✉)
Division of Clinical Protein Science and Imaging, Department of Measurement Technology and Industrial Electrical Engineering, Lund University, Biomedical Center C13,
SE-211 84 Lund, Sweden

1st Department of Surgery, Tokyo Medical University, Tokyo, Japan
e-mail: gyorgy.markovarga@elmat.lth.se

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_7,

7.1 Introduction

Society has an increasing demand and expectations on health-care quality and efficiency, which happens in a time of fast technology development and milestone achievements in science. This opens up for new opportunities that can meet challenges to the medical research community and to further drive the patient-centric and technology-driven research strategy.

The development of new problem-solving biomarkers has a great prospective, in where both the industry and the academic fields are investing and exploring approaches to exploit technology to make innovative discoveries (Anderson 2005; Marko-Varga and Fehniger 2004).

Recently, the protein science research field has developed new and complementary technologies, like protein shotgun sequencing and quantitative mass spectrometry platforms that are readily available and also used in everyday clinical chemistry assays within hospitals around the world (Mann 2009; Olsen et al. 2009; Schmidt et al. 2009). With about 20,300 gene products and their multiple structural variations, involved in disease pathophysiology, measuring these proteins is a real challenge to research society. The coding of the human genome is with its splice-forming variants expected to produce a much higher number of human proteins (Kato et al. 2011a; Rezeli et al. 2011). The exact human protein number is still not known today. However, considering the posttranslational modifications that occur in the body, it is expected to be many millions of proteins. The cellular activity of all these proteins in any given disease state forms a high degree of complexity that holds the Holy Grail of many diseases.

As the expression of proteins in human biofluid is a rich source of clinical patient material, health care has given blood analysis and protein quantitation a high degree of attention over the last decades. Still today, considering blood samples like plasma and serum constitutes an expression range of 10–12 orders of magnitude. This is a real challenge, since there is no analytical technology platform today that can measure such a large dynamic range of any given protein in patient samples (Anderson and Anderson 2002; Domon and Aebersold 2006). The limitation of sensitivity of currently available protein platforms does not allow us to perform the assay. The alternative solution that most laboratories apply is to build the assays in a way that covers 3–5 orders of magnitude and then covers the entire range by crossover solutions that will cover the concentration area of interest.

Right now, there is a large numbers of putative markers that are used for various diagnostic predictions. The objective is to be assessing which combination of markers has the greatest diagnostic and prognostic power. As there is an ever-increasing attention in the clinical field, trying to understand and predict the disease status and malfunctions of patients, the focus on optimal drug treatment of patients has become a central point of attention.

The lifestyle with alcohol consumption and smoking habits constitutes a unique global public health problem with increasing burden on the health-care system due

to alcohol- and smoking-related morbidities. These habits in relation to an alternative lifestyle with exercise and healthy food, such as functional food products entering into the everyday life of people, will have a global impact to the health of our communities. In addition, increased life expectancy is leading to geriatric and chronic illnesses. The predominant diseases in this regard are:

- Cardiovascular disease
- Cancer diseases
- Neurodegenerative diseases
- Obesity and diabetes type 2
- Pulmonary diseases, including chronic obstructive pulmonary disease (COPD)

There is an increasing international interest to strengthen and progress research areas that can aid disease understanding to improve patient care. This will include novel medicines such as "personalized medicine." The targeted drug treatments have opened up a whole new field where patients within a given disease are considered as a heterogeneous group rather than one uniform that can be treated in the same way. As a result, an alternative dedicated solution is targeted for a specific phenotype. This new generation of medications is at the frontline of modern medicine, and most drug companies have the pipeline activated in clinical trials to evaluate and adapt these treatments to new patient groups. Personalized medicine is closely linked to the diagnostic arm of the treatment, where the selection is made. Proteomics and the clinical protein science fields are devoting lots of resources to outline the optimal path, linking treatment and diagnosis together. In this respect, early indication of disease diagnosis along with alternative treatment technologies can utilize both imaging techniques and biomarker diagnostics (Kato et al. 2011b).

Although considerable progresses are being made, there is still an unmet need for solutions how to manage and address best these health-care challenges. These tribulations are well known by governments and health-care institutions. These are reasons why for instance the European Commission has dedicated large-scale research programs to address the developments of disease mechanism research within dedicated research area (Andrejevs et al. 2009), and so has National Institute of Health (NIH) in USA, as well as other sponsoring bodies in the world. In a joint effort between Europe and the USA, there have been common strategies of how systems biology can be useful in cancer research (Aebersold et al. 2009).

7.2 Protein Biomarkers and Targeted Analysis

Patient needs and disease symptoms are very often multifactorial. Consequently, the demands are driven by more than one biological mechanism. In addition, disease presentations are in most cases not a one-point position of a treatment event. On the contrary, it is a situation where multifold of disease presentation forming the ongoing pathophysiology is cooperatively making out these disease effects. These are reasons why the diagnosis of the disease is a highly important part of the health-care

process. Diagnosis as such appears in the very beginning when the patients first meet the doctor. A number of analysis and investigations are performed, most commonly involving clinical chemistry assay measurements. In many cases, imaging tests are made, such as X-ray, CT, PET scan, and MRI. The responsible physician will have consultation throughout the disease progression and treatments, where the diagnosis will be repeated several times.

7.2.1 Disease-Targeted Concepts

Today, the diagnosis is made in the hospital, where the number of available analysis tests are in the order of 400–600 different clinical measures. In these clinical chemistry tests, protein quantitations are preferably performed with an immunoreagent-based assay technique, with a large number of methodologies at hand. At present, ELISA and clinical flow injection assays are the most common, and they are being used in highly automated analysis platforms where robotic instrumentation is state of the art.

As there is an interest in trying to understand multifactorial biology, such as stratification within disease, the systems biology approach is gaining more and more interest. A systematic and holistic approach where the biology is considered as more of a matrix-related event rather than biological event, needs to be treated. The systems biology as a concept builds all of the biology components into one common view, explaining how the system works. As a consequence, it will allow for an opening to a new and efficient future for the patients with alternative treatments, where the entire system is taken into consideration. In this context, the information, which is needed in the clinical field for improved medical treatments, is increasing.

We see the rise of high-density protein arrays (Schwenk et al. 2008; Weissenstein et al. 2006; Rinn et al. 2004; Cho et al. 2006), as well as tissue microarrays (Apweiler et al. 2009; Paavilainen et al. 2010). The technology-driven disease biology cataloguing exercise is a more extensive challenge than was foreseen. To address this, the idea of the Human Proteome Project (HPP) was launched in September 2010 in Sydney at the HUPO World Congress.

The purpose of the international Chromosome-Centric Human Proteome Project (C-HPP) is to map and identify all human proteins that are encoded by the genes on all human chromosomes. In addition, it is the objective of the C-HPP consortium to establish and organize a joint network among the research society, responsible for protein mapping of individual chromosomes, and to identify compelling biological and genetic mechanisms influencing colocated genes and their protein products.

The C-HPP aims to foster integrating these findings, from related molecular omics technology platforms, through collaborations among universities, industries, and private research groups (Legrain et al. 2011; Paik et al. 2012a, b; Marko-Varga et al. 2011).

7.3 Biomarkers

Biomarkers are most well known from the past as surrogate markers that have successfully been used in patient treatment. In general, the biomarker concept is extremely wide in use and application, and it is clear from our current experience that one size does not fit all; instead, there is a clear strategy how biomarkers should be used in combination with patient drug treatment that is focused on *use of the right diagnosis to get the right drug to the right patient at the right time.*

The surrogate markers have high scores both respect to sensitivity and specificity. Clinicians at hospitals that meet patients on an everyday basis will be able to make high efficiency scoring, as the prediction predictive value of surrogate markers is high. As a consequence of poor unsatisfactory developments of surrogate markers, the high-tech science field opened up for the development of biomarkers that are predictive, but to a lower degree. These biomarkers are commonly used as multiple marker assay approaches. Most of the new technology platforms generate multitude of data from a biology of interest, where the informatics and mathematics groups are needed to help elucidate the value and the interpretation of these results. The health-care area and clinical organizations are utilizing these developments and are expecting payback, by means of new and useful protein assays, in the treatment of patients. In addition, there are extensive developments ongoing, where ELISA tests are being challenged by new mass spectrometry-based assay principles. Biomarker perspective and overviews where new technology progresses play a central role in cancer, and other disease areas were recently presented by joint research initiatives. Biomarkers are utilized for various purposes within the health-care sector, as well as for drug developments where the type of biomarkers is linked to the development phase within the drug development process:

(a) Primary biomarkers, commonly with low abundant protein expression.
(b) Secondary biomarkers, commonly with medium abundant protein expression, often with an indirect biomarker that is a resulting outcome of the signaling pathway biology.
(c) Tertiary biomarkers are, commonly with low-to-medium abundant protein expression.

These processing steps of biomarker developments are depicted and presented in Fig. 7.1. The concept whereby these biomarkers are used in clinical drug studies is shown in Table 7.1. Here, the biomarkers have specific aims and outcomes as shown in Table 7.1. In early drug studies, the drug impact needs to be verified by a proof of mechanism biomarker, followed in a later phase study by a proof of principle biomarker. In the final phase of drug developments, in clinical phase 3 drug studies, the proof of concept biomarkers are applied. At this stage, the biomarker needs to provide evidence with a given assay that the drug concept is valid. At this stage, the number of patients participating in the study is also high, and consequently the statistical evidence and outcome of the biomarker data are critical.

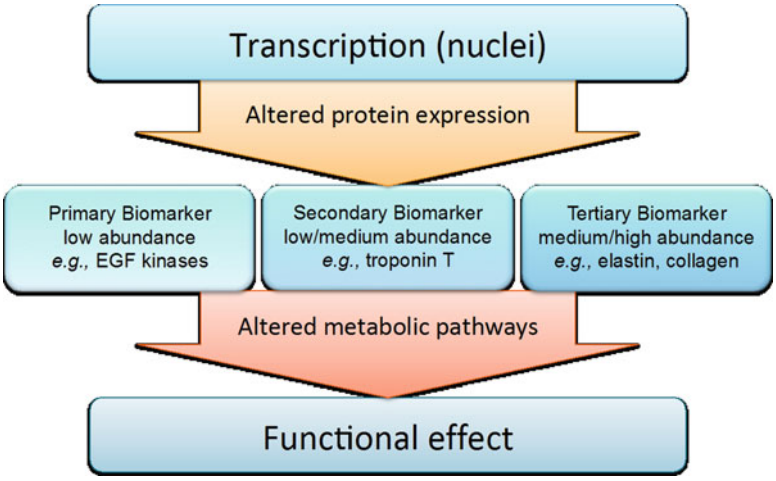


Fig. 7.1 Biomarker structure and expression abundance areas in clinical studies

Table 7.1 The biomarker utilization within clinical drug studies

Biomarkers for "Proof of Mechanism: POC"
A biomarker demonstrates an effect, which results in a functional change related to the proposed mechanism of action. The proof of mechanism effects can be measured for instance with an in vivo assay, following an appropriate stimuli.
Biomarkers for "Proof of Principle: POP"
A biomarker demonstrates an effect that results in a biological change, closely related to the proposed mechanism of action and known to be associated with disease activity in patients. The proof of principle biomarker readouts is proven in a dedicated patient study. It can be a measure of, e.g., an acute phase marker regulation in patient studies after drug intervention.
Biomarkers for "Proof of Concept: POC"
A study demonstrates an effect on a clinical end point. Proof of concept biomarker evaluation needs to be carried out in patients with the disease in question. In cancer studies for instance, a tumor reduction would be a positive effect where the biomarker quantitation provides a positive effect that has been achieved.

In addition to the above, within the drug development process, there is a successive use of biomarkers such as:

- Biomarkers in translational medicine
- Biomarkers for clinical pharmacology
- Safety and toxicity biomarkers
- Monitoring of response to therapy
- Patient stratification biomarkers
- Molecular diagnostic biomarkers

In order to have a better understanding, bridging academic and industrial biomarker research progression, a collaborative effort has been made in-between Duke University and LabCorp. This effort is established to optimize the speed whereby the discovery of interesting biomarkers is taken to commercialization.

7.4 The MRM Technology Platform

In most situations, biomarkers with protein origin are identified as differentially expressed in clinical samples comparing for instance clinical status of disease and health. The standard hospital technique of today is ELISA, where most commonly a recombinant protein is being used as the reference. The recombinant protein is used for quantitative analysis of patient samples which by itself is a possible source of analysis errors to be made in the diagnosis of the disease. The reason is that most often, the recombinant protein is different from the endogenous biomarker.

By utilizing mass spectrometry platforms, a protein sequence-related readout is generated, which is indifferent to the ELISA method, as the ELISA will quantify, based on antigen and antigen-like proteins and resulting in a final readout, which is not necessarily biomarker specific.

The MRM/SRM mass spectrometry technique, on the other hand, identifies and quantifies specific peptides within the digested samples, which is a complex mixture. In addition, the MRM/SRM mass spectrometry technology is inherently easy to multiplex. Multiple biomarker analysis assays are straightforward to develop, where the fast scan rate of modern mass spectrometry platforms of today easily handles multiple protein sequence quantifications. In addition, MRM/SRM assays will offer high sensitivity and speed, which is a future requirement, and high-throughput screening of clinical samples for candidate biomarkers within the clinical study area.

In the SRM or MRM methods, a given series of transitions are made with target peptides that are the precursors being ionized after LC separation and interface to MS. These precursor ions are then next fragmented ion pairs, where the corresponding data are picked up by the instrument. A multitude of peptides are quantified during a single LC-MS experiment. Looking back to decades of biological mass spectrometry and the recent years with proteomics studies, it is obvious that all the generated protein sequence data, compiled and built within databases, is a great access.

These experimental reports provide in many instances not only sequences of importance but also in plentiful cases quantitative information with respect to specific biology of clinical relevance. These clinical assay data have been generated both within Academia, where they are made public, as well as within pharma and biotech industry. The value of all this information, also highlighted by HUPO in the standardization program, Proteomics Standards Initiative (http://www.HUPO.org/research/psi/), is a strong fundament for MRM/SRM assay developments. By the utilization of stable isotope standards, MRM platforms are becoming a complementary protein assay technology to ELISA and other antibody-based assay techniques. A multifold of research groups in laboratories globally is picking up on the developments and utility of MRM platforms and assays. The relative ease of multifold reagent (stable isotope peptide) procurement is an attractive feature for MRM assays when compared with ELISAs. Most of the MRM assays developed for absolute quantitations make use of a digestion step followed by spiking of isotope-labeled heavy peptides (Fig. 7.2).

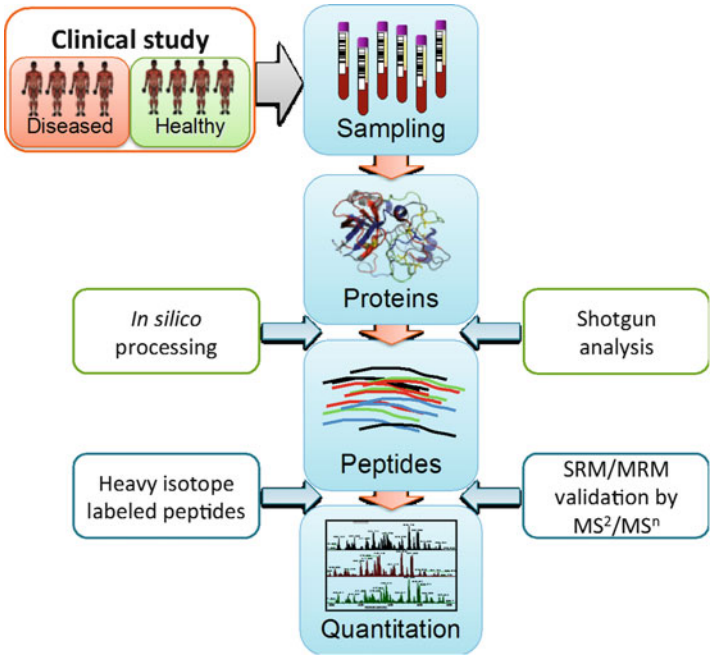


Fig. 7.2 Illustration of the MRM work flow that includes disease and healthy patient cohorts, sample preparation, peptide separation, and protein sequencing and quantification

MRM assay developments have been proven to be a highly useful technology principle for clinical protein and peptide analysis. In recent years, there is constant increase in reports on the assay developments that can be made by modern MRM platforms. The mass spectrometry instrumentation platforms are rapidly developing, becoming faster, more sensitive, and more user friendly. These technology platforms offer robust and reliable quantifications of proteins and peptides with good accuracy and precision. At the moment, MRM applications are the fastest growing targeted protein analysis area, with multiplex assays for absolute quantifications in clinical disease areas. The cardiovascular area holds, besides cancer, some of the fastest growing disease area and growth.

The MRM technology is able to quantify proteins down to femtomole levels and higher concentration regions, which means high and medium abundant and upper low abundant area. These are the applicable expression regions for proteins of interest. Importantly, the precision, as well as stability of these platforms, needs to be met for practical use.

By the use of isotope labeling technology, uniformly ^{13}C-^{15}N-labeled blood plasma levels of 100 ng/mL biomarkers can be quantified. A first report came out very recently where a magnetic bead-based immunoaffinity sample preparation method could reach relevant medium abundant concentration levels at ng/mL sensitivities.

By automation of the peptide antibody-capturing process step, high-throughput capacities were reached with low statistical variations (median CV 12.6%) for quantifying biomarkers using only 10 μL of plasma. Interestingly, by increasing the sample volumes to 1 mL, an improvement of the detection to the low pg/mL range was reached.

7.5 Applications to Prostate Cancer

As an example to cancer studies, we summarize our results on prostate cancer (PCa), where prostate-specific antigen (PSA) is the main biomarker for diagnosis. The worldwide prevalence of PCa is increasing leading to the diagnosis of every sixth men. However, the clinical course of the disease broadly dispersed, resulting in a larger group of patients who eventually die of other causes. The quantitative measure of PSA in blood is a cornerstone both for diagnosing and monitoring the disease (Lilja et al. 2007). There are difficulties associated with clinical applicability of PSA since its values can be elevated due to malignant as well as benign prostate disease, e.g., hyperplasia or prostatitis. As a result, 65–75% of men with a moderate PSA elevation (≈3–10 ng/mL, ref. value <3 ng/mL) do not have evidence of cancer (Schröder et al. 2009), whereas every fourth PCa patient has normal PSA levels (Rittenhouse et al. 1998). Additionally, the contribution of the lack of precise understanding of which molecular form(s) of PSA is measured by different commercially available clinical routine assays could be significant. Therefore, the microheterogeneity of PSA molecular forms should be closely investigated as it may reflect its diversity in the biology of prostate disease, providing important diagnostic values.

Mass spectrometry with high-resolving nanoseparation is a technique that we have developed specific methods and assays around (Végvári and Marko-Varga 2010). We have combined various analytical principles in order to improve the resolving power of PSA identification, utilizing 1-D gel electrophoresis, high-resolution MALDI-MS, and MALDI-MS/MS, which could confirm complex patterns of PSA forms in seminal plasma (Végvári et al. 2010). The PSA expression profiles in clinical samples were thus determined by MALDI-MS generating accurate masses and peptide sequences (using a high-end MALDI LTQ Orbitrap XL mass spectrometer), as well as monitored by Western blot analysis as shown in Fig. 7.3. Furthermore, we have used high-resolving MS (FT-ICR), resulting in unambiguous protein annotations in zymogram gel bands, which has enabled us to identify enzymatically active PSA isoforms even in the presence of a high chemical background.

As a latest addition to the MS platforms, we utilized for clear-cut identification of PSA, and we developed an MRM assay that included several molecular forms of PSA found in the UniProt/KB database, including both reviewed and nonreviewed sequence variants. All listed sequence variations were used for further processing of *in silico* digestion choosing trypsin as protease. The digestion was performed by

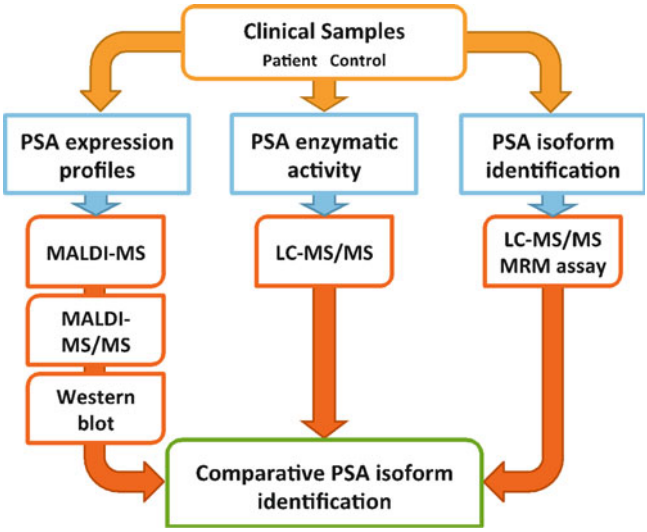


Fig. 7.3 Illustration of the mass spectrometric-based proteomics strategy. This approach was built extensively in order to improve identification of molecular forms of prostate-specific antigen in clinical samples, combining various mass spectrometric techniques

using the following settings: iodoacetamide as alkylation agent without oxidation on methionine and no miss-cleavage. The resulted tryptic peptides of all PSA isoforms were investigated for uniqueness by blast search. The isoform specificity of the proteotypic peptides was also noticed at this step. Finally, a list of tryptic PSA peptides was prepared filtering by size (not longer than 26 amino acids) for synthesis with and without heavy isotope labeling and alkylation at cysteine residues. Nine tryptic peptides were repeatedly identified in biological sample covering the entire range of useful sequences in MRM assays, and the top candidate for quantitative analysis was determined (LSEPAELTDAVK), as shown in Fig. 7.4. Further peptides could be selected (SVILLGR and FMLCAGR) suitable to measure PSA levels in clinical samples that can provide improved confidence and a useful tool control and filter possible interferences in the matrix background. Ultimately, these assays will be run in parallel to the standard measurements performed by ELISA used in clinical practice today and benchmarked.

Today, there is a high unmet need in the medical area for methods, instrumentation, and diagnostic capabilities that can fulfill the demand for improvements of the clinical health care, including early indicators of disease, disease severity, the evolvement phases of disease, and therapeutic efficacy. By the nine PSA forms we identified until today (Végvári et al. 2012), it is clear in our experience that the details of any given target, such as PSA in our case, the bioinformatics data at hand, and the "*in silico*" predictions that are experimentally verified, are powerful combinations. It allows us to reach statistical power with significance scoring in clinical situations that previously have been unknown.

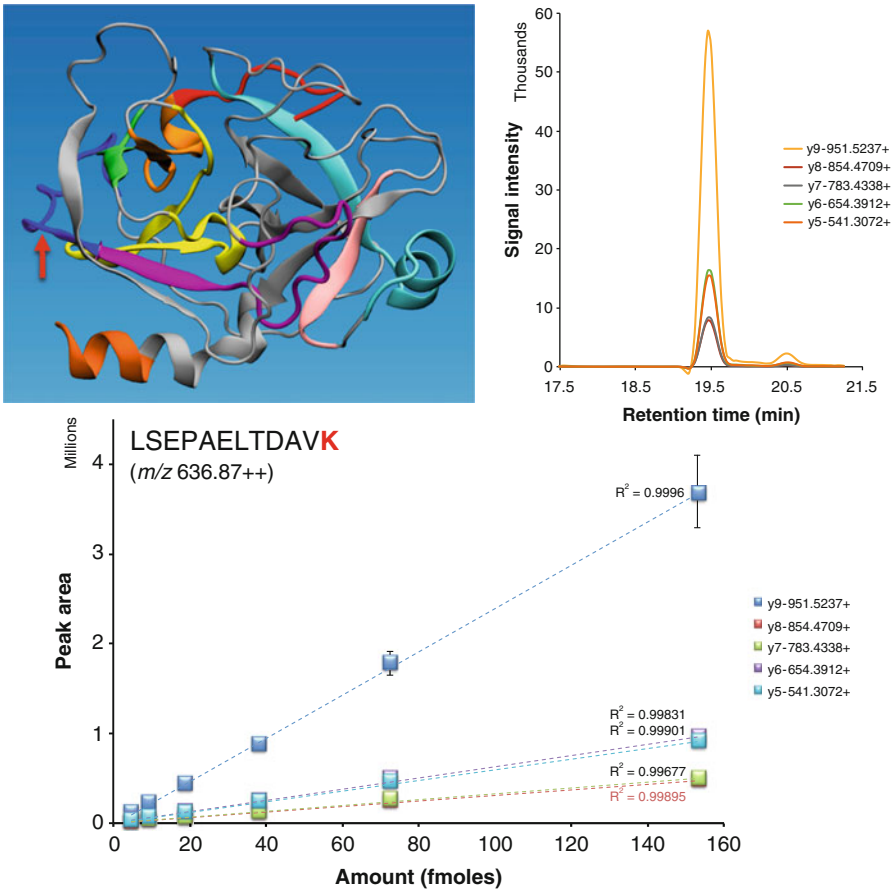


Fig. 7.4 Selection of the most suitable signature peptide of prostate-specific antigen. Among the accessible tryptic peptides of PSA (shown *color coded in the upper left panel*), LSEPEALTDAVK was identified as top candidate for quantification of PSA due to its high signal intensity and excellent linearity (presented in the *upper right and lower panels*, respectively)

7.6 Applications to Cardiovascular Diseases

The cardiovascular disease area is extensively investigated both in the discovery of new biomarkers and the verification of these markers using quantitative assays. MRM assays were published recently for the quantitation of a numerous cardiovascular biomarkers with LOQs below 1,000 ng/mL (Domanski et al. 2012) and with the application of multidimensional separations in the range of 1–20 ng/mL (Keshishian et al. 2009). With the application of the stable isotope standards with capture by antipeptide antibodies (SISCAPA) technique, the limit of quantitation

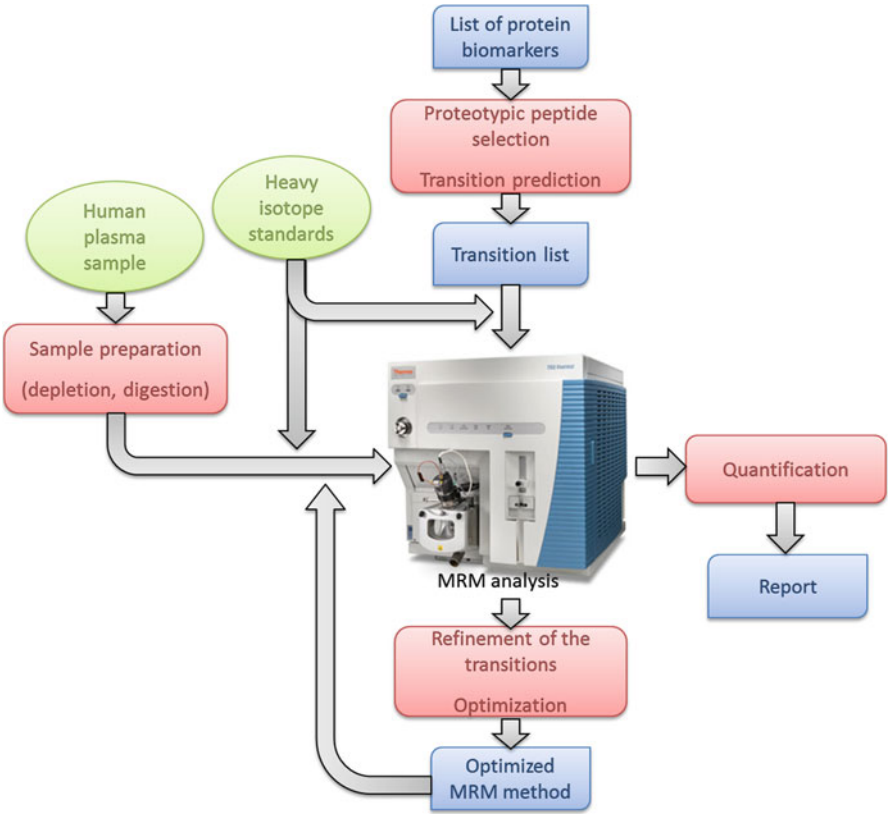


Fig. 7.5 Work flow of the cardiovascular assay development

can be improved dramatically, as it was demonstrated by Carr's group using this approach for the quantification of troponin I and interleukin 33 (Kuhn et al. 2009).

In our assay development (represented in Fig. 7.5), we have focused on 11 cardiovascular disease biomarkers in the high and medium abundant plasma level, selected by clinicians. Unique tryptic peptide sequences were selected for each protein biomarker based on previous experimental data (nontargeted LC-MS/MS trials) and database search (Peptide Atlas, GPMDB) according to the common filtering criteria. The selection of sequences was then synthesized in light and heavy labeled forms (C-terminal [$^{13}C_6$, $^{15}N_4$]-Arg or Lys), and these peptides were used for selection and optimization of the transitions. The transitions were tested in real background matrix, depleted and nondepleted plasma digest, as well. Stable isotope-labeled standards spiked into the samples permit validation of the transitions and help in filtering out the inaccurate transitions and, in addition, allow absolute quantitation. The endogenous and the heavy standard peptide fragmented identically under the same conditions, generating the same pattern of the daughter

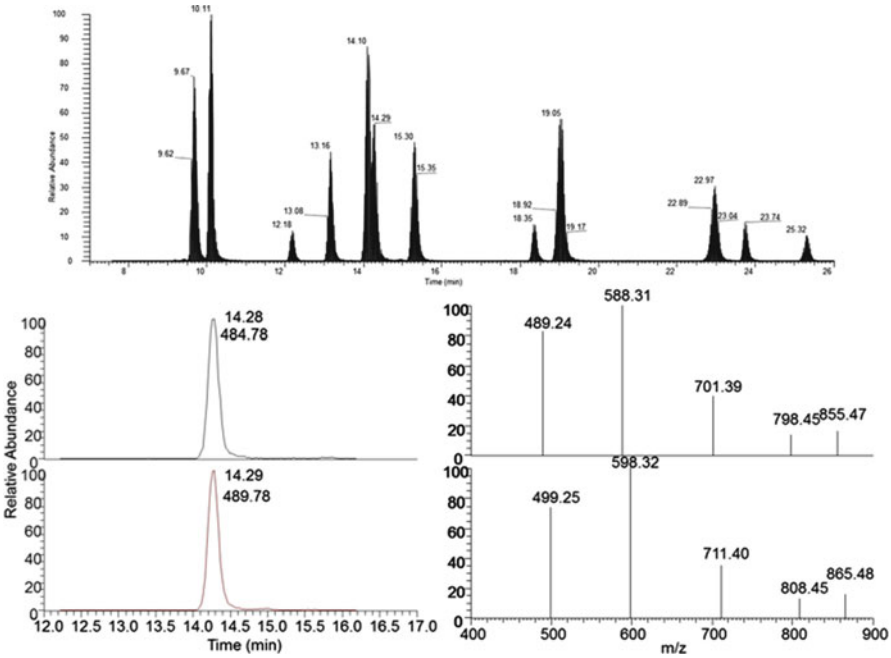


Fig. 7.6 MRM assay readout. TIC of a spiked plasma sample and a representative endogenous/heavy peptide pair with coeluting peaks and identical fragmentation patterns

ions, reflecting to the mass differences due to the labeled AS. The relative intensities of the product ions for a given precursor ion along with the peak shape and the retention time are used to the refinement of the transitions in order to generate a high-selectivity assay. In the final assay, at least three transitions, free-form matrix interferences, per precursor ion were monitored and used for quantitation. Plasma concentrations of these 11 biomarkers were measured in patients with myocardial infarct and control samples. Figure 7.6 shows the MRM assay readout in plasma samples.

7.7 Conclusion

In the near future, we expect to have reached a point where protein biomarkers and SRM-MRM technology have taken a larger percentage of the market with a targeted drug approach. Improved clinical chemistry diagnosis with gene- and protein sequence-based assays will also become state of the art in future medical health care, as it is inherently linked to the increasing use of personalized medicine. These changes will result in a much higher number of overall analysis throughput and patient data generation. Consequently, high-throughput multiplexed biomarker

assay platforms will play an important clinical role as becoming a complement to traditional immunoassays for future use in clinical health care and targeted medicine. The introduction of new biomarkers of tumors and cardiovascular diseases will assist in early identification of disease and in monitoring the effect of therapeutic agents on disease progression.

We also envision that key issues are related to an extensive role that biobanks will play in the development of new paradigms of disease pathogenesis and in the establishment of new treatment protocols for unmet needs in the clinic that will only be learned in time. New resources in stored samples, representing milestones of health and illness, deserve attention by the public and the political institutions that protect the public's interest. Lastly, whether such future solutions will be able to provide the remedy and become the Holy Grail of disease understanding still remains to be proven by all of us within our societies.

References

Aebersold R, Auffray C, Baney E, Barillot E, Brazma A, Brett C, et al. Report on EU-USA workshop: how systems biology can advance cancer research (27 October 2008). Mol Oncol. 2009;3(1):9–17.

Anderson NL. The roles of multiple proteomic platforms in a pipeline for new diagnostics. Mol Cell Proteomics. 2005;4(10):1441–4.

Anderson NL, Anderson NG. The human plasma proteome – history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1(11):845–67.

Andrejevs G, Celis JE, Guidi G, Peterle A, Sullivan R, Wilson R. Tackling cancer in the EU: the role of innovation. Mol Oncol. 2009;3(1):18–23.

Apweiler R, Aslanidis C, Deufel T, Gerstner A, Hansen J, Hochstrasser D, et al. Approaching clinical proteomics: current state and future fields of application in cellular proteomics. Cytometry A. 2009;75(10):816–32. Epub 2009/09/10.

Cho CR, Labow M, Reinhardt M, van Oostrum J, Peitsch MC. The application of systems biology to drug discovery. Curr Opin Chem Biol. 2006;10(4):294–302. Epub 2006/07/11.

Domanski D, Percy AJ, Yang J, Chambers AG, Hill JS, Freue GVC, et al. MRM-based multiplexed quantitation of 67 putative cardiovascular disease biomarkers in human plasma. Proteomics. 2012;12(8):1222–43.

Domon B, Aebersold R. Review – mass spectrometry and protein analysis. Science. 2006;312(5771):212–17.

Kato H, Nishimura T, Hirano T, Nomura M, Tojo H, Fujii K, et al. A clinician view and experience of proteomic studies in the light of lung cancer in Japanese healthcare. J Proteome Res. 2011a;10(1):51–7.

Kato H, Nishimura T, Ikeda N, Yamada T, Kondo T, Saijo N, et al. Developments for a growing Japanese patient population: facilitating new technologies for future health care. J Proteomics. 2011b;74(6):759–64.

Keshishian H, Addona T, Burgess M, Mani DR, Shi X, Kuhn E, et al. Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics. 2009;8(10):2339–49.

Kuhn E, Addona T, Keshishian H, Burgess M, Mani DR, Lee RT, et al. Developing multiplexed assays for troponin I and interleukin-33 in plasma by peptide immunoaffinity enrichment and targeted mass spectrometry. Clin Chem. 2009;55(6):1108–17.

Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, et al. The human proteome project: current state and future direction. Mol Cell Proteomics. 2011;10(7):M111.009993.

Lilja H, Ulmert D, Bjork T, Becker C, Serio AM, Nilsson JA, et al. Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. J Clin Oncol. 2007;25(4):431–6.

Mann M. Comparative analysis to guide quality improvements in proteomics. Nat Methods. 2009;6(10):717–19.

Marko-Varga G, Fehniger TE. Proteomics and disease – the challenges for technology and discovery. J Proteome Res. 2004;3(2):167–78.

Marko-Varga GA, Végvári Á, Fehniger TE. A protein shake-up. Public Serv Rev Eur Union. 2011;21:250–2.

Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, et al. A dual pressure linear ion trap orbitrap instrument with very high sequencing speed. Mol Cell Proteomics. 2009;8(12):2759–69.

Paavilainen L, Edvinsson A, Asplund A, Hober S, Kampf C, Ponten F, et al. The impact of tissue fixatives on morphology and antibody-based protein profiling in tissues and cells. J Histochem Cytochem. 2010;58(3):237–46. Epub 2009/11/11.

Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, et al. The chromosome-centric human proteome project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012a;30(3):221–3.

Paik YK, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, et al. Standard guidelines for the chromosome-centric human proteome project. J Proteome Res. 2012b;11(4):2005–13.

Rezeli M, Végvári Á, Fehniger TE, Laurell T, Marko-Varga G. Moving towards high density clinical signature studies with a human proteome catalogue developing multiplexing mass spectrometry assay panels. J Clin Bioinformatics. 2011;1(1):7.

Rinn JL, Rozowsky JS, Laurenzi IJ, Petersen PH, Zou K, Zhong W, et al. Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. Dev Cell. 2004;6(6):791–800.

Rittenhouse HG, Finlay JA, Mikolajczyk SD, Partin AW. Human Kallikrein 2 (hK2) and prostate-specific antigen (PSA): two closely related, but distinct, kallikreins in the prostate. Crit Rev Clin Lab Sci. 1998;35(4):275–368.

Schmidt A, Claassen M, Aebersold R. Directed mass spectrometry: towards hypothesis-driven proteomics. Curr Opin Chem Biol. 2009;13(5–6):510–17.

Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360(13):1320–8.

Schwenk JM, Gry M, Rimini R, Uhlen M, Nilsson P. Antibody suspension bead arrays within serum proteomics. J Proteome Res. 2008;7(8):3168–79.

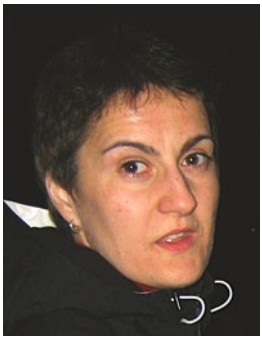
Végvári Á, Marko-Varga G. Clinical protein science and bioanalytical mass spectrometry with an emphasis on lung cancer. Chem Rev. 2010;110(5):3278–98.

Végvári Á, Rezeli M, Welinder C, Malm J, Lilja H, Marko-Varga G, et al. Identification of prostate-specific antigen (PSA) isoforms in complex biological samples utilizing complementary platforms. J Proteomics. 2010;73(6):1137–47.

Végvári Á, Rezeli M, Sihlbom C, Häkkinen J, Carlsohn E, Malm J, et al. Molecular microheterogeneity of prostate specific antigen in seminal fluid by mass spectrometry. Clin Biochem. 2012;45(4–5):331–8.

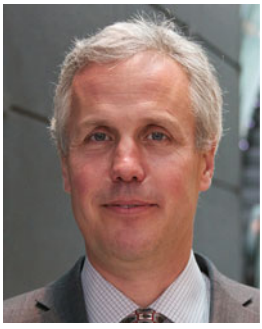
Weissenstein U, Schneider MJ, Pawlak M, Cicenas J, Eppenberger-Castori S, Oroszlan P, et al. Protein chip based miniaturized assay for the simultaneous quantitative monitoring of cancer biomarkers in tissue extracts. Proteomics. 2006;6(5):1427–36.

Ákos Végvári, Ph.D., Sweden Dr. Végvári is a docent and senior research scientist at Lund Institute of Technology. He is working on analytical methodology development for separation of peptides, proteins, and nucleic acids. He joined a biotechnology company in Stockholm, working on the analytical evaluation of a newly developed AIDS drug. He is currently working at the Biomedical Center of Lund University. His main research interests focus on disease linked, mass spectrometry-based proteome analysis, including targeted cancer proteomics as well as fundamental method development for localization of drug compounds in tissue sections by MALDI imaging mass spectrometry. Dr. Végvári has published more than 45 scientific papers within internationally well-recognized journals and has written three book chapters. He has an h-index of 13 and a total of more than 500 independent citations.



Melinda Rezeli, Ph.D., Sweden Dr. Rezeli got her Ph.D. degree in chemistry at the University of Pécs in 2008, with the thesis on the development of acrylamide-based separation matrices. During her postgraduate studies, she spent 2 years in the prestigious Stellan Hertén's laboratory at Uppsala University.

In 2007, she joined a biotech company in Pécs, Hungary, as a researcher and took part in the development of immune-based multiplex assays in the field of life sciences and health care. Currently, she is working at the Biomedical Center of Lund University as a research scientist. Her main research interest focuses on mass spectrometry-based protein biomarker quantitation, including multiplex assay development and biomarker discovery. Dr. Rezeli has published 19 scientific papers in international well-recognized journals.



David Erlinge, Ph.D., Professor, Sweden David Erlinge is a professor at the Department of Cardiology at Lund University. Prof. Erlinge was the imitator and started LUNDHEARTGENE, which is a genetic and proteomic biobank coupled to the SWEDEHEART register in Sweden. The LUNDHEARTGENE has over the years analyzed microRNA expression in plasma patients with acute myocardial infarction and found release of cardiac-specific miRNA in patients with STEMI that has prognostic information for the development of heart failure. The *identified clinical problems, related to* the outcome after acute myocardial infarction, are major research areas of Prof. Erlinge.

György Marko-Varga, Ph.D., Professor, Sweden and Japan Marko-Varga is head of Clinical Protein Science/Imaging Division, Lund University, Sweden, senior scientist at University Hospital of Lund, and professor at the Tokyo Medical University Hospital. Marko-Varga has been working within senior drug discovery/development positions and responsibilities within Astra and AstraZeneca for a period of 17 years.

His current research is focused on the development of novel diagnostic assays and platforms, and interfacing high-resolution separation with mass spectrometry, to build an understanding of the mode of drug action and disease mechanisms for lung cancer and COPD. To that end, he is building new research groups involving collaborations with the pharmaceutical industry, academia, and clinical hospitals. For example, he participated in protein biomarker discovery research involving 52 lung cancer centers in Japan, in the biggest biomarker study ever in the industry. One of his major new challenges is responsibility for the biobanking of the "big three" study, lung cancer, cardiovascular disease, and COPD, which involve 100,000 patients and six million samples in Sweden. Additionally, he is the president of European Proteomics Association, president of the Swedish Proteomics Society, and the coordinator of Chromosome 19, within the Human Proteome Project initiative. Prof. Marko-Varga has filed for 22 patents worldwide and has more than 250 publications with h-index: 34. He has edited 3 books and written 19 book chapters. He is also the European editor of Journal of Proteome Research, an American Chemical Society journal, and member of the editorial board of additionally 9 international journals. Prof. Marko-Varga has since 2005 developed 3 start-up companies.

Chapter 8
Disease Biomarkers: Modelling MR Spectroscopy and Clinical Applications

Luis Martí-Bonmatí and A. Alberich-Bayarri

Abstract Clinical MRS has become a reference technique for in vivo evaluating the metabolism of different tissues, with special application to brain and prostate lesion characterization and tumour's follow-up. It allows detecting relevant changes that cannot be appreciated in the conventional MR images. Nowadays, MRS has been widely applied in many different brain pathologies with excellent results as a disease biomarker. Since the different diseases and grades have different manifestations in the spectroscopic profile, a deep understanding of the subjacent biology is needed for the signal interpretation. The development of high-field (≥3 T) scanners has permitted the acquisition of high-quality MRS also in other organs and regions outside the brain. The most non-brain extended application of spectroscopy nowadays is in the detection and monitoring of prostate cancer. The technique is increasingly being applied also for the detection and diagnosis of breast lesions, the quantification of the hepatic fat fraction in steatosis and the characterization of muscle metabolism. The evolution of the acquisition technology with spectroscopic imaging, higher spatial resolution, lower acquisition times and the automation of the spectra processing analysis will encourage the wider application of this technique in many degenerative and oncologic diseases.

Keywords Magnetic resonance • Spectroscopy • Disease biomarkers

L. Martí-Bonmatí, Ph.D., M.D. (✉)
Department of Radiology, Hospital Quirón Valencia,
Blasco Ibáñez 14, 46010 Valencia, Spain

Radiology, Department of Medicine, University of Valencia, Valencia, Spain
e-mail: luis.marti@uv.es

A. Alberich-Bayarri, Ph.D.
Quantification Group, Department of Radiology, Hospital Quirón Valencia,
Blasco Ibáñez 14, 46010 Valencia, Spain
e-mail: luis.marti@uv.es

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_8,

8.1 Introduction

Magnetic resonance spectroscopy (MRS) was initially developed in many research teams after the discovery of the nuclear magnetic resonance phenomenon in 1946 by Felix Bloch and Edward Purcell. Many researchers started to study the differences in the resonant frequency related to the type of nucleus (mainly hydrogen, phosphorus, carbon and sodium). However, they noticed that even the same type of atoms had a different resonant frequency depending on the molecular structure in which atoms were integrated, due to the chemical shift effect, discovered by Proctor and Yu in 1950, and to the J-coupling effect. Chemistry scientists quickly realized the great potential of this new analytical technique and incorporated MRS as a routine method to explore the structural composition of chemical compounds and to monitor the changes that occur during the synthesis processes.

The development of more powerful magnets and more uniform magnetic fields allowed the study of compounds with higher molecular weight. MRS rapidly expanded in the field of biochemistry for the structural analysis of macromolecules and molecular interactions. However, the boom of magnetic resonance imaging (MRI) in medicine in the 1980s relegated MRS to a second term, since the in vivo acquisition was not trivial. Only with the development and implementation of clinical MR systems with higher field strengths, the acquisition of spectra with sufficient resolution and sensitivity became feasible.

The key factors for the clinical use of MRS in patients are high field strength and linear, stable and selective gradients to accurately locate and position the MR signals from a given volume of tissue inside the human body. As an early example, the magnetic field gradients allowed the acquisition of spectroscopic signals from a voxel located in the centre of the brain to characterize tumoral cells. In clinical MRS, the signal localization techniques are similar to those used in imaging although the spectroscopic signals are acquired from larger voxel volumes when compared with the higher spatial resolutions of MRI.

While the main voxel signal observed in MRI is mainly generated from water and also fat molecules, the MRS voxel signals are derived from molecules and metabolites that are from 1,000 to 100,000 times less abundant than water molecules in the biological tissues. Therefore, the water resonance peak may obscure other metabolites, being one of the main challenges in the quantification of MR spectra. Different methods to suppress the water signal and properly tuning the MR equipment are a key element in the spectroscopic MRS analysis.

The aim of this chapter is to provide a comprehensive vision of in vivo MRS, by means of understanding the main sequences and configurations used in the spectra acquisition, processing methods applied to the signal, quantification of different metabolite concentration and finally analysing the typical spectrum alterations in different pathological abnormalities.

8.2 Physical Principles and Concepts of MR Relaxation

The MR phenomenon is mainly governed by excitation-relaxation processes, in which the protons of the sample are excited at their resonant frequency by a radio-frequency (RF) pulse. Immediately after the excitation, protons progressively lose and emit the previously received energy in the so-called relaxation process.

During the proton excitation phase by RF pulses, the position of the main magnetization component, which in the equilibrium state is in the longitudinal (z) axis, is tilted to the transverse (x–y) plane. After the RF transmission, the relaxation process starts, allowing the capture of the generated MR signal. The received signal displays the decay of the magnetization in the transverse plane, and it is known as the free induction decay (FID) signal.

The most extended representation of an FID signal can be appreciated in Fig. 8.1a. However, in real conditions, the FID signal is a mixture of a high number of resonant frequencies (an example of a real FID signal can be appreciated in Fig. 8.1b).

In order to properly analyse the FID signal and its frequency components, a Fourier transform is applied to convert from the temporal to the frequency domain (Fig. 8.2).

There are two main mechanisms that contribute to the different resonance frequencies that the same atoms have at different parts of a molecule: the electronic molecular shielding and the J-coupling phenomena. The electronic shielding is mainly the responsible of the chemical shift frequency differences between different metabolites and even different parts of the molecules. In the J-coupling, the magnetic field of a nucleus changes the external magnetic field that acts on a neighbouring nucleus. This effect is conducted by the electrons of the existing chemical bond between the pair of nuclei.

8.2.1 Chemical Shift

According to the Larmor equation, the precession frequencies of the different nuclei depend on the perceived main magnetic field. The main magnetic field that is

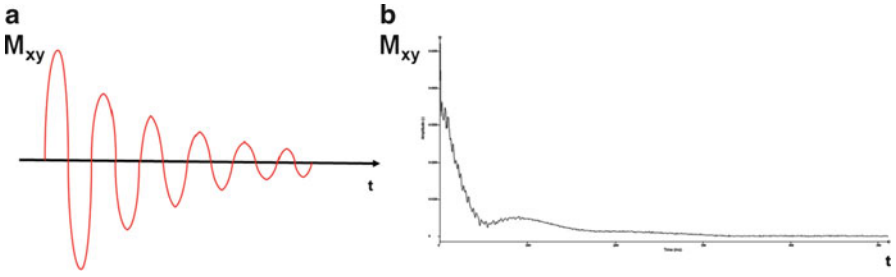


Fig. 8.1 Free induction decay (*FID*) signal obtained from the relaxation process of magnetization. In (**a**), a theoretical depiction of the decay of the FID signal. In (**b**), real FID signal acquired in an MRS sequence

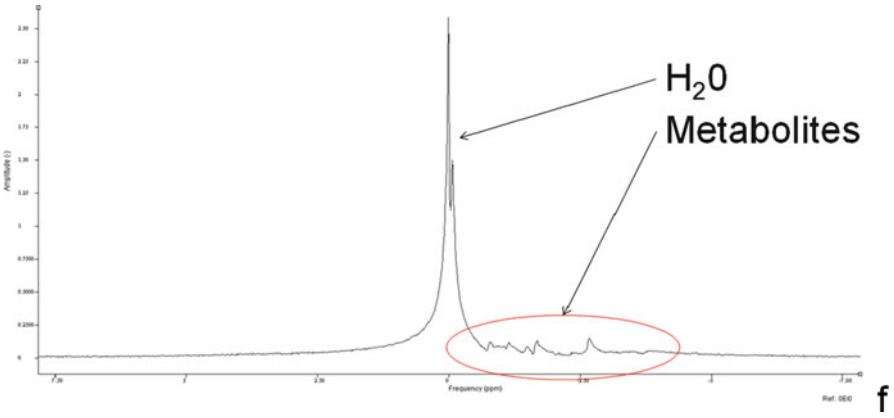


Fig. 8.2 Raw spectrum acquired in a 3 T MR scanner. Spectrum of the signal calculated after the MR acquisition with water suppression. The water concentration is significantly higher than the metabolites of interest, and an additional filtering of the water signal will be needed

present in MRS during an acquisition is formed by three different components: the main magnetic field (B_0), a lower magnetic field produced by the gradients system (G) and a really small magnetic field associated to the microenvironment in the molecule (B_{molecule}). Since B_0 and G are equal for all the protons of a spectroscopy acquisition from a given volume of tissue, the frequency differences of hydrogen atoms will be given by the small magnetic field variations that depend on the molecular environment. These differences are very small when compared with the basic resonance frequency but sufficient to identify individual molecules. These differences, usually expressed in (Hz) or parts per million (ppm), define the chemical shift concept.

These variations in frequency between nuclei of the same elements, such as hydrogen, within a molecule are due to the electronic shielding present in certain parts of the molecule. Thus, larger molecules with a higher number of bonds and atoms will have denser electron 'clouds'. Such clouds act shielding the external magnetic field and slightly decreasing the overall magnetic field feel by the atomic protons and therefore changing their resonance frequency.

The chemical shift range in proton clinical MRS is very small, approximately between 5 and 10 ppm, meaning that the frequency differences between the metabolites are in the order of a 100 Hz in a signal of 128 MHz, which is the frequency precession for a 3 T MR system. As an example, the water and fat resonance peaks are separated approximately 225 Hz in a 1.5 T system and 445 Hz in a 3 T system. To avoid confusion in relationship with B_0 of the chemical shift and also to unify the criteria of MRS studies, the chemical shift values are usually expressed in ppm units, which are normalized to a reference frequency by the following expression:

$$\delta_a(\text{ppm}) = \frac{f_a - f_r}{f_r} \times 10^6$$

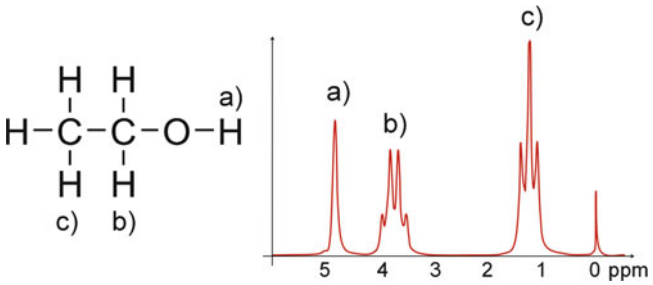


Fig. 8.3 Multiplets in ethanol spectrum. The different parts of the molecule resonate at different frequencies and divide in multiplets according to the $2{\cdot}N{\cdot}I+1$ rule from quantum physics for J-coupling effect

where f_a is the absolute frequency and f_r is the reference frequency. The most common compound used as a reference in organic MRS is tetramethylsilane. As an example, after normalization, the difference between the water and fat peaks is 3.4 ppm with complete independence of the main magnetic field where measurements are done.

8.2.2 Multiplets

A single nucleus within a molecule can also present different resonance peaks as a result of coupling with other nuclei. This coupling (also called indirect spin-spin coupling or J-coupling) between a pair of nuclei only takes place if they are linked by a chemical bond. The bonding electrons moving between the two nuclei have an influence in the external magnetic field and produce subtle differences in the precession frequency of each nucleus. In this way, the J-coupling effect influences the direct relationship between the observed spectrum and the distribution of the atoms in the molecule (Salibi and Brown 1997). In fact, from quantum physics, it is known that the number of resonant peaks for each element is equal to $2{\cdot}N{\cdot}I+1$, being N the equivalent number of protons of the neighbour group and I the spin number of the nucleus being excited, which in the case of hydrogen is 1/2. An example of this relationship can be observed in the ethanol molecule and the corresponding spectra in Fig. 8.3.

8.3 Clinical MRS Acquisition

8.3.1 Preparation Phase

The first technical consideration in a clinical MRS examination is the proper planning of the voxel geometry in order to cover the region of interest. It is recommended to use always the last imaging sequences acquired in the patient to avoid positioning errors due to movement misregistration.

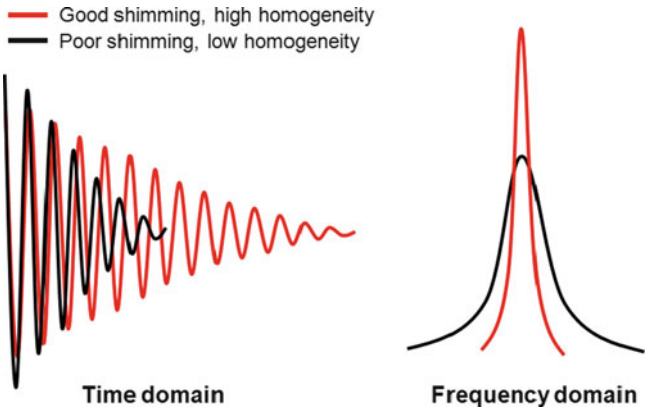


Fig. 8.4 Shimming. Characteristics of the decay of the transverse magnetization in the relaxation process during the shimming preparation. A short FID decay (*wide water peak*) represents a high magnetic susceptibility and signal heterogeneities. A long FID decay (*narrow water peak*) represents homogeneous magnetization and efficient shimming

Although the different MR manufacturers, in order to minimize human interaction and reduce acquisition time, are progressively automating the initial preparation steps in an MR acquisition sequence, it is important to understand their relevance for the quality of the final spectra. The preparation phase can be divided in three main different steps: the centre frequency tuning, the volume shimming and the water suppression.

In the selection of the centre frequency, the water resonance is placed at the centre of the band. If the volume of interest is located in a region subject to magnetic susceptibility, such as air, the tuning may be incorrect and should be adjusted manually (Gili 2009).

All the MR systems are equipped with a set of coils that generate magnetic field variations used to adjust the field homogeneity ('shimming'). The current through these coils is varied to offset the main magnetic field homogeneity variations. Although all current tools are optimized to obtain quality images for diagnostic purposes in MRI, the degree of homogeneity is less than the necessary to obtain a spectrum at MRS. Magnetic field uniformity adjustments may be global or regional, that is, in the sensitive region of the coil or located just in the volume of interest. For an optimum shimming in MRS acquisitions, it is recommended to avoid tissues and components with high susceptibility variation (such as bone, air or iron) within the volume of interest or next to its boundaries. If the tissue under study contains MR susceptibility changes, the relaxation of the magnetization will occur much faster, and the spectrum of the water signal will be wider. On the contrary, a good shimming will make the magnetization relax for a longer time with a narrower water signal peak (Fig. 8.4). In the shimming process, the system will try to minimize the full width at half maximum (FWHM) of the water bandwidth. The different manufacturers usually give these FWHM values during the preparation of the spectroscopy sequences. A good shimming is considered when the FWHM values are <10 Hz.

The last step in acquisition preparation is the water suppression, a need in proton spectroscopy since metabolites presence is 1,000–100,000 times lower than water. There are different techniques to suppress water signal. As water suppression effectiveness depends mainly in magnetic field homogeneity, it must be optimized for every patient, as the tissues, organs and anatomy differ between individuals. The most extended method consists on the selective excitation of the water resonance just before spectrum acquisition (Gili 2009). A narrow bandwidth guarantees that these pulses only affect the water signal and preserve for quantification all the different resonances of other metabolites.

In terms of frequency domain, as the water selective excitation pulses need to be centred in the water component, the frequency tuning is performed at the beginning of the preparation phase.

The inversion technique can also be applied for water suppression; it consists in the application of a selective inversion pulse in water, while the spectrum acquisition starts when the water magnetization signal is passing through the zero line during the recovery process. The use of adiabatic pulses for optimization of water suppression in conditions of inadequate homogeneity can also be employed but is not the object of this chapter.

8.3.2 Acquisition

As in MRI, many different approaches can be considered for the acquisition of MRS examinations. The different acquisition methods can be grouped according to the sequence type and excitation volume or depending on the echo time.

8.3.2.1 Sequence Type and Acquisition Volume

There are two main different sequence configurations for the acquisition of single-voxel spectrum, the point resolved spectroscopy (PRESS) and the stimulated echo acquisition mode (STEAM) (see Fig. 8.5 for sequence diagrams). The PRESS sequence is based on spin echo and provides a higher signal-to-noise ratio than the STEAM. On the contrary, STEAM is stimulated echo sequences that allow shorter echo times compared to PRESS sequences, permitting the acquisition of spectra with a higher amount of information.

Regarding the acquisition of multi-voxel spectrum, also called chemical shift imaging (CSI) or spectroscopic imaging (MRSI), some considerations must be taken in acquisition. The main problems arise from fat contamination. As an example, in brain CSI spectroscopy, the subcutaneous fat can contaminate the signal in nearby voxels. However, either PRESS or STEAM configurations can be combined with CSI in order to properly delimitate the excitation volume. Another approach to avoid fat contamination in the volume of interest is to use saturation slabs. Depending on the number of phase encoding steps, sequences can be 2D or 3D. Nowadays,

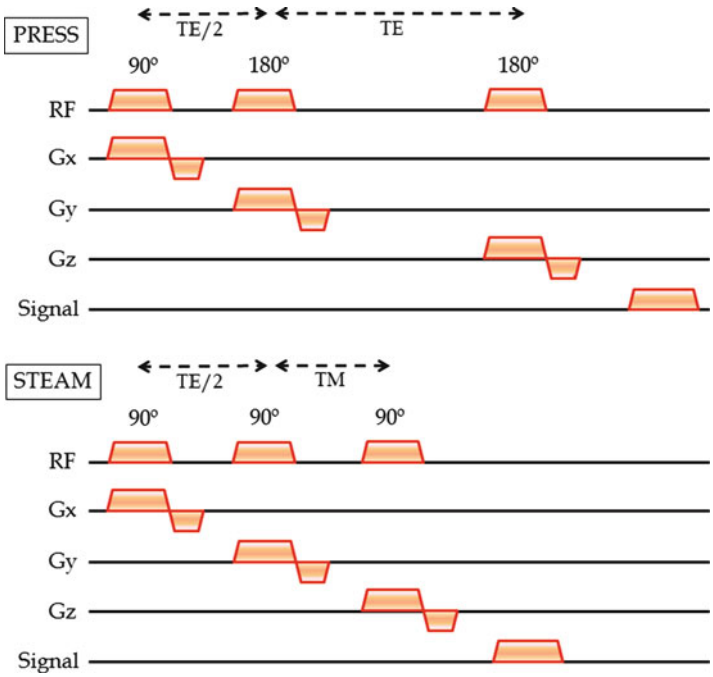


Fig. 8.5 Single-voxel MRS sequences. Sequence diagrams for PRESS and STEAM sequences. Note the shorter duration of *STEAM* if compared to *PRESS*

3D-CSI sequences can be routinely acquired in reasonable acquisition times, such as in the diagnosis of prostate cancer where a 3D multi-voxel MRS provides the metabolic profile of the prostate at different levels. An example of single-voxel and multi-voxel acquisition can be appreciated in Fig. 8.6.

8.3.2.2 Echo Time

The echo time (TE) is a crucial parameter for signal acquisition of the different metabolites. The adequate TE for each acquisition is a non-straightforward aspect in proton MRS, balancing the signal-to-noise versus the contribution of metabolites to the signal. If a short TE is used, metabolites with either short or long T2 relaxation times will be depicted. The typical value for short TE is around 30 ms in MRS. If a long TE is used, close to 270 ms, the relaxation of metabolites with short T2 will have ended and will not contribute to the signal. A useful long TE is 144 ms since lactate, which is an indicator of hypoxia, is inverted at this TE and can be differentiated from the lipid resonance. An example of a healthy brain spectrum acquired with both short (TE = 32 ms) and long (TE = 136 ms) TE can be appreciated in Fig. 8.7.

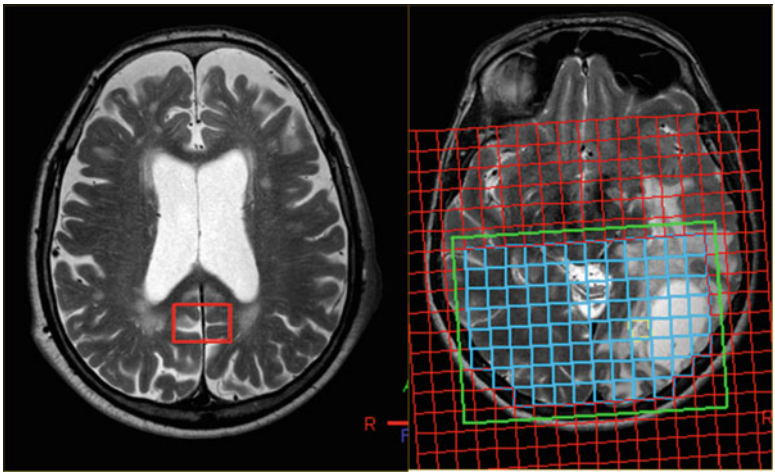


Fig. 8.6 Single voxel vs. multi-voxel. Single-voxel MRS acquisition positioning in the posterior cingulate in a patient with Alzheimer's disease (*left*). Multi-voxel acquisition in a patient with glioblastoma multiforme (*right*)

8.4 Signal Processing and Analysis

After the proper signal acquisition, the spectrum needs to be prepared and analysed for the extraction of quantitative information in order to answer a clinical dilemma or evaluation. There are different software applications for the processing and analysis of MRS data. The different manufacturers have developed quite automated applications that fasten the MR spectrum analysis tasks, although these applications are difficult to manipulate in case of nonoptimal fitting of the spectra. As a result, other third-party applications such as jMRUI (Naressi et al. 2001) and LCmodel (Provencher 2001) have been extensively used among the MRS research and advanced clinical applications community.

The first step is the elimination of the residual water signal as despite the use of water suppression techniques in the sequence preparation phase, some residual water signal exists and must be filtered. A selective filter is applied typically between 4.30 and 5.10 ppm (van den Boogaart et al. 1994). Afterwards, an apodization filtering is applied in order to reduce spectrum noise. Once the spectrum is de-noised and the metabolite resonances are identified, the phase of the signal must be adjusted in order to obtain good peaks depiction and a similar baseline level all over the spectral baseline. Finally, the offset level of the baseline is corrected.

The different steps can be summarized as:

- Water peak filtering
- Spectrum apodization filtering
- Phase correction
- Baseline correction

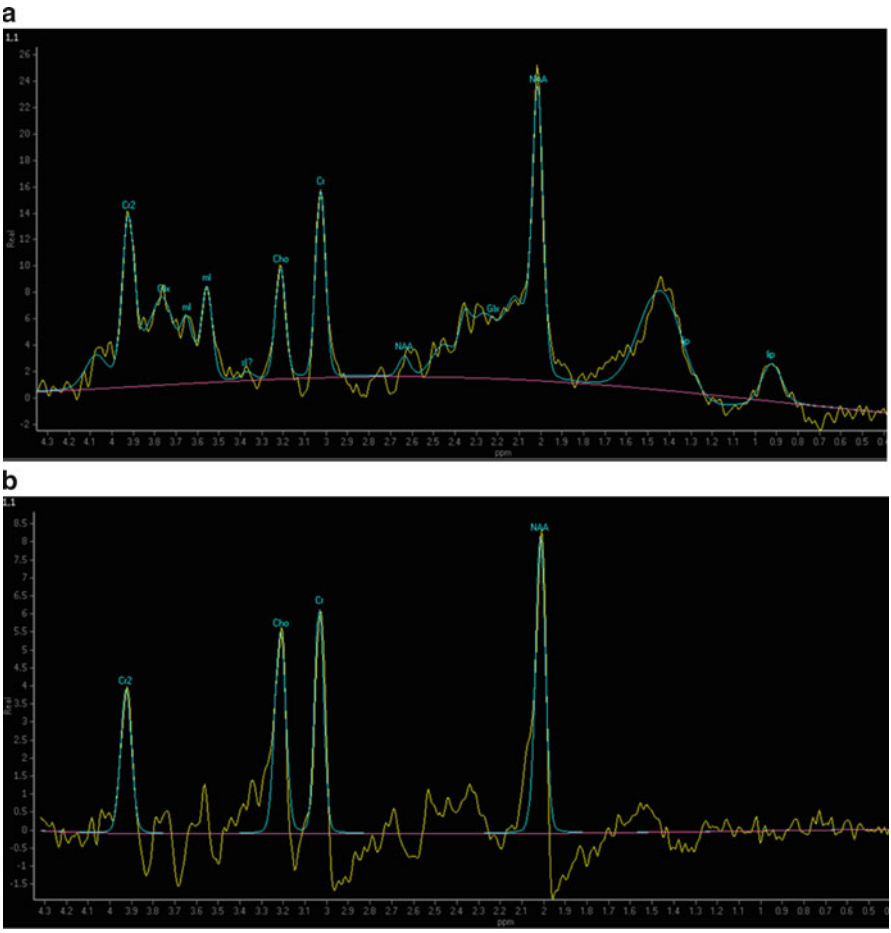


Fig. 8.7 Short TE vs. long TE. In (**a**), brain MRS acquisition in a healthy subject performed with a short TE. In (**b**), acquisition in the same region than (**a**) but using a long TE. Note the differences in the myoinositol resonances between short and long TE

After the processing phase, the relevant identified metabolite concentrations must be quantified. Although there exist both time- and frequency-based algorithms for spectrum quantification, the method can be intuitively explained better in the frequency domain. The area delimited by each peak and the baseline must be calculated since it is directly related to the metabolite concentration (Vanhamme et al. 1997).

The quantification can be performed either in absolute or relative values. For the absolute values quantification, either an internal or external reference is needed. The main drawback of this method is that it increases the acquisition time, since ideally the repetition time (TR) of the sequence needs to be at least of 5 s (approximately the relaxation time of water if water is used as a reference) in order to

Table 8.1 Main metabolites and frequency positions that are commonly analysed and studied as disease biomarkers in brain

Metabolite	Abbreviation	Frequency (ppm)	Function
N-Acetyl-aspartate	NAA	2.02	Marker of neuronal viability
Choline	Cho	3.21	Cell proliferation and membrane synthesis
Creatine	Cr	3.03	Energy storage and metabolism
Glutamate	Glx	2.1–2.5 3.1–3.8	Mediation of excitatory signals
Lactate	Lac	1.3	Cell hypoxia
Myoinositol	mI	3.55	Glial activity and neurodegeneration
Lipids	Lip	1.3	Cell necrosis, aggressiveness

Table 8.2 Main metabolites and frequency positions that are commonly analysed and studied as disease biomarkers in prostate

Metabolite	Abbreviation	Frequency (ppm)	Function
Citrate	Cit	2.5–2.8	Synthesized by the prostate gland
Creatine	Cr	3.03	Energy storage and metabolism
Choline	Cho	3.21	Cell proliferation and membrane synthesis

accurately measure the water concentration. In the relative quantification, the ratios of the areas of the metabolites are provided. With this method, reduced acquisition times can be obtained since TR can be shorter (commonly around 2 s). Creatine (Cr) is the preferred tissue metabolite to be used as a reference for the relative quantification, since it is related to the levels of energy stores within the human body and has quite stable values between subjects.

The metabolites of interest are different depending on the organ or tissue under study. In Table 8.1, the most important metabolites for MRS analysis in brain are listed:

In Table 8.2, the main metabolites for the diagnosis of prostate abnormalities by MRS are presented.

The quantification of choline concentration as an indicator of cell proliferation and tumour aggressiveness can be extended to other organs and tissues, such as breast and rectum with low movement artefacts and safe from signal heterogeneities.

8.5 Disease Biomarkers in Brain

The MRS technique has been widely applied in many different brain disorders with excellent results as a disease biomarker. An understanding of the lesion subjacent biology is needed for the spectra interpretation.

Table 8.3 1H MRS ranges of the metabolite concentration in the posterior cingulate acquired with a TE of 32 ms for the diagnosis of cognitive impairment versus normal aging

	NAA/Cr		mI/Cr		NAA/mI
MCI	<1.4	or	>0.65	and	1.5–2.0
AD	<1.3	and	>0.7	and	<1.5

8.5.1 Neurodegenerative Diseases

Various 1H MRS studies in Alzheimer's disease (AD) have shown a decrease of NAA and an increased in mI compared to healthy subjects (Martínez-Bisbal et al. 2004; Pfefferbaum et al. 1999; Jessen et al. 2000). In mild cognitive impairment (MCI), metabolic abnormalities have been also described, with increased mI/Cr and a decrease of the NAA/Cr ratio with respect to the healthy subjects (Martínez-Bisbal et al. 2004). The MR images show structural differences in these cases in relation to cognitive impairment, but only in advance and severe situations.

The most extended evaluated regions for spectroscopy acquisition in patients with cognitive impairment are the posterior cingulate (see Fig. 8.6 left) and the temporal lobe, which are significantly affected in MCI and AD (Kantarci et al. 2000).

The experience of the authors in the spectroscopy analysis of cognitive impairment patients has permitted the definition of some thresholds to differentiate between healthy, MCI and AD patients that can be observed in Table 8.3.

8.5.2 Brain Tumours

The study of brain tumours has been the most extended application of MRS. The concentration of the different metabolites is directly related to the tumour biology and malignancy. The metabolic profile of more aggressive tumours tends to present significantly reduced NAA/Cr levels as an expression of the reduced neuronal expression, high Cho/Cr levels related to the processes of cell proliferation and membrane synthesis and also increased lipids content due to cell necrosis (Fig. 8.8). Sometimes, in very aggressive lesions, when tumour oxygen demand exceeds the vessels' blood supply and cells are under hypoxia conditions, the lactate is also produced and can be only detected at long TE MRS, since at short TE lactate is coupled with fat signal and cannot be isolated. Low-grade tumours present a moderate decrease in the NAA/Cr levels and a slight increase in the Cho/Cr levels, without increase in cell necrosis and lipid content (see Fig. 8.9).

The common assessment of a brain tumour by MRS is to initially acquire a single voxel with two acquisitions at short and long TE in the enhancement region of the

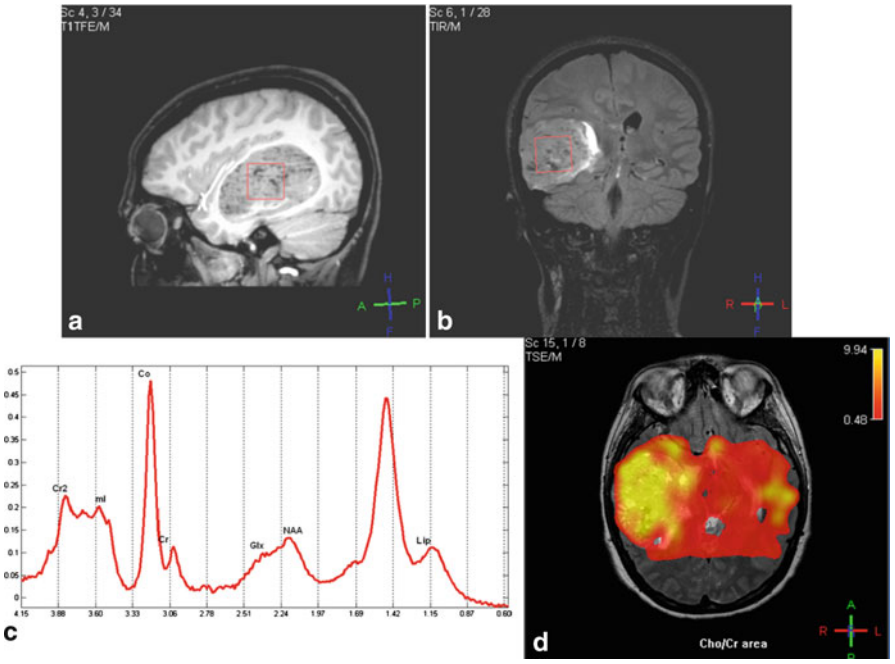


Fig. 8.8 High-grade astrocytoma. In (**a**) and (**b**), single-voxel acquisition positioning for the study of the lesion. In (**c**), spectroscopic profile of the lesion, showing a significantly increased Cho and lipids within the tumour. In (**d**), parametric map of the Cho/Cr ratio obtained from the chemical shift imaging acquisition

lesion, trying to avoid as much as possible the necrotic regions and the areas with high susceptibility differences such as those with haemorrhage and close to bone and air borders. In order to compare the tumour metabolic profile with the normal non-affected parenchyma, a contralateral reference single voxel is also acquired. After single-voxel acquisitions, a multi-voxel approach must be performed to detect infiltration beyond the signal abnormal tumour margins, since in these regions tumoral infiltration is present but not seen in many situations. As an example, the single-voxel acquisitions of a high-grade glioma and a metastasis have a similar spectroscopic profile; however, in the multi-voxel acquisitions, the peripheral oedematous region of a high-grade glioma has a significant increase of the Cho, while metastasis has a decay of the Cho levels in this area since metastatic margins are better defined and have a relatively low infiltration ratio.

One of the major issues in brain tumour follow-up after radiotherapy is the differentiation between radionecrosis effects and tumour recurrence. A high Cho is a manifestation of tumour recurrence, while radionecrosis produces a general decrease in all the metabolite concentrations with significant reductions in Cho, Cr and NAA and an increase in lipid content due to the necrosis.

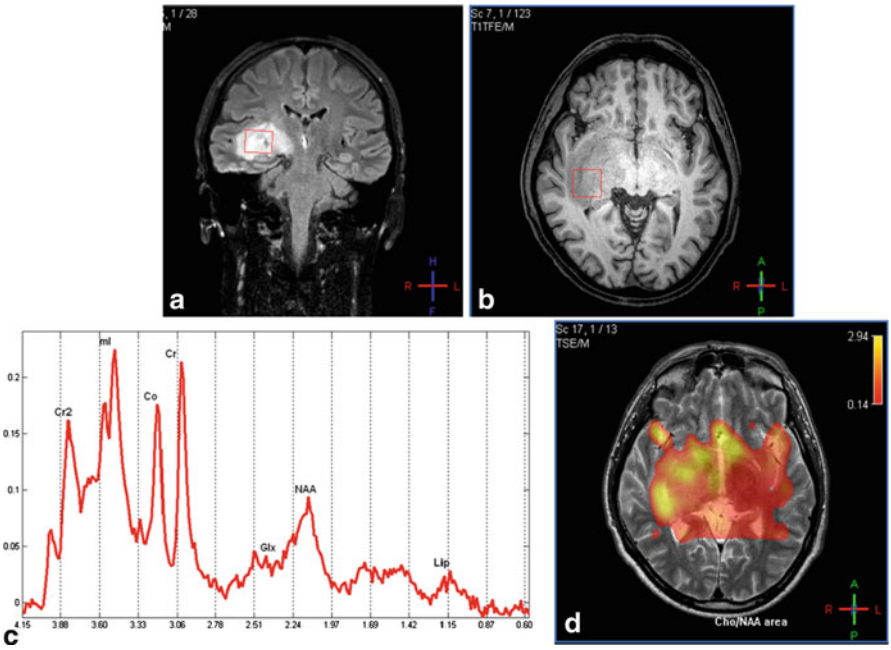


Fig. 8.9 Low-grade glioma. In (**a**) and (**b**), single-voxel acquisition positioning for the study of the lesion. In (**c**), spectroscopic profile of the lesion, showing a decreased NAA and slight increase in Cho in the lesion. In (**d**), parametric map of the Cho/NAA ratio obtained from the chemical shift imaging acquisition

Some automated classification systems for clinical decision support have been developed from large spectra data sets analysed by stepwise and peak integration feature extraction methods (Fuster-Garcia et al. 2011). These tools can aid in the classification of the tumour type and in the evaluation of the degree of malignancy.

8.5.3 Other Diseases

MRS in multiple sclerosis has emerged as a valuable tool to demonstrate not only changes in lesions related to activity but also modifications in the normal-appearing white matter, even early in the process (Narayana 2005). Main research results indicate that there is a clear relationship between the NAA decline and the clinical disability in patients with normally appearing white matter. It seems that MRS may be a stronger predictor than lesional load in the assessments of very early disease stages (Wolinsky and Narayana 2002).

The sensitivity of MRS has also been applied to temporal lobe epilepsy. Although it is known that a significant reduction of NAA exists in the primary epileptogenic focus in the hippocampal region associated to neuronal cell loss

and hippocampal atrophy, a general decrease in NAA has also been observed earlier and beyond the primary epileptogenic focus in brain white matter (Mueller et al. 2002).

8.5.4 Paediatric Disorders

The developing brain in the first years of life has a continuously changing spectroscopic profile. Newborns usually present high Cho concentrations and low NAA when compared to the adult brain. After approximately 4 years of age, brain maturation process stabilizes and metabolite concentrations are close to those observed in adults with a decrease in Cho and increase in NAA, reflecting the proper brain maturation. These variations allow the use of MRS in the diagnosis and follow-up of maturation-related disorders (Xu and Vigneron 2010). Even more, MRS has been also found to have a potential application in the evaluation of foetal brain maturation (Kok et al. 2002).

8.6 Disease Biomarkers in Prostate Cancer

The development of high-field scanners permitted the acquisition of MRS also in other organs and regions outside the brain where the technique is more cumbersome. The most extended application of MRS outside the brain is in the detection and monitoring of prostate cancer. Prostate MRS acquisition is not straightforward, since the gland is relatively small and far from the skin and surface coils that are placed for signal reception in imaging. To increase signal, endorectal coils are frequently used for the acquisition of MRS of the prostate. However, if some considerations are taken in the preparation of the patient, MRS can be properly acquired also with the use of surface coils and 3 T field strengths (Scheenen et al. 2007).

The multi-voxel acquisition is performed using long TE in order to have a good baseline for the quantification of Cho, Cr and Cit concentrations. Malignant prostate lesions present an increase of Cho and a decrease in the Cit concentration (Fig. 8.10). Usually, a metabolic quotient given by (Cho + Cr)/Cit is considered a good biomarker of a potential malignant region. The classification that has been widely extended in prostate cancer diagnosis by MRS differentiates between ‘normal voxel’, ‘suspicious of cancer’ and ‘highly suspicious of cancer’ depending on the metabolic quotient results (Kurhanewicz et al. 1996) . Other approaches to the prostate cancer classification have been considered, which take into consideration the (Cho+Cr)/Cit value in the normally appearing prostate tissue (usually contralateral) as the reference (Scheenen et al. 2007) (Table 8.4).

Apart from the classification of the different prostate gland MRSI voxels in different levels of risk of having a prostate cancer, it is also important to localize the region of risk within the prostate volume. In order to minimize the difficulty

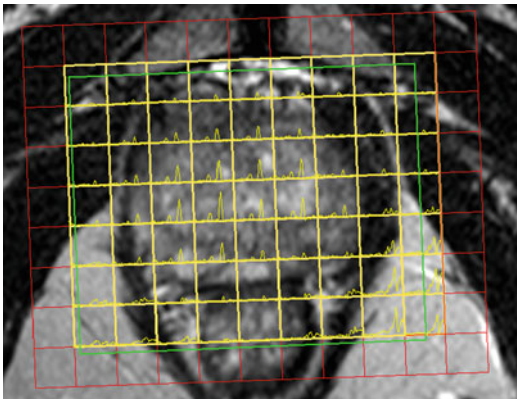


Fig. 8.10 Multi-voxel prostate MRS acquisition. Representative multi-voxel acquisition of the prostate in a patient with suspicious cancerous lesion in the left periphery. Note the decrease in the Cit signal

Table 8.4 Prostate MRS voxel classification (Kurhanewicz et al. 1996)

Voxel-based classification	(Cho + Cr)/Cit
Normal	<0.75
Suspicious of cancer	>0.75
Highly suspicious of cancer	>0.86

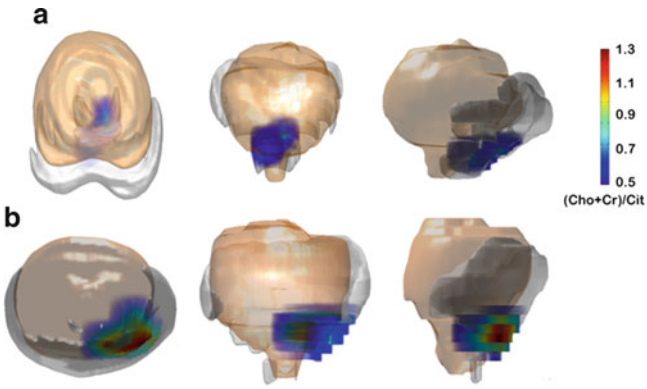


Fig. 8.11 3D volumetric reconstructions of the prostate with metabolic quotient. In (**a**), 3D parametric visualization of a region with slight increase of Cho and decrease of Cit in a patient with benign prostatic hyperplasia. In (**b**), 3D parametric visualization of a region with high probability of malignancy that can aid in the biopsy procedures

in the analysis of the prostate MRS results, semiautomated pipelines for MRS quantification with volumetric representation can be implemented (Fig. 8.11). These volumetric reconstructions with parametric superimposition of the metabolic quotient can be also applied to guide biopsy and interventional therapeutic procedures. The ongoing technical developments in multiparametric prostate MR acquisitions and image-guided procedures will allow the inclusion of all the imaging biomarkers in ultrasound-navigated prostate procedures, converting the current blind biopsies and interventions into obsolete techniques.

8.7 Other Applications

MRS has also been applied to the liver parenchyma as a disease biomarker of steatosis with great success. It has been proven to be one of the most accurate methods for the quantification of the hepatic fat fraction (Cowin et al. 2008).

The application to the detection and aggressiveness characterization of breast lesions is also of growing interest (Bolan et al. 2005). The main difficulties in breast MRS arise from the fact that both a high water and fat concentrations exist and it is difficult to isolate the Cho peak for the characterization of the small areas of tumour proliferation.

Finally, some results have also been shown in the use of MRS for the characterization of lipids in muscle, permitting the differentiation between intra-myocyte lipids (long-term lipids reserve) and extra-myocyte lipids (interstitial triglycerides) (Fayad et al. 2010).

8.8 Conclusion

MRS has become a reference technique for evaluating the metabolism of different tissues in vivo, with special application to brain and prostate lesions characterization. It allows detecting metabolic changes related to degeneration and cancer that cannot be appreciated visually in conventional imaging techniques. The evolution of the acquisition technology with higher spatial resolution and the automation of the spectra processing analysis will encourage the wider application of the technique in many neurodegenerative and oncologic applications including grading and follow-up.

References

Bolan PJ, et al. Imaging in breast cancer: magnetic resonance spectroscopy. Breast Cancer Res. 2005;7:149–52.

Cowin GJ, et al. Magnetic resonance imaging and spectroscopy for monitoring liver steatosis. J Magn Reson Imaging. 2008;28(4):937–45.

Fayad LM, et al. Quantification of muscle choline concentrations by proton MR spectroscopy at 3 T: technical feasibility. AJR Am J Roentgenol. 2010;194:W73–9.

Fuster-Garcia E, et al. Compatibility between 3 T 1H SV-MRS data and automatic brain tumour diagnosis support systems based on databases of 1.5 T 1H SV-MRS spectra. MAGMA. 2011;24:35–42.

Gili J. Introducción Biofísica a la Resonancia Magnética aplicada a la clínica. V(09-1). RPI: B-49123; 2009.

Jessen F, Block W, Traber F, Keller E, Flacke S, Papassotiropoulos A, et al. Proton MR spectroscopy detects a relative decrease of N-acetylaspartate in the medial temporal lobe of patients with AD. Neurology. 2000;55:684–8.

Kantarci K, Jack CR, Xu YC, Campeau NG, O'Brien PC, Smith GE, et al. Regional metabolic patterns in mild cognitive impairment and Alzheimer's disease. A 1HMRS study. Neurology. 2000;55:210–17.

Kok RD, van den Berg PP, van den Bergh AJ, Nijland R, Heerschap A. Maturation of the human fetal brain as observed by 1H MR spectroscopy. Magn Reson Med. 2002;4:611–16.
Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Carroll P, Nelson SJ. Three-dimensional H-1 MR spectroscopic imaging of the in situ human prostate with high (0.24–0.7-cm^3) spatial resolution. Radiology. 1996;198:795–805.
Martínez-Bisbal MC, Arana E, Martí-Bonmatí L, Molla E, Celda B. Cognitive impairment: classification by 1H magnetic resonance spectroscopy. Eur J Neurol. 2004;11:187–93.
Mueller SG, Suhy J, Laxer KD, Flenniken DL, Axelrad J, Capizzano AA, Weiner MW. Reduced extrahippocampal NAA in mesial temporal lobe epilepsy. Epilepsia. 2002;43:1210–16.
Narayana PA. Magnetic resonance spectroscopy in the monitoring of multiple sclerosis. J Neuroimaging. 2005;15:46S–57S.
Naressi A, Couturier C, Devos JM, Janssen M, Mangeat C, de Beer R, Graveron-Demilly D. Java-based graphical user interface for the MRUI quantitation package. MAGMA. 2001;12:141–52.
Pfefferbaum A, Adalsteinsson E, Spielman D, Sullivan EV, Lim KO. In vivo brain concentrations of N-acetyl compounds, creatine, and choline in Alzheimer disease. Arch Gen Psychiatry. 1999;56:185–92.
Provencher SW. Automatic quantitation of localized in vivo 1H spectra with LCModel. NMR Biomed. 2001;14:260–4.
Salibi N, Brown MA. Clinical MR spectroscopy: first principles. New York: Wiley-Liss; 1997.
Scheenen TW, Heijmink SW, Roell SA, de Kaa CA H-V, Knipscheer BC, Witjes JA, Barentsz JO, Heerschap A. Three-dimensional proton MR spectroscopy of human prostate at 3T without endorectal coil: feasibility. Radiology. 2007;245:507–16.
van den Boogaart A, van Ormondt D, Pijnappel WWF, de Beer R, Ala-Korpela M. HLSVD water filtering. In: McWhirter JG, editor. Mathematics in signal processing III. Oxford: Clarendon; 1994. p. 175–95.
Vanhamme L, van den Boogaart A, Van Huffel S. Improved method for accurate and efficient quantification of MRS data with use of prior knowledge. J Magn Reson. 1997;129:35–43.
Wolinsky JS, Narayana PA. Magnetic resonance spectroscopy in multiple sclerosis: window into the diseased brain. Curr Opin Neurol. 2002;15:247–51.
Xu D, Vigneron D. Magnetic resonance spectroscopy for imaging the newborn brain – a technical review. Semin Perinatol. 2010;34:20–7.

Luis Marti-Bonmati, Ph.D., M.D., Professor, Spain Dr. Marti-Bonmati is a professor and chairman of Radiology Department at Quirón Hospital and professor of Radiology at the Valencia University, Spain. He served as the president of ESMRMB in 2002, Spanish Society of Radiology during 2009–2010, and Spanish Society of Abdominal Radiology during 2000–2005, vice-president of ESGAR in 2011–2013, chairman of Research Committee in European Society of Radiology in 2010, and president of Association for the Development and Research of MR.

A. Alberich-Bayarri, Ph.D., Spain Dr. Alberich-Bayarri is a biomedical research engineer in the Department of Radiology of Quiron Valencia Hospital and member of the European Society for Magnetic Resonance in Medicine and Biology (ESMRMB) Institute of Electric and Electronic Engineers (IEEE), Engineering in Medicine and Biology Society (EMBS), and the Association for the Development and Research of MR.

Chapter 9
Processing of Mass Spectrometry Data in Clinical Applications

Dario Di Silvestre, Pietro Brunetti, and Pier Luigi Mauri

Abstract Mass spectrometry-based proteomics has become the leading approach for analyzing complex biological samples at a large-scale level. Its importance for clinical applications is more and more increasing, thanks to the development of high-performing instruments which allow the discovery of disease-specific biomarkers and an automated and rapid protein profiling of the analyzed samples. In this scenario, the large-scale production of proteomic data has driven the development of specific bioinformatic tools to assist researchers during the discovery processes. Here, we discuss the main methods, algorithms, and procedures to identify and use biomarkers for clinical and research purposes. In particular, we have been focused on quantitative approaches, the identification of proteotypic peptides, and the classification of samples, using proteomic data. Finally, this chapter is concluded by reporting the integration of experimental data with network datasets, as valuable instrument for identifying alterations that underline the emergence of specific phenotypes. Based on our experience, we show some examples taking into consideration experimental data obtained by multidimensional protein identification technology (MudPIT) approach.

Keywords Mass spectrometry-based proteomics • Disease-specific biomarkers • Bioinformatic tools • Algorithms • Integration • Multidimensional protein identification technology

D. Di Silvestre • P. Brunetti • P.L. Mauri (✉)
Proteomics and Metabolomics Laboratory, Institute
for Biomedical Technologies – National Research Council,
Via F.lli Cervi 93, 20090 Segrate, Milan, Italy
e-mail: pierluigi.mauri@itb.cnr.it

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_9,

9.1 Introduction

The increasing availability of fully sequenced genomes is making the high-throughput proteomics research more and more possible. Developments in fractionation approaches coupled to advances in liquid chromatography (LC), mass spectrometry (MS), and bioinformatic tools have made proteomic approaches mature to analyze complex proteomes, such as *Homo sapiens* (Nilsson et al. 2010). In fact, although proteome complexity prevents the quantitative profiling of all proteins expressed in a cell or tissue at a given time, higher sensitivity, accuracy, and resolution of new MS instruments allow routine analysis, reaching limit of detection of attomole and dynamic range of $1e^6$ (Yates et al. 2009).

High-throughput proteomics approaches allow to identify and quantify hundreds of proteins per sample, giving a snapshot of cells or tissues associated with different phenotypes. This wealth of data has driven strategies of investigation based on systems biology approaches, allowing insight into disease, taking into consideration functional relationship among proteins (Gstaiger and Aebersold 2009). In addition, highly specific biomarkers represent also key features for improving methods of diagnosis and prognosis or for monitoring disease progression under appropriate therapeutic approaches (Palmblad et al. 2009; Simpson et al. 2009). In this context, MS has been introduced as a tool for enhancing the current clinical application practices and potentially for targeting the development of personalized medicine (Brambilla et al. 2012).

The ultimate success of MS-based proteomics analysis, both for research and for clinical applications, may be affected by several aspects. Like sample preparation, pre-fractionation methodologies, or instrument setup, data processing procedures represent an important step for obtaining good results and their correct interpretation. Evaluation of thousands of data by hand/eye is time consuming and subjected to biases and missed results. Therefore, to assist researchers during the different stages of analysis and to improve understanding of biological systems, an increasing number of tools and procedures are continually developing, giving rise to a specific bioinformatics area for proteomic applications (Di Silvestre et al. 2011).

In this chapter, we make an overview of the computational trends for processing proteomic data obtained by MS-based proteomics approaches. Based on our experience, we focused primarily on strategies related to multidimensional protein identification technology (MudPIT) approach (Mauri and Scigelova 2009), (Fig. 9.1). In particular, we have explored methods, algorithms, and procedures used for biomarker discovery, by means of label and label-free methods. In this context, we then introduce the main advances of the targeted proteomics (Lange et al. 2008) by investigating the bioinformatics aspects concerning the identification of proteotypic peptides (Craig et al. 2005; Kuster et al. 2005). In the second part of the chapter, we discuss recent advances regarding clinical proteomics application for discriminating sample, such as diseased and healthy. Finally, since most known mechanisms leading to disease involve multiple molecules, we conclude with a discussion of the integration of proteomic data with network datasets, as a promising framework for identifying subnetwork that underlines the emergence of specific phenotypes.

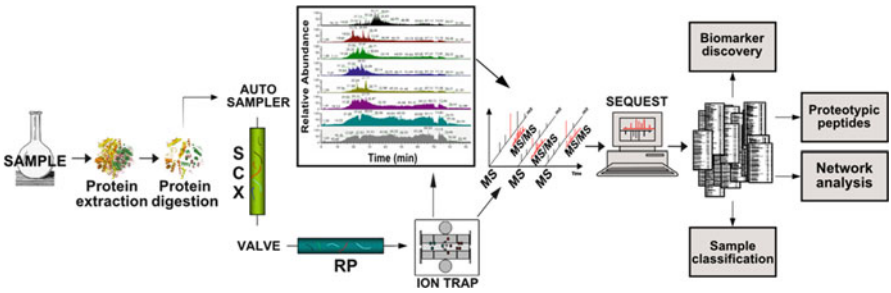


Fig. 9.1 Multidimensional protein identification technology represents a fully automated technology that simultaneously allows separation of digested peptides, their sequencing, and identification of the corresponding proteins. Peptides are separated by means of strong ion exchange (*SCX*), using steps of increasing salt concentration, followed by C18 reverse phase (*RP*) chromatography, using an acetonitrile gradient. Finally, eluted peptides are directly analyzed by MS and raw spectra processed by specific algorithms and bioinformatics tools (see Supplemental Information Table 1). In this way, MudPIT permits simultaneous identification of hundreds, or even thousands, of proteins without limits related to pI, MW, or hydrophobicity. This huge amount of data represents a rich source of information, and their content may be exploited for discovery and classification approaches

9.2 Biomarker Discovery

Quantification of proteomic differences between samples at different biological condition, such as healthy and diseased, is a helpful strategy for providing important biological and physiological information concerning disease state (Simpson et al. 2009; Abu-Asab et al. 2011). For this purpose, MS-based approaches are applied for identifying proteins changing their abundance by comparing two or more samples. They consist of different strategies basically belonging to two categories which rely on stable isotope-labeling and label-free methodologies (Domon and Aebersold 2010).

As for labeling approaches, isotopes are introduced in the peptides to create a specific mass tag recognized by MS (Kline and Sussman 2010). Accordingly, quantification is achieved by measuring the ratio of the signal intensities between the unlabeled peptide and its identical counterpart enriched with isotopes (further details on stable isotope-labeling methods are reported in Supplemental Information). Absolute measurements of protein concentration may be achieved with spiked synthetic peptides, as in QconCAT (Mirzaei et al. 2008), AQUA (Gerber et al. 2003), SISCAPA (Anderson et al. 2004), VICAT (Lu et al. 2007), and PC-IDMS (Barnidge et al. 2004). Quantification is obtained by adding into the sample a known amount of an isotopically labeled peptide. In this way, the level of the endogenous form of peptide can be calculated. Of course, the identity of the peptide must be known prior to analysis by MS. Sometimes, if the m/z ratio of the spiked standard is the same of other peptides, it may lead to an inaccurate quantification. In this case, the ambiguity of the results may be minimized combining these approaches with the selected reaction monitoring (SRM) (Lange et al. 2008).

Although the approaches by labeling, with or without internal standard, allow a highly reproducible and accurate quantification of proteins, most of them have potential limitations, such as the complexity of sample preparation, the requirement of a large amount of time, the requirement of specific bioinformatics tools, and the high cost. As opposite, a simpler alternative concerns label-free approaches (Zhu et al. 2010). They are basically based on counting of peptides identified by means of tandem mass spectrometry (MS/MS) or by evaluating the signal intensity of peptides. Spectral sampling is directly proportional to the relative abundance of the protein in the mixture and therefore represents an attractive methodology, thanks to their intrinsic simplicity, throughput, and low cost.

For these reasons, researchers are increasingly turning to label-free shotgun proteomics approaches (Zhu et al. 2010). Even if they are less accurate, due to the systematic and nonsystematic variations between the experiments, they represent an attractive alternative for their high-throughput setting that also allows the comparison of an unlimited number of experiments with less time consumed. However, efforts should be made to improve experimentally reproducibility and so consequently the reliability of differentially expressed proteins.

A variety of label-free methodologies for semiquantitative evaluation of proteins have been described in literature by reporting a direct relationship between the protein abundance and the sampling parameters associated with identified proteins and peptides (Florens et al. 2002; Gao et al. 2003; Wang et al. 2003; Bridges et al. 2007). One of the most diffused approaches uses the spectral count (SpC) value (Liu et al. 2004) and is based on the empirical observation that more is the quantity of a protein in a sample and more tandem MS spectra may be collected for its peptides. In this context, the normalized spectral abundance factor (NSAF), or its natural log transformation, has been used for the quantitative evaluation with t-test analysis (Zybailov et al. 2006). Other authors have used the protein abundance index (PAI or emPAI) that is calculated by dividing, for each protein, the number of observed peptides by the number of all possible detectable tryptic peptides (Ishihama et al. 2005), while Zhang and colleagues processed SpC values by means of the statistical G-test as previously described (Zhang et al. 2006).

The need to automate the procedure for identifying biomarkers has driven many research groups to develop algorithms and in-house software for identification, visualization, and quantification of mass spectrometry data. Census (Park et al. 2008) and MSQuant (Mortensen et al. 2010) software allows protein quantification by processing MS and MS/MS spectra and they are compatible with label and label-free analysis as well as with high- and low-resolution MS data. In addition, Protein-Quant Suite (Mann et al. 2008) and ProtQuant (Bridges et al. 2007) software are attractive because they allow processing of data in different file formats, therefore collected by different types of mass spectrometers. This aspect focuses attention on the standardization of mass spectrometric data for their sharing and dissemination. In fact, over the years, MS instrument manufacturers have developed proprietary data formats, making it difficult. However, to address this limitation, several tools, such as Trans-Proteomic Pipeline (Deutsch et al. 2010),

allow the conversion of MS data in standard format, like mzData, mzXML, or mzML (Orchard et al. 2010).

The list of computational tools developed for label-free quantitative analysis, by using LC-MS data, is very long. In addition to Corra (Brusniak et al. 2008) and APEX (Braisted et al. 2008) tools, PatternLab (Carvalho et al. 2008) allows different data normalization strategies, such as Total Signal, log preprocessing (by ln), Z normalization, Maximum Signal, and Row Sigma, for implementing ACFold and nSVM (natural support vector machine) methods to identify protein expression differences.

Based on our experience on proteomic analysis based on MudPIT approach, we developed a simple tool, called MAProMA (Multidimensional Algorithm Protein Map) (Mauri and Dehò 2008). It is based on a label-free quantitative approach based on processing of score/SpC values, by means of Dave and DCI algorithms (see Supplemental Information). Its effectiveness has been demonstrated in various studies (Mauri et al. 2005; Regonesi et al. 2006; Bergamini et al. 2012; Simioniuc et al. 2011). In addition, MAProMa allows the comparison of up to 125 protein lists and data visualization in a format more comprehensible to biologists (Fig. 9.2).

9.3 Proteotypic Peptides

A limitation of shotgun proteomics is due to potential inference problem that may affect protein quantification (Nesvizhskii and Aebersold 2005). In addition, limit of detection "may" exclude the identification of biologically relevant molecules. For identifying, validating, and transferring them to the routine clinical analysis, targeted proteomics or "selected reaction monitoring" (SRM) has been recently developed (Lange et al. 2008; Shipkova et al. 2008; Yang and Lazar 2009). The robustness and the simplicity of its data analysis are ideally suited for detecting and quantifying with high confidence up to 100 proteins per sample. For this purpose, mass spectrometers and bioinformatics tools are set to explore a defined number of proteins of interest, following, for each one, a set of representative peptides with a known *m/z* value. They are fragmented, and the monitoring of a specific daughter fragments allow a combination precursor-product, called "transition," that is highly specific for each amino acid sequence.

These peptides, called proteotypic peptides, describe something typical of a protein. Initially, they were defined as the most observed peptides by the current MS-based proteomics approaches (Craig et al. 2005). Then, other authors added the uniqueness condition for a protein (Kuster et al. 2005), while more recently an empirical definition that defines proteotypic peptide as a peptide observed in more than 50% of all identifications of the corresponding parent protein was appended (Mallick et al. 2007). In other words, these peptides have to be previously identified, with a known MS/MS fragmentation pattern and specific for each targeted protein.

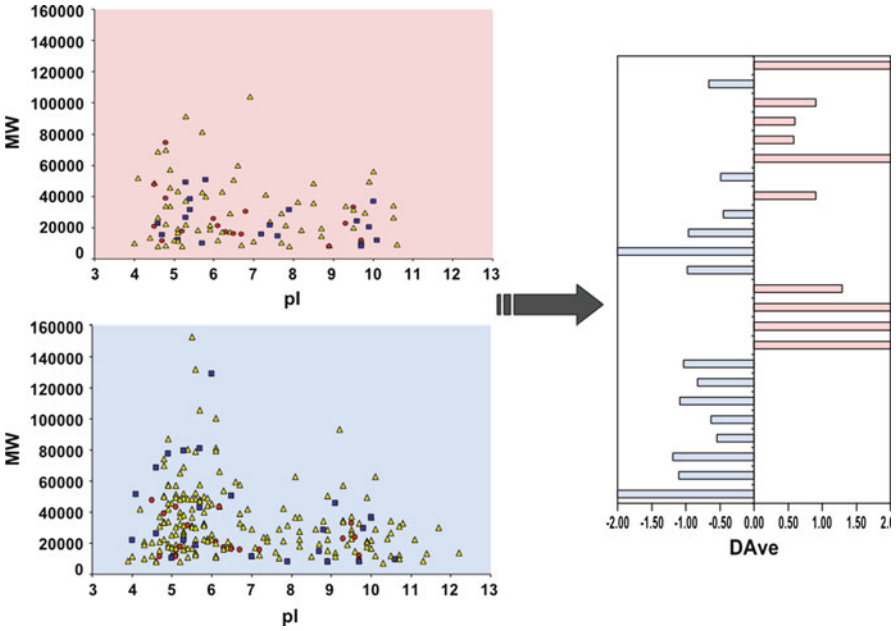


Fig. 9.2 Virtual 2D MAP tool of MAProMa software allows a rapid evaluation of proteins identified by MudPIT by presenting them in the usual form for biologists (maps). It automatically plots in a virtual 2D map the Mw vs. pI for each protein identified, assigning it a color/shape according to a range of a sampling statistics (score or SpC) derived by SEQUEST data handling. This representation permits to have a rapid visual of the proteins that change comparing two or more conditions. In addition, using DAve and DCI algorithms, MAProMa reports a histogram that shows the differentially expressed proteins, their identifier, and their DAve value

The identification of proteotypic peptides useful for targeted proteomics is based on three different methods:

1. By experimental MS/MS data
2. By searching in specific databases
3. By using predictive strategies

The first one is based on the selection of peptides by using the adopted definitions. It allows also the investigation of organisms with proteotypic peptide data not stored inside specific data repositories. In this context, several databases have been developed. In particular:

- Global Proteome Machine Database (Craig et al. 2004) allows users to quickly compare their experimental results with the results previously observed by other scientists. For each dataset, it is possible to view observed spectra for the design of SRM experiments. Query of data may be performed by protein name or Ensembl identifier with the possibility to restrict search to a specific data source, such as eukaryotes, prokaryotes, virus, or precise organism. In addition, further filters may be set by keywords comprising organs, cell location, protein function, or PubMed id.

GPM project is linked to X! Software series (Craig and Beavis 2004) and, of course, with X! P3, the algorithm that makes possible the use of their spectra for profiling proteotypic peptide.

- PeptideAtlas (Desiere 2006) is a publicly accessible source of peptides experimentally identified by tandem mass spectrometry. Raw data, search results, and full builds may be also downloaded. User may browse data, selecting different sources, and few of these need the permission to access. Protein may be searched by different protein identifiers, such as Ensembl and IPI. In addition to general information like GO terms, orthologs, or description, a graphical description indicates the unique peptides found and their occurrence. For each one, it is possible to reach information, like spectra, modification, or genome mapping.

PeptideAtlas is linked to Trans-Proteomic Pipeline (Deutsch et al. 2010) that is used for processing data passed to PeptideAtlas and SBEAMS (Marzolf et al. 2006). In particular, data are processed for deriving the probability of a correct identification and therefore for insuring a high-quality database.

Other databases, designed for data warehousing, store MS/MS spectra collected from proteomics experiments. Even if they are not useful to find proteotypic peptides, they may be used in the comparison with own experimental data.

In particular:

- PRIDE (Martens et al. 2005) stores experiments, identified proteins and peptides, unique peptides, and spectra. In addition to protein (name or various identifiers) and PRIDE experiment identifier, it is possible to browse PRIDE by species, tissue, cell type, GO terms, and disease.
- Proteome Commons (Hill et al. 2010) is a public proteomics database linked to the Tranche (Falkner and Andrews 2007), a powerful open-source web application designed to store and exchange data. A public access to free, open-source proteomics tools, articles, data, and annotations is provided.
- Proteomexchange (Hermjakob and Apweiler 2006) is a work package for encouraging the data exchange and dissemination. Its consortium has been set up to provide a single point of submission of MS data concerning to the main existing proteomics repositories (at the moment PRIDE, PeptideAtlas, and Tranche).

Experimental data stored in the described repositories represent a wealthy source of information, useful for bioinformaticians which attempt to design algorithms for predicting peptides most observable using MS. For this purpose, the STEPP software contains an implementation of a trained support vector machine (SVM) (Cristianini and Shawe-Taylor 2000; Vapnik 1999) that uses a simple descriptor space, based on 35 properties of amino acid, to compute a score representing how proteotypic a peptide is by LC-MS (Webb-Robertson 2009). Similarly to STEPP, a predictor was developed, called Peptide Sieve (Mallick et al. 2007), by studying physicochemical properties of more than 600,000 peptides identified by four different proteomic platforms. This predictor has the ability to accurately

identify proteotypic peptides from any protein sequence and offer starting points for generating a physical model describing the factors that govern elements of proteomic workflows such as digestion, chromatography, ionization, and fragmentation. Other authors, like Tang et al., used neural networks (Riedmiller and Braun 1993) to develop the DetectabilityPredictor software that uses 175 amino acid properties (Tang et al. 2006)**.** In the same way, artificial neural networks were used to predict peptides potentially observable for a given set of experimental, instrumental, and analytical conditions concerning multidimensional protein identification technology datasets (Sanders et al. 2007). Finally, random forest (Breiman 2001) was used to develop enhanced signature peptide (ESP) predictor. It was specifically designed for facilitating the development of targeted MS-based assays for biomarker verification or any application where protein levels need to be measured (Fusaro et al. 2009).

9.4 Classification and Clustering Algorithms

Clinical proteomics aims to use relevant data for improving disease diagnosis or for monitoring its progression (Palmblad et al. 2009; Brambilla et al. 2012). In this context, biomarkers represent a key aspect to develop methods for classifying samples according to their phenotypes (e.g., healthy-diseased, early-late stage).

In addition, to address the biological questions, technologies for high-throughput proteomics allow long lists of spectra, sequenced peptides, and parent proteins that represent a wealthy source of data for identifying predictive biomarkers. For these purposes, most studies have used spectra, generated by MALDI and SELDI technology, in combination with a wide variety of prediction algorithms. On the contrary, fewer cases have taken into consideration data obtained by LC-MS analysis (see Supplemental Information Table 2). However, results of LC-MS analysis (or by MudPIT) are formatted in an $m \times n$ matrix, with a structure very reminiscent of the output of microarray genomics experiments (Fig. 9.3). Hence, the software packages and the tools useful for analyzing genomics data may easily be used for proteomics (Ressom et al. 2008; Dakna et al. 2009).

Even if some properties of proteomic datasets are related to the analytical technology used to generate them, procedures for sample classification basically consist of four steps, such as data preprocessing, feature selection, classification, and cross-validation (Ressom et al. 2008; Dakna et al. 2009; Sampson et al. 2011; Barla et al. 2008). The first one aims to achieve reproducible results by minimizing errors due to the experimental-designed methodology. Mass spectral profiles may be influenced by several factors, such as baseline effects, shifts in mass-to-charge ratio, alignment problem, or differences in signal intensities that may be corrected by specific computational procedures (Yu et al. 2006; Arneberg et al. 2007; Pluskal et al. 2010). In the same way, variation of sampling parameters

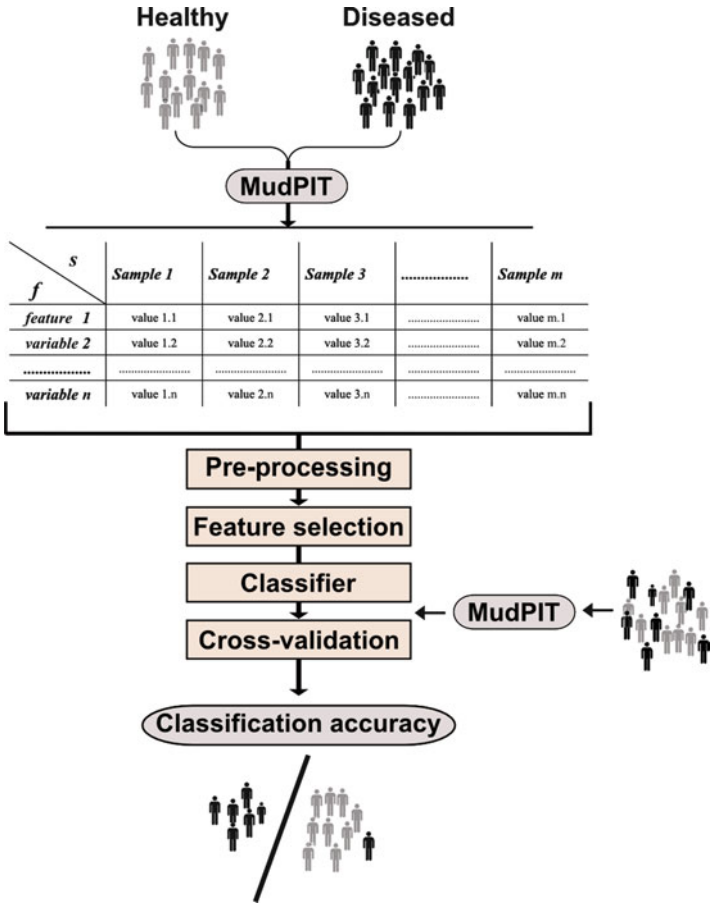


Fig. 9.3 Data matrix is obtained aligning features identified by analyzing sample by MudPIT approach. In this context, MAProMa software allows a rapid alignment of up to 125 protein lists. Rows in data matrix represent features (e.g., proteins, peptides, or *m*/*z* values) while columns indicate samples. Each cell of data matrix is represented by a value corresponding to parameter associated with features. In particular, spectral count, Xcorrelation (Xcorr), and signal intensity are used for protein, peptides, and *m*/*z* mass features, respectively

associated with sequenced protein, such as spectral count or score, is adjusted using related strategies of data normalization (e.g., Total Signal, log preprocessing (by ln), Z normalization, Maximum Signal, or Row Sigma) (Carvalho et al. 2008).

Typically, MudPIT analysis generates a number of variables usually bigger than the number of analyzed samples ($f >> s$). This complexity represents a key problem of computational proteomics, and most classification methods require the reducing of the dimensionality prior to classification. It is obtained by

discarding the irrelevant variables for obtaining a combination of features ($f << s$), highly correlated and with a more informative lower dimensional space that maximizes the quality of the hypothesis learned from these features (Guyon et al. 2006).

Feature selection procedures may be classified in three different approaches based on different processes to rank features: filter, wrapper, and embedded (Levner 2005). A number of techniques have been used for the analysis of proteomic data, and these include methods such as support vector machines (SVM) and artificial neural networks (ANN) as well as approaches like partial least squares (PLS), principal component regression (PCR), and principal component analysis (PCA). A good overview of statistical and machine learning-based feature selection and pattern classification algorithms is reported by Ressom and colleagues (Ressom et al. 2008). Of course, different combinations of them show different sensitivity to noisy data and outliers as well as different susceptibility to the over-fitting problem (Sampson et al. 2011).

A limitation of many machine learning-based classification algorithms is that they are not based on a probabilistic model; therefore, there is no confidence associated with the predictions of new datasets. Inadequate performance could be attributed to different reasons (e.g., insufficient or redundant features, inappropriate model classifier, few or too many model parameters, under- or overtraining, and code error, as well as presence of highly nonlinear relationships, noise, and systematic bias). Thus, with the purpose of testing the adequacy/inadequacy of a classifier, after learning is completed, its performances are evaluated through validation set, previously unseen. For this purpose, various methods, such as *k*-fold cross-validation, bootstrapping, and holdout methods, have been used (Ressom et al. 2008). The most common performance measures to evaluate the performances of classifiers are a confusion matrix and a receiver operating characteristic (ROC) curve. The first one shows information about actual and predicted classifications of a classifier and assesses its performances using standard indices, such as sensitivity, specificity, PPV, NPV, and accuracy values (see Supplemental Information). On the other hand, ROC is a plot of the sensitivity of a classifier against 1-specificity for multiple decision thresholds.

9.5 From Proteomics to Systems Biology

Proteomics is a holistic science that refers to the investigation of the entire systems. Before the advent of -omics technologies, reductionism has dominated the biological research for over a century by investigating individual cellular components. Despite its enormous success, it is more and more evident that most molecular functions occur from a concerted action of multiple molecules, and their investigation implies the examination of an ensemble of elements (Barabási and Oltvai 2004). In fact, biomolecular interactions play a role in the majority of cellular

processes that are regulated connecting numerous constituents, such as DNA, RNA, proteins, and small molecules.

Data abstraction in pathways or networks is the natural result of the desire to rationalize knowledge of complex systems. More recently, their use has changed from purely illustrative to an analytic purpose. In fact, even if it is purely virtual and not related to any intrinsic structure in the cell or organism, understanding how, where, and when single components interact is fundamental to facilitate the investigation of experimental data by taking into consideration the functional relationship among molecules.

A major challenge for biologists and bioinformaticians is to gain tools, procedures, and skills for integrating data into accurate models that can be used to generate hypotheses for testing. This objective is partially the result of the confluence in systems biology of advances in computer science and -omics technologies. In this context, systems biology approaches have evolved in different strategies basically belonging to two categories, such as computational systems biology, which uses modeling and simulation tools (Barrett et al. 2006; Kim et al. 2012), and data-derived systems biology, which relies on "-omics" datasets (Rho et al. 2008; Li et al. 2009; Jianu et al. 2010; Pflieger et al. 2011).

For deciphering mechanisms of complex and multifactorial diseases, such as those concerning heart failure, recent studies have coupled proteomic and systems biology approaches (Wheelock et al. 2009; Isserlin et al. 2010; Arrell et al. 2011). From a standpoint of the data visualization, the possibility to map protein expression onto pathway or network reveals how they are modulated under different conditions, such as healthy and disease states (Gstaiger and Aebersold 2009; Sodek et al. 2008). About that, an unbiased procedure to identify subnetworks which change consistently between different states involves three key steps:

1. The execution of high-throughput proteomic experiments
2. The identification of candidate biomarkers by label or label-free methods
3. The integration of data into network model to identify clusters of proteins with under, over, and normal expression

In addition, subnetwork selected using experimental data may be analyzed by computing network centrality parameters (Scardoni et al. 2009) for identifying proteins with a relevant biological and topological significance (Fig. 9.4). However, some limitations concerning this kind of approaches could be represented by measurements that cover only a small fraction of the network or by organisms with a limited dataset of cataloged protein sequences and interactions.

To date, in order to visualize and analyze biological networks, a wide set of bioinformatic tools are available (Suderman and Hallett 2007) and include well-known examples, such as Cytoscape (Shannon et al. 2003), VisANT (Hu et al. 2005), Pathway Studio (Nikitin et al. 2003), PATIKA (Demir et al. 2002), Osprey (Breitkreutz et al. 2003), and ProViz (Iragne et al. 2005). Among these, Cytoscape is a Java application whose source code is released under the Lesser General Public License (LGPL). It is probably the most famous open-source software platform for

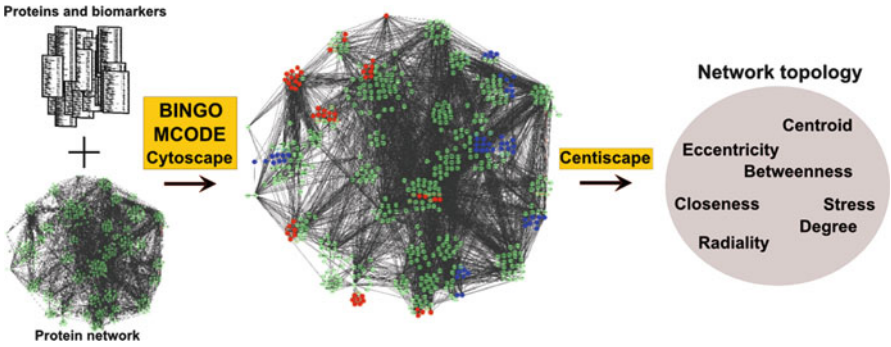


Fig. 9.4 By means of Cytoscape software and its plugins, proteins and biomarkers identified by MudPIT are integrated in protein network for identifying pathways or subnetworks that underline the emergence of specific biological states. In addition, networks identified by experimental data may be analyzed by plugins, such as CentiScape, for calculating centrality parameters that indicate nodes with relevant biological and topological significance

visualizing network datasets and biological pathways and for integrating them with annotations or gene and protein expression profiles. Its core distribution provides a basic set of features. However, additional features are available as plugins, thanks to a big community of developers which uses the Cytoscape open API based on Java technology.

Most of the plugins are freely available and concern tasks like the importing and the visualizing of networks from various data formats, the generating of networks from literature searches, and the analysis or the filtering of them by selecting subsets of nodes and/or interactions in relation with topological parameters, GO annotation, or expression levels. In particular, for analyzing large set of proteomics data, we suggest some plugins, such as:

- CentiScape (Scardoni et al. 2009) that computes specific centrality parameters describing the network topology
- MCODE (Bader and Hogue 2003) that finds clusters or highly interconnected regions
- BiNGO (Maere et al. 2005) that determines the Gene Ontology (GO) categories statistically overrepresented in a set of genes or a subgraph of a biological network
- BioNetBuilder (Avila-Campillo et al. 2007) that offers a user-friendly interface to create biological networks integrated from several databases such as BIND (Alfarano et al. 2005), BioGRID (Stark et al. 2006), DIP (Xenarios et al. 2000), HPRD (Mishra et al. 2006), KEGG (Kanehisa et al. 2004), IntAct (Kerrien et al. 2007), MINT (Zanzoni et al. 2002), MPPI (Pagel et al. 2005), and Prolinks (Bowers et al. 2004) as well as interolog networks derived from these sources for all species represented in NCBI HomoloGene

Other important repositories for protein-protein interaction are STRING (von Mering et al. 2007), Reactome (Joshi-Tope et al. 2005), Pathway Commons (Cerami et al. 2011), and WikiPathways (Pico et al. 2008). However, an exhaustive overview of existing databases is available through the Pathguide website (http://www.pathguide.org/), a useful web resource where about 300 biological pathways and interaction database are described.

9.6 Conclusion

In the last few years, developments in MS instrumentation have increased both the number of identified proteins, reaching hundreds to thousands in a single experiment, and the confidence of such identifications. Thanks to this relevant amount of data, researchers are characterizing the discovery processes by integrating large set of experimental data into models used to generate hypothesis for testing. For this purpose, systems biology approaches provide a powerful strategy for linking biomarker expression with biological processes that can be segmented and linked to disease presentation. Mass spectrometry-based proteomics is emerging also as a powerful approach suitable to face clinical questions. Even if it is an area of still unrealized potential, clinical proteomics offers the promise of diagnosis, prognosis, and therapeutic follow-up of human diseases. However, given the current status of measurement reproducibility and lack of standardization, further comparative investigations are of great importance.

As widely emerged by this chapter, both for basic or clinical research, bioinformatics and statistical tools have a primary importance to support the discovery processes at various levels of sophistication, or for improving the performances of the technologies themselves. In particular, the relevant amount of data produced by the high-throughput proteomics technologies require powerful informatics supports for their organization and interpretation. In this context, several topics concerning data storage, their processing, their visualization, and their interpretation have been faced. However, the need of standards is considered fundamental, and some projects for sharing experimental data between research groups have been launched (e.g., MIAPE, CDISC, and HL7). These should increase meta-analysis, by using raw data from different centers, for helping the development that was grossly underestimated in the initial studies. In addition, as a proteomics community, we believe proteomics methodologies mature for tackling future challenges in clinical proteomics. However, the production of valuable data should rise in step with cooperation with medically focused groups.

Acknowledgments This study was supported by the Italian Ministry of Economy and Finance to the CNR for the Project "FaReBio di Qualita," by Italian Ministry of University and Research for the Project FAR and by Fondazione Cariplo (2010-0653).

Supplemental Information

Introduction

Table 1 Bioinformatics platforms for processing proteomics data

Name	Description	References
Corra	Frameworks for LC-MS analysis	Brusniak et al. (2008)
ATAQS	Pipeline implemented for SRM	Brusniak et al. (2011)
Central Proteomics Facilities Pipeline	Pipeline for the analysis of MS/MS proteomic data	Trudgian et al. (2010)
SASHIMI	Suite of tools for MS/MS proteomics	Deutsch et al. (2010)
MS Data Miner	A web-based software that accepts data from Mascot or other software	Dyrlund et al. (2012)
Katsura	Overlays -omics empirical data onto metabolic pathways	Kanehisa et al. (2004)
ProteoWizard	Set of libraries and tools to perform proteomics data analysis	Kessner et al. (2008)
ProteoConnections	Web-based set of tools using for analyzing proteomic data	Courcelles et al. (2011)
Chipster	A Java Web Start framework that organizes workflows for -omics data	Kallio et al. (2011)
Multiplierz	A scriptable framework that access to manufacturer data files	Parikh et al. (2009)
Proteomatic	A framework that permits to create concatenate scripts in pipeline	Specht et al. (2011)
DAnTE/Inferno	Software to perform statistical analysis on proteomics data	Polpitiya et al. (2008)

Biomarker Discovery (Stable Isotope Labeling)

As for labeling approaches, the stable isotopes may be introduced in the peptide using different methods based on the metabolic, chemical, or enzymatic incorporation (Ong and Mann 2005). Metabolic labeling was described for marking protein of yeast by means of ^{15}N-enriched cell culture medium (Oda et al. 1999). Since the number of labeled nitrogen atoms may vary from peptide to peptide, stable isotope labeling by amino acids in cell culture (SILAC) approach (Ong et al. 2002) was then introduced. In this case, culture medium contains $^{13}C_6$-Lys and $^{13}C_6$ and $^{15}N_4$-Arg, ensuring at least one labeled amino acid of tryptic cleavage products. In addition, samples treated with different isotopes may be combined prior to sample preparation minimizing the potential errors introduced by their handling. However, a low cellular growth in adapted media may represent a potential drawback.

Methods which allow labeling by chemical or enzymatic incorporation overcome some limitations associated with metabolic labeling. They include isotope-coded affinity tags (ICAT), where free cysteine residues are tagged by a reagent containing eight or zero deuterium atoms (Gygi et al. 1999). Tags are linked to the biotin which may be exploited for enriching the labeled peptides using affinity purification prior to MS analysis. Although this strategy reduces the complexity of the peptide mixture, proteins that do not contain cysteine are excluded from the analysis. For this reason, other approaches use reactive residues that occur more frequently in proteins. At this class belong the isobaric tags for relative and absolute quantitation (iTRAQ) (Ross et al. 2004), the tandem mass tags (TMT) (Thompson et al. 1999), and the isotope-coded protein label (ICPL) (Schmidt et al. 2005). In particular, iTRAQ approach is widely used and it is based on the covalent labeling of the N-terminus and side chain amines of peptides. Multiplexing tagging allows the analysis of up to 8 samples per experiment (Choe et al. 2007). In fact, samples differently tagged are pooled and usually fractionated by LC and analyzed by MS/MS. However, problem of co-elution of peptides with similar mass could interfere with the quantification. Finally, as for enzymatic tagging of peptides, recently trypsin-catalyzed ^{18}O labeling has grown in popularity due to its simplicity, its cost, and its ability to universally label peptides. Both C-terminal carboxyl group atoms of tryptic peptides can be enzymatically exchanged with ^{18}O providing a labeled peptide with a 4-Da mass shift from the ^{16}O-labeled sample (Qian et al. 2011).

Biomarker Discovery (DAve and DCI Algorithms)

A direct correlation between the SEQUEST-based score value and the relative abundance of the identified proteins has been previously demonstrated (Mauri et al. 2005; Regonesi et al. 2006). Based on this finding, protein profiles of healthy and diseased samples are semiquantitatively compared using a label-free proteomic approach based on DAve (differential average) and DCI (differential confidence index) algorithms of MAProMA software (Mauri and Dehò 2008).

In particular, DAve, which evaluates changes in protein expression, is defined as

$$\left(X-Y\right)/\left(X+Y\right)*0.5,$$

while DCI, which describes the confidence of differential expression, is defined as

$$\left(X+Y\right)\times\left(X-Y\right)/2$$

where X and Y represent the SEQUEST-based score or SpC values of a given protein in two compared samples.

Conventionally, signs (+/−) of DAve and DCI indicate if proteins are up-regulated in the first or in the second sample, respectively. A value of DAve >0.4 (or ≤ −0.4) corresponds to SCORE ratio ≥1.5. Coupled to a threshold value ≥400 (or ≤ −400)

for DCI, it allows, with a good reliability, to identify differentially expressed (Mauri et al. 2005; Simioniuc et al. 2011; Bergamini et al. 2012) proteins. However, DAve and DCI threshold values may be decreased when calculated, considering mean SCORE values derived from replicate analyses (DAve ≥0.2 or < −0.2 and DCI ≥200 and ≤ −200). On the contrary, when a single analysis per sample/condition is available, a better reliability of the differentially expressed proteins may be assured increasing the threshold values (DAve ≥0.8 or < −0.8 and DCI ≥800 and ≤ −800).

Classification and Clustering Algorithms

Table 2 Classification studies published in the last few years

Sample	References	Technology	Data	Algorithm	Biological condition
Urine	Dawson et al. (2012)	MALDI	**Proteins/ peptides**	**SVM**	Ischemic stroke
Serum	Timms et al. (2011)	MALDI	–	–	Ovarian cancer
Virus	Wong et al. (2010)	MALDI	**Spectra**	**Bayes classifiers**	Influenza viruses
Tissues	Le Faouder et al. (2011)	MALDI-IMS	**Protein peaks**	**SVM**	Carcinoma
Tissues	M'Koma et al. (2011)	MALDI-IMS	**Protein Peaks**	**k-nearest-neighbor**	Colitis
Tissues	Djidja et al. (2010)	MALDI-IMS	**Proteins/ peptides**	**PCA-DA**	Tumor
Acinetobacter spp.	Alvarez-Buylla et al. (2012)	MALDI	–	–	Acinetobacter spp.
Serum	Fan et al. (2012)	MALDI	**Raw data**	**SVM**	Breast cancer
Plasma	Fassbender et al. (2012)	MALDI	–	**SVM**	Endometriosis
Cerebrospinal fluid	Ishigami et al. (2012)	MALDI	**Spectra**	**SVM-PCA**	Parkinson
Cerebrospinal fluid	Komori et al. (2012)	MALDI	**Spectra**	**SVM-PCA**	Multiple sclerosis disorder
Pollen	Krause et al. (2012)	MALDI	**Spectra**	–	Pollen
Tissue	Meding et al. (2012)	MALDI	**Spectra**	**SVM–RF**	Tumor
Serum	Pecks et al. (2012)	MALDI	–		Preeclampsia
Bronchoal-veolar fluid	Frenzel et al. (2011)	MALDI	**Protein peaks**	**SVM**	ALI/ARDS
Urine	Gao et al. (2011)	MALDI	**Protein peaks**	**SVM**	–

(continued)

Table 2 (continued)

Sample	References	Technology	Data	Algorithm	Biological condition
Serum	Han (2010)	MALDI	**Spectra profile**	**SVM**	Hepatitis B
Tissue	Waloszczyk et al. (2011)	MALDI	**Protein peaks**	**–**	Lung cancer
Urine	Balog et al. (2010)	MALDI	**Peptide**	**SVM**	Schistosoma mansoni infection
Tissue	Kim et al. (2012)	MALDI	**Spectra**	**SVM-PCA**	Gastric cancer
Bacterial suspensions	Lasch et al. (2010)	MALDI	**Spectra**	**ANN**	Yersinia
Tissue	Liao et al. (2010)	MALDI	**Spectra**	***k*NN**	Colorectal cancer
Serum	Camaggi et al. (2010)	MALDI	**–**	**RF**	Hepato carcinoma
Plasma	Lin et al. (2012)	SELDI	**Protein peaks**	**PCA**	Lung adenocarcinoma
Serum	Van Gorp et al. (2012)	SELDI	**Protein peaks**	**LS-SVM**	Lymph node status in cerebral cancer
Serum	Zhu et al. (2012)	SELDI	**–**	**SVM**	Pancreatic cancer
Endometrial samples	Kyama et al. (2011)	SELDI	**Protein peaks**	**SVM**	Endometriosis
Serum	Fan et al. (2010)	SELDI	**Protein peaks**	**SVM**	Breast cancer
Serum	Liu et al. (2010)	SELDI	**Spectra**	**Decision tree**	Tuberculosis
Serum	Tang et al. (2010)	SELDI	**Spectra**	**Kernel PLS models**	Ovarian cancer
Tissues	Wang et al. (2010)	SELDI	**Spectra**	**ANN**	Endometriosis
Serum	Song et al. (2012)	SELDI and ELISA	**Protein peaks**	**SVM**	Biliary atresia
Serum	Ahn et al. (2012)	Multiplex array	–	**–**	Gastric adenocarcinoma
Plasma	Izbicka et al. (2012)	Multiplex immunoassays		**SVM**	Lung cancer
Tissue	Lazova et al. (2012)	IMS	**Spectra**	**–**	Spitzoid malignant melanoma
Serum	Sui et al. (2010)	–	**–**	**Genetic algorithm**	Urinemia

SVM support virtual machine, *ANN* artificial neural network, *PCA* principal component analysis, *DA* discriminant analysis, *RF* random forest, *PLS* partial least square

Classification and Clustering Algorithms (Sensitivity, Specificity, PPV, NPV, and Accuracy)

A confusion matrix presents information about actual and predicted classifications made by a classifier. It assesses the classification performance of the classifier. TP, TN, FP, and FN indicate the number of true-positive, true-negative, false-positive, and false-negative samples, respectively. A false positive is when the outcome is incorrectly classified as positive. A false negative is when the outcome is incorrectly classified as negative. True positives and true negatives represent correct classifications. In particular:

Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
Positive predictive value = TP/(TP + FP)
Negative predictive value = TN/(TN + FN)
Overall classification accuracy = (TP + TN)/(TP + TN + FP + FN)

References

Abu-Asab MS, Chaouchi M, Alesci S, Galli S, Laassri M, Cheema AK, Atouf F, VanMeter J, Amri H. Biomarkers in the age of omics: time for a systems biology approach. OMICS. 2011;15:105–12.

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H, Garderman E, Gong Y, Gonzaga R, Grytsan V, Gryz E, Gu V, Haldorsen E, Halupa A, Haw R, Hrvojic A, et al. The biomolecular interaction network database and related tools 2005 update. Nucleic Acids Res. 2005;33:D418–24.

Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW. Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). J Proteome Res. 2004;3:235–44.

Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, Myhr K-M, Vedeler CA, Ulvik RJ, Kvalheim OM. Pretreatment of mass spectral profiles: application to proteomic data. Anal Chem. 2007;79:7014–26.

Arrell DK, Zlatkovic Lindor J, Yamada S, Terzic A. K(ATP) channel-dependent metaboproteome decoded: systems approaches to heart failure prediction, diagnosis, and therapy. Cardiovasc Res. 2011;90:258–66.

Avila-Campillo I, Drew K, Lin J, Reiss DJ, Bonneau R. BioNetBuilder: automatic integration of biological networks. Bioinformatics. 2007;23:392–3.

Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.

Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Brief Bioinform. 2008;9:119–28.

Barnidge DR, Hall GD, Stocker JL, Muddiman DC. Evaluation of a cleavable stable isotope labeled synthetic peptide for absolute protein quantification using LC-MS/MS. J Proteome Res. 2004;3:658–61.

Barrett CL, Kim TY, Kim HU, Palsson BØ, Lee SY. Systems biology as a foundation for genome-scale synthetic biology. Curr Opin Biotechnol. 2006;17:488–92.

Bergamini G, Di Silvestre D, Mauri P, Cigana C, Bragonzi A, De Palma A, Benazzi L, Döring G, Assael BM, Melotti P, Sorio C. MudPIT analysis of released proteins in Pseudomonas aeruginosa laboratory and clinical strains in relation to pro-inflammatory effects. Integr Biol (Camb). 2012;4:270–9.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 2004;5:R35.

Braisted JC, Kuntumalla S, Vogel C, Marcotte EM, Rodrigues AR, Wang R, Huang S-T, Ferlanti ES, Saeed AI, Fleischmann RD, Peterson SN, Pieper R. The APEX quantitative proteomics tool: generating protein quantitation estimates from LC-MS/MS proteomics results. BMC Bioinformatics. 2008;9:529.

Brambilla F, Lavatelli F, Di Silvestre D, Valentini V, Rossi R, Palladini G, Obici L, Verga L, Mauri P, Merlini G. Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue. Blood. 2012;119:1844–7.

Breiman L. Random forests. Mach Learn. 2001;45:5–32.

Breitkreutz B-J, Stark C, Tyers M. Osprey: a network visualization system. Genome Biol. 2003;4:R22.

Bridges SM, Magee GB, Wang N, Williams WP, Burgess SC, Nanduri B. ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. BMC Bioinformatics. 2007;8 Suppl 7:S24.

Brusniak M-Y, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V, Vitek O, Zhang N, Aebersold R, Watts JD. Corra: computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. BMC Bioinformatics. 2008;9:542.

Carvalho PC, Fischer JSG, Chen EI, Yates 3rd JR, Barbosa VC. PatternLab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics. 2008;9:316.

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C. Pathway commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685–90.

Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7.

Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004;3:1234–42.

Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. Rapid Commun Mass Spectrom. 2005;19:1844–50.

Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.

Dakna M, He Z, Yu WC, Mischak H, Kolch W. Technical, bioinformatical and statistical aspects of liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) based clinical proteomics: a critical assessment. J Chromatogr B Analyt Technol Biomed Life Sci. 2009;877:1250–8.

Demir E, Babur O, Dogrusoz U, Gursoy A, Nisanci G, Cetin-Atalay R, Ozturk M. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics. 2002;18:996–1003.

Desiere F. The PeptideAtlas project. Nucleic Acids Res. 2006;34:D655–8.

Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. A guided tour of the trans-proteomic pipeline. Proteomics. 2010;10:1150–9.

Di Silvestre D, Daminelli S, Brunetti P, Mauri P. Bioinformatics tools for mass spectrometry-based proteomics analysis. In: Li P, editor. Reviews in pharmaceutical and biomedical analysis. Bussum: Bentham Science Publishers; 2011. p. 30–51.

Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol. 2010;28:710–21.

Falkner JA, Andrews PC. P6-T Tranche: secure decentralized data storage for the proteomics community. J Biomol Technol. 2007;18:3.

Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. A proteomic view of the *Plasmodium falciparum* life cycle. Nature. 2002;419:520–6.

Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol. 2009;27:190–8.

Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA. Changes in the protein expression of yeast as a function of carbon source. J Proteome Res. 2003;2:643–9.

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003;100:6940–5.

Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet. 2009;10:617–27.

Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature extraction: foundations and applications. Berlin: Springer; 2006.

Hermjakob H, Apweiler R. The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. Expert Rev Proteomics. 2006;3:1–3.

Hill JA, Smith BE, Papoulias PG, Andrews PC. ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository. J Proteome Res. 2010;9:2809–11.

Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C. VisANT: data-integrating visual framework for biological networks and modules. Nucleic Acids Res. 2005;33:W352–7.

Iragne F, Nikolski M, Mathieu B, Auber D, Sherman D. ProViz: protein interaction visualization and exploration. Bioinformatics. 2005;21:272–4.

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics. 2005;4:1265–72.

Isserlin R, Merico D, Alikhani-Koupaei R, Gramolini A, Bader GD, Emili A. Pathway analysis of dilated cardiomyopathy using global proteomic profiling and enrichment maps. Proteomics. 2010;10:1316–27.

Jianu R, Yu K, Cao L, Nguyen V, Salomon AR, Laidlaw DH. Visual integration of quantitative proteomic data, pathways, and protein interactions. IEEE Trans Vis Comput Graph. 2010;16:609–20.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428–32.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct–open source resource for molecular interaction data. Nucleic Acids Res. 2007;35:D561–5.

Kim HU, Sohn SB, Lee SY. Metabolic network modeling and simulation for drug targeting and discovery. Biotechnol J. 2012;7:330–42.

Kline KG, Sussman MR. Protein quantitation using isotope-assisted mass spectrometry. Annu Rev Biophys. 2010;39:291–308.

Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol. 2005;6:577–83.

Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol. 2008;4:222.

Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics. 2005;6:68.

Li J, Zimmerman LJ, Park B-H, Tabb DL, Liebler DC, Zhang B. Network-assisted protein identification and data interpretation in shotgun proteomics. Mol Syst Biol. 2009;5:303.
Liu H, Sadygov RG, Yates 3rd JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004;76:4193–201.
Lu Y, Bottari P, Aebersold R, Turecek F, Gelb MH. Absolute quantification of specific proteins in complex mixtures using visible isotope-coded affinity tags. Methods Mol Biol. 2007;359:159–76.
Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21:3448–9.
Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol. 2007;25:125–31.
Mann B, Madera M, Sheng Q, Tang H, Mechref Y, Novotny MV. ProteinQuant suite: a bundle of automated software tools for label-free quantitative proteomics. Rapid Commun Mass Spectrom. 2008;22:3823–34.
Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. PRIDE: the proteomics identifications database. Proteomics. 2005;5:3537–45.
Marzolf B, Deutsch EW, Moss P, Campbell D, Johnson MH, Galitski T. SBEAMS-microarray: database software supporting genomic expression analyses for systems biology. BMC Bioinformatics. 2006;7:286.
Mauri P, Dehò G. A proteomic approach to the analysis of RNA degradosome composition in Escherichia coli. Methods Enzymol. 2008;447:99–117.
Mauri P, Scigelova M. Multidimensional protein identification technology for clinical proteomic analysis. Clin Chem Lab Med. 2009;47:636–46.
Mauri P, Scarpa A, Nascimbeni AC, Benazzi L, Parmagnani E, Mafficini A, Peruta MD, Bassi C, Miyazaki K, Sorio C. Identification of proteins released by pancreatic cancer cells by multidimensional protein identification technology: a strategy for identification of novel cancer markers. FASEB J. 2005;19:1125–7.
Mirzaei H, McBee JK, Watts J, Aebersold R. Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics. Mol Cell Proteomics. 2008;7: 813–23.
Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, et al. Human protein reference database–2006 update. Nucleic Acids Res. 2006;34:D411–14.
Mortensen P, Gouw JW, Olsen JV, Ong S-E, Rigbolt KTG, Bunkenborg J, Cox J, Foster LJ, Heck AJR, Blagoev B, Andersen JS, Mann M. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. J Proteome Res. 2010;9:393–403.
Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data the protein inference problem. Mol Cell Proteomics. 2005;4:1419–40.
Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio–the analysis and navigation of molecular networks. Bioinformatics. 2003;19:2155–7.
Nilsson T, Mann M, Aebersold R, Yates 3rd JR, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods. 2010;7:681–5.
Orchard S, Albar J-P, Deutsch EW, Eisenacher M, Binz P-A, Hermjakob H. Implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. Proteomics. 2010;10:1895–8.
Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes H-W, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005;21:832–4.
Palmblad M, Tiss A, Cramer R. Mass spectrometry in clinical proteomics – from the present to the future. Proteomics Clin Appl. 2009;3:6–17.
Park SK, Venable JD, Xu T, Yates 3rd JR. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods. 2008;5:319–22.

Pflieger D, Gonnet F, de la Fuente van Bentem S, Hirt H, de la Fuente A. Linking the proteins – -elucidation of proteome-scale networks using mass spectrometry. Mass Spectrom Rev. 2011;30:268–97.

Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6:e184.

Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395.

Regonesi ME, Del Favero M, Basilico F, Briani F, Benazzi L, Tortora P, Mauri P, Dehò G. Analysis of the Escherichia coli RNA degradosome composition by a proteomic approach. Biochimie. 2006a;88:151–61.

Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R. Classification algorithms for phenotype prediction in genomics and proteomics. Front Biosci. 2008;13:691–708.

Rho S, You S, Kim Y, Hwang D. From proteomics toward systems biology: integration of different types of proteomics data into network models. BMB Rep. 2008;41:184–93.

Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: IEEE international conference on neural networks, 1993, vol. 1. Piscataway: IEEE Service Center; 1993. p. 586–91.

Sampson DL, Parker TJ, Upton Z, Hurst CP. A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. PLoS One. 2011;6:e24973.

Sanders WS, Bridges SM, McCarthy FM, Nanduri B, Burgess SC. Prediction of peptides observable by mass spectrometry applied at the experimental set level. BMC Bioinformatics. 2007;8 Suppl 7:S23.

Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. Bioinformatics. 2009;25:2857–9.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

Shipkova P, Drexler DM, Langish R, Smalley J, Salyan ME, Sanders M. Application of ion trap technology to liquid chromatography/mass spectrometry quantitation of large peptides. Rapid Commun Mass Spectrom. 2008;22:1359–66.

Simioniuc A, Campan M, Lionetti V, Marinelli M, Aquaro GD, Cavallini C, Valente S, Di Silvestre D, Cantoni S, Bernini F, Simi C, Pardini S, Mauri P, Neglia D, Ventura C, Pasquinelli G, Recchia FA. Placental stem cells pre-treated with a hyaluronan mixed ester of butyric and retinoic acid to cure infarcted pig hearts: a multimodal study. Cardiovasc Res. 2011;90:546–56.

Simpson KL, Whetton AD, Dive C. Quantitative mass spectrometry-based techniques for clinical use: biomarker identification and quantification. J Chromatogr B Analyt Technol Biomed Life Sci. 2009;877:1240–9.

Sodek KL, Evangelou AI, Ignatchenko A, Agochiya M, Brown TJ, Ringuette MJ, Jurisica I, Kislinger T. Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT). Mol Biosyst. 2008;4:762–73.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.

Suderman M, Hallett M. Tools for visually exploring biological networks. Bioinformatics. 2007;23:2651–9.

Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P. A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics. 2006;22:e481–8.

Vapnik V. The nature of statistical learning theory. New York: Springer; 1999.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P. STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 2007;35:D358–62.

Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem. 2003;75:4818–26.

Webb-Robertson B-JM. Support vector machines for improved peptide identification from tandem mass spectrometry database search. Methods Mol Biol. 2009;492:453–60.

Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, van Erk M, Kleemann R, Haeggström JZ, Goto S. Systems biology approaches and pathway tools for investigating cardiovascular disease. Mol Biosyst. 2009;5:588–602.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res. 2000;28:289–91.

Yang X, Lazar IM. MRM screening/biomarker discovery with linear ion trap MS: a library of human cancer-specific peptides. BMC Cancer. 2009;9:96.

Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng. 2009;11:49–79.

Yu W, Li X, Liu J, Wu B, Williams KR, Zhao H. Multiple peak alignment in sequential data analysis: a scale-space-based approach. IEEE/ACM Trans Comput Biol Bioinform. 2006;3:208–19.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett. 2002;513:135–40.

Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. J Proteome Res. 2006;5:2909–18.

Zhu W, Smith JW, Huang C-M. Mass spectrometry-based label-free quantitative proteomics. J Biomed Biotechnol. 2010;2010:840518.

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res. 2006;5:2339–47.

Supplementary Information References

Ahn HS, Shin YS, Park PJ, Kang KN, Kim Y, Lee H-J, Yang H-K, Kim CW. Serum biomarker panels for the diagnosis of gastric adenocarcinoma. Br J Cancer. 2012;106:733–9.

Alvarez-Buylla A, Culebras E, Picazo JJ. Identification of Acinetobacter species: is Bruker biotyper MALDI-TOF mass spectrometry a good alternative to molecular techniques? Infect Genet Evol. 2012;12:345–9.

Balog CIA, Alexandrov T, Derks RJ, Hensbergen PJ, van Dam GJ, Tukahebwa EM, Kabatereine NB, Thiele H, Vennervald BJ, Mayboroda OA, Deelder AM. The feasibility of MS and advanced data processing for monitoring *Schistosoma mansoni* infection. Proteomics Clin Appl. 2010;4:499–510.

Bergamini G, Di Silvestre D, Mauri P, Cigana C, Bragonzi A, De Palma A, Benazzi L, Döring G, Assael BM, Melotti P, Sorio C. MudPIT analysis of released proteins in Pseudomonas aeruginosa laboratory and clinical strains in relation to pro-inflammatory effects. Integr Biol (Camb). 2012;4:270–9.

Brusniak M-Y, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V, Vitek O, Zhang N, Aebersold R, Watts JD. Corra: computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. BMC Bioinformatics. 2008;9:542.

Brusniak M-YK, Kwok S-T, Christiansen M, Campbell D, Reiter L, Picotti P, Kusebauch U, Ramos H, Deutsch EW, Chen J, Moritz RL, Aebersold R. ATAQS: a computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. BMC Bioinformatics. 2011;12:78.

Camaggi CM, Zavatto E, Gramantieri L, Camaggi V, Strocchi E, Righini R, Merina L, Chieco P, Bolondi L. Serum albumin-bound proteomic signature for early detection and staging of hepatocarcinoma: sample variability and data classification. Clin Chem Lab Med. 2010; 48:1319–26.

Choe L, D'Ascenzo M, Relkin NR, Pappin D, Ross P, Williamson B, Guertin S, Pribil P, Lee KH. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. Proteomics. 2007;7:3651–60.

Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P. ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses. Proteomics. 2011;11: 2654–71.

Dawson J, Walters M, Delles C, Mischak H, Mullen W. Urinary proteomics to support diagnosis of stroke. PLoS One. 2012;7:e35879.

Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, Carapito C, von Haller PD, Tasman N, Mendoza L, Farrah T, Aebersold R. Trans-proteomic pipeline supports and improves analysis of electron transfer dissociation data sets. Proteomics. 2010;10:1190–5.

Djidja M-C, Claude E, Snel MF, Francese S, Scriven P, Carolan V, Clench MR. Novel molecular tumour classification using MALDI-mass spectrometry imaging of tissue micro-array. Anal Bioanal Chem. 2010;397:587–601.

Dyrlund TF, Poulsen ET, Scavenius C, Sanggaard KW, Enghild JJ. MS Data Miner: a web-based software tool to analyze, compare and share mass spectrometry protein identifications. Proteomics. 2012;12(18):2792–6.

Fan Y, Wang J, Yang Y, Liu Q, Fan Y, Yu J, Zheng S, Li M, Wang J. Detection and identification of potential biomarkers of breast cancer. J Cancer Res Clin Oncol. 2010;136:1243–54.

Fan N-J, Gao C-F, Zhao G, Wang X-L, Liu Q-Y. Serum peptidome patterns of breast cancer based on magnetic bead separation and mass spectrometry analysis. Diagn Pathol. 2012;7:45.

Fassbender A, Waelkens E, Verbeeck N, Kyama CM, Bokor A, Vodolazkaia A, Van de Plas R, Meuleman C, Peeraer K, Tomassetti C, Gevaert O, Ojeda F, De Moor B, D'Hooghe T. Proteomics analysis of plasma for early diagnosis of endometriosis. Obstet Gynecol. 2012;119:276–85.

Frenzel J, Gessner C, Sandvoss T, Hammerschmidt S, Schellenberger W, Sack U, Eschrich K, Wirtz H. Outcome prediction in pneumonia induced ALI/ARDS by clinical features and peptide patterns of BALF determined by mass spectrometry. PLoS One. 2011;6:e25544.

Gao B-X, Li M-X, Liu X-J, Cai J-F, Fan X-H, Yang X-L, Li X-M, Li X-W. Analyzing urinary proteome patterns of metabolic syndrome patients with early renal injury by magnet bead separation and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Zhongguo Yi Xue Ke Xue Yuan Xue Bao. 2011;33:511–16.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999;17:994–9.

Han H. Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery. BMC Bioinformatics. 2010;11 Suppl 1:S1.

Ishigami N, Tokuda T, Ikegawa M, Komori M, Kasai T, Kondo T, Matsuyama Y, Nirasawa T, Thiele H, Tashiro K, Nakagawa M. Cerebrospinal fluid proteomic patterns discriminate Parkinson's disease and multiple system atrophy. Mov Disord. 2012;27:851–7.

Izbicka E, Streeper RT, Michalek JE, Louden CL, Diaz 3rd A, Campos DR. Plasma biomarkers distinguish non-small cell lung cancer from asthma and differ in men and women. Cancer Genomics Proteomics. 2012;9:27–35.

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12:507.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.

Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008;24:2534–6.

Kim HU, Sohn SB, Lee SY. Metabolic network modeling and simulation for drug targeting and discovery. Biotechnol J. 2012;7:330–42.

Komori M, Matsuyama Y, Nirasawa T, Thiele H, Becker M, Alexandrov T, Saida T, Tanaka M, Matsuo H, Tomimoto H, Takahashi R, Tashiro K, Ikegawa M, Kondo T. Proteomic pattern analysis discriminates among multiple sclerosis-related disorders. Ann Neurol. 2012;71: 614–23.

Krause B, Seifert S, Panne U, Kneipp J, Weidner SM. Matrix-assisted laser desorption/ionization mass spectrometric investigation of pollen and their classification by multivariate statistics. Rapid Commun Mass Spectrom. 2012;26:1032–8.

Kyama CM, Mihalyi A, Gevaert O, Waelkens E, Simsa P, Van de Plas R, Meuleman C, De Moor B, D'Hooghe TM. Evaluation of endometrial biomarkers for semi-invasive diagnosis of endometriosis. Fertil Steril. 2011;95:1338–43.e1–3.

Lasch P, Drevinek M, Nattermann H, Grunow R, Stämmler M, Dieckmann R, Schwecke T, Naumann D. Characterization of Yersinia using MALDI-TOF mass spectrometry and chemometrics. Anal Chem. 2010;82:8464–75.

Lazova R, Seeley EH, Keenan M, Gueorguieva R, Caprioli RM. Imaging mass spectrometry – a new and promising method to differentiate Spitz nevi from Spitzoid malignant melanomas. Am J Dermatopathol. 2012;34:82–90.

Le Faouder J, Laouirem S, Chapelle M, Albuquerque M, Belghiti J, Degos F, Paradis V, Camadro J-M, Bedossa P. Imaging mass spectrometry provides fingerprints for distinguishing hepatocellular carcinoma from cirrhosis. J Proteome Res. 2011;10:3755–65.

Liao CCL, Ward N, Marsh S, Arulampalam T, Norton JD. Mass spectrometry protein expression profiles in colorectal cancer tissue associated with clinico-pathological features of disease. BMC Cancer. 2010;10:410.

Lin Q, Peng Q, Yao F, Pan X-F, Xiong L-W, Wang Y, Geng J-F, Feng J-X, Han B-H, Bao G-L, Yang Y, Wang X, Jin L, Guo W, Wang J-C. A classification method based on principal components of SELDI spectra to diagnose of lung adenocarcinoma. PLoS One. 2012;7:e34457.

Liu Q, Chen X, Hu C, Zhang R, Yue J, Wu G, Li X, Wu Y, Wen F. Serum protein profiling of smear-positive and smear-negative pulmonary tuberculosis using SELDI-TOF mass spectrometry. Lung. 2010;188:15–23.

M'Koma AE, Seeley EH, Washington MK, Schwartz DA, Muldoon RL, Herline AJ, Wise PE, Caprioli RM. Proteomic profiling of mucosal and submucosal colonic tissues yields protein signatures that differentiate the inflammatory colitides. Inflamm Bowel Dis. 2011;17:875–83.

Mauri P, Dehò G. A proteomic approach to the analysis of RNA degradosome composition in Escherichia coli. Methods Enzymol. 2008;447:99–117.

Mauri P, Scarpa A, Nascimbeni AC, Benazzi L, Parmagnani E, Mafficini A, Peruta MD, Bassi C, Miyazaki K, Sorio C. Identification of proteins released by pancreatic cancer cells by multidimensional protein identification technology: a strategy for identification of novel cancer markers. FASEB J. 2005;19:1125–7.

Meding S, Nitsche U, Balluff B, Elsner M, Rauser S, Schöne C, Nipp M, Maak M, Feith M, Ebert MP, Friess H, Langer R, Höfler H, Zitzelsberger H, Rosenberg R, Walch A. Tumor classification of six common cancer types based on proteomic profiling by MALDI imaging. J Proteome Res. 2012;11:1996–2003.

Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci U S A. 1999;96:6591–6.

Ong S-E, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005;1:252–62.

Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1:376–86.

Parikh JR, Askenazi M, Ficarro SB, Cashorali T, Webber JT, Blank NC, Zhang Y, Marto JA. Multiplierz: an extensible API based desktop environment for proteomics data analysis. BMC Bioinformatics. 2009;10:364.

Pecks U, Schütt A, Röwer C, Reimer T, Schmidt M, Preschany S, Stepan H, Rath W, Glocker MO. A mass spectrometric multicenter study supports classification of preeclampsia as heterogeneous disorder. Hypertens Pregnancy. 2012;31:278–91.

Polpitiya AD, Qian W-J, Jaitly N, Petyuk VA, Adkins JN, Camp 2nd DG, Anderson GA, Smith RD. DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics. 2008;24:1556–8.

Qian W-J, Petritis BO, Nicora CD, Smith RD. Trypsin-catalyzed oxygen-18 labeling for quantitative proteomics. Methods Mol Biol. 2011;753:43–54.

Regonesi ME, Del Favero M, Basilico F, Briani F, Benazzi L, Tortora P, Mauri P, Dehò G. Analysis of the Escherichia coli RNA degradosome composition by a proteomic approach. Biochimie. 2006;88:151–61.

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004;3:1154–69.

Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics. 2005;5:4–15.

Simioniuc A, Campan M, Lionetti V, Marinelli M, Aquaro GD, Cavallini C, Valente S, Di Silvestre D, Cantoni S, Bernini F, Simi C, Pardini S, Mauri P, Neglia D, Ventura C, Pasquinelli G, Recchia FA. Placental stem cells pre-treated with a hyaluronan mixed ester of butyric and retinoic acid to cure infarcted pig hearts: a multimodal study. Cardiovasc Res. 2011;90: 546–56.

Song Z, Dong R, Fan Y, Zheng S. Identification of serum protein biomarkers in biliary atresia by mass spectrometry and ELISA**.** J Pediatr Gastroenterol Nutr. 2012;55(4):370–5.

Specht M, Kuhlgert S, Fufezan C, Hippler M. Proteomics to go: proteomatic enables the user-friendly creation of versatile MS/MS data evaluation workflows. Bioinformatics. 2011;27: 1183–4.

Sui W, Dai Y, Zhang Y, Chen J, Liu H, Huang H. Proteomic profiling of uremia in serum using magnetic bead-based sample fractionation and MALDI-TOF MS. Ren Fail. 2010;32:1153–9.

Tang K-L, Li T-H, Xiong W-W, Chen K. Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data. BMC Bioinformatics. 2010;11:109.

Thompson D, Pepys MB, Wood SP. The physiological structure of human C-reactive protein and its complex with phosphocholine. Structure. 1999;7:169–77.

Timms JF, Menon U, Devetyarov D, Tiss A, Camuzeaux S, McCurrie K, Nouretdinov I, Burford B, Smith C, Gentry-Maharaj A, Hallett R, Ford J, Luo Z, Vovk V, Gammerman A, Cramer R, Jacobs I. Early detection of ovarian cancer in samples pre-diagnosis using CA125 and MALDI-MS peaks. Cancer Genomics Proteomics. 2011;8:289–305.

Trudgian DC, Thomas B, McGowan SJ, Kessler BM, Salek M, Acuto O. CPFP: a central proteomics facilities pipeline. Bioinformatics. 2010;26:1131–2.

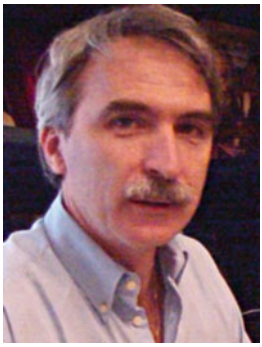
Van Gorp T, Cadron I, Daemen A, De Moor B, Waelkens E, Vergote I. Proteomic biomarkers predicting lymph node involvement in serum of cervical cancer patients. Limitations of SELDI-TOF MS. Proteome Sci. 2012;10:41.

Waloszczyk P, Janus T, Alchimowicz J, Grodzki T, Borowiak K. Proteomic patterns analysis with multivariate calculations as a promising tool for prompt differentiation of early stage lung tissue with cancer and unchanged tissue material. Diagn Pathol. 2011;6:22.

Wang L, Zheng W, Ding X, Yu J, Jiang W, Zhang S. Identification biomarkers of eutopic endometrium in endometriosis using artificial neural networks and protein fingerprinting. Fertil Steril. 2010;93:2460–2.

Wong JWH, Schwahn AB, Downard KM. FluTyper-an algorithm for automated typing and subtyping of the influenza virus from high resolution mass spectral data. BMC Bioinformatics. 2010;11:266.

Zhu Y-W, Wang Y-D, Ye Z-Y, Hu X, Yu J-K. Application of serum protein fingerprint in diagnosis of pancreatic cancer. Zhejiang Da Xue Xue Bao Yi Xue Ban. 2012;41:289–97.



Pier Luigi Mauri, Principal Investigator, Italy
Dr. Mauri has a notable experience in the development and application of several technologies (HPLC, capillary electrophoresis, mass spectrometry) for the identification and quantification of several biomolecules in complex matrices. He has been involved in the study of the bioavailability of several drugs in humans and animals.

He has set up a proteomics laboratory based on both traditional proteomic technologies (2D gel and off-line identification LC-MS/MS) and innovative approaches. In particular, he is among the first Italian researchers to use the MudPIT methodology (2DC-MS/MS) and a parallel computing system for proteomics. The latter instrumentation allows for the proteomic profiling of biological samples, without limitations, according to protein molecular weight, pI, or hydrophobicity, and guarantees high productivity. Using this methodology, secreted proteins from tumor (*FASEB* 2005) and immune system cells have been characterized. Furthermore, direct analysis of biological samples (*JMS* 2007), enzymatic complexes (*Biochimie* 2006), development of novel methodological approaches for studying switch-redox proteins (*JBC* 2005 and 2006, *JMB* 2006), and structural and functional characterization of proteins (*Biol. Chem.* 2004, *FEBS* 2006) have been performed.

In addition, he has developed analytical methods for quali-quantitative characterization of natural bioactive compounds (such as polyphenols and terpenes from Ginkgo biloba and other medicinal plants), and bioavailability studies in human and animals have been performed. He is coauthor of more than 100 scientific publications; he has participated in numerous international congresses and has taught in several metabolomic and proteomic courses.

Chapter 10
Bioinformatics Approach for Finding Target Protein in Infectious Disease

Hemant Ritturaj Kushwaha and Indira Ghosh

Abstract With the technological advancements, biological data pertaining to various infectious organisms is getting abundant. This copiousness has been useful for developing abysmal understanding of the complex biological process which leads to the diseased condition in human race. Existing mode of treatments for such infectious diseases currently faces various challenges such as drug resistance. Thus, identification of new drug targets has become one of the major objectives of the scientific community involved in drug designing. These novel drug targets can provide effective know-how of the infectious organisms in order to develop novel therapeutic agents in order to contain the spread of the disease. Systems biology approach has been considered as one of the promising approach that can effectively lead to novel drug target identification. It provides the conceptual framework for the analysis using the amalgamation of variety of data obtained from conglomeration of advanced molecular biology techniques. In this chapter, we have elaborated the systems biology approaches which can be used for identification of novel drug targets for various infectious diseases. Apart from emphasizing systems biology leads in the area of drug target identification, we have highlighted some *in silico* experiments performed using these techniques for the identification of novel drug targets in infectious organisms such as *P. falciparum* and *M. tuberculosis*. This chapter might help in devising the effective systems biology strategies in order to develop hypothesis toward identification of novel drug targets.

Keywords Drug targets • Systems biology • Conceptual framework • Amalgamation • Infectious diseases • In silico experiments

H.R. Kushwaha, Ph.D.
Synthetic Biology and Biofuel Group, International Center for Genetic Engineering and Biotechnology (ICGEB), New Delhi, India

I. Ghosh, Ph.D. (✉)
School of Computational and Integrative Sciences,
Jawaharlal Nehru University, New Delhi 110067, India
e-mail: ighdna@yahoo.com

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_10,

10.1 Introduction and Overview of Target-Based Drug Discovery

For several years, infectious diseases have been a substantial global threat to the human civilization. According to WHO reports, in developing countries, one of the leading causes of morbidity and mortality in humans is infectious diseases such as tuberculosis (WHO 1999). The situation becomes worrisome with the recognition of new infectious diseases and resistance of infectious agents against the existing drugs. Several factors have been considered as responsible to the emerging problem of infectious diseases like changes in demography, environmental changes, and various technological advancements which are leading to problems of pollution, thus weakening the immune response. Thus, there is a serious need for quick and efficient therapeutic agents which can control the menace of infectious disease.

The process of drug discovery is considered as one of the longest and costliest efforts. The classical approach to drug discovery process is the identification of drug target, then designing the lead compound, and finally to the validation of the compounds through clinical trials. It was considered that on an average one new drug takes approximately 12–13 years to reach to a patient from a research lab. Millions have been spent to find some new drug candidates for the particular disease, but the success rate is very low: the reason is the limited knowledge of complex biological systems. Therefore, it was perceived that until in-depth knowledge of complex biological process that leads to the diseased state is considered, the discovery of new drug will be a time-consuming and challenging process.

Methodological advancements in molecular biology have led to the prodigious rise in the availability of high-resolution biological data. Advance technologies such as "*high-throughput*" generation of data have taken center stage in both academia and industry, in order to unveil the novel pathways and develop understanding of the biological systems. New and robust sequencing technologies have led to the growth of high-resolution data pertaining to the genome of various organisms. Therefore, twenty-first century is considered as the era of "*omics*," i.e., genomics, study of genomes; proteomics, large-scale study of proteins; and metabolomics, analysis of biochemical processes involving metabolites. It is considered that careful interpretation and integration of such data can yield novel insights and conclusions about the crucial biological processes. Systems biology is one rapid advancing interdisciplinary subject that combines data obtained from the experimental means along with the computational and mathematical techniques. Using the concoction of such disciplines, it is assumed that it can develop crucial understanding of the complex biological and physiological phenomenon. It produces the detailed route map of the subcellular networks, thus creating an opportunistic window for the smarter therapeutic strategies.

Drug target identification is considered as one of the most crucial and strenuous processes in drug designing. Effective drug target identification not only assists in improving the efficacy of the drug but also helps in avoiding the potential side effects; many failures could be avoided in early phase (Drews 2000). In order for the

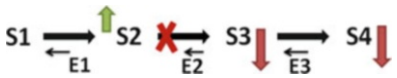


Fig. 10.1 A schematic diagram of a series of reaction in cellular system. The *filled arrows* show the increase (*green*) and decrease (*red*) of metabolites due to selective inhibition of enzyme *E2* represented by a *cross*

drug target to be effective, it must be essential for its growth, replication, and survival of the microorganism (Sakharkar et al. 2004). According to the pharmacological definition, drug targets can either be inhibited or activated by the drug molecule by the binding of the drug molecule (Jiang and Zhou 2005). The sequencing of various genomes including those organisms which are involved in various infectious diseases has paved the way to identify novel drug targets from the analysis of gene networks (Whittaker 2003; Peterson and Ringner 2003). In case of signal transduction-related drug targets, network-based drug target identification approach has been in spotlight in recent years (Farkas et al. 2011). With the availability of advance methods of data generation, it is comparatively apparent to validate the model generated for a process with the experimental results. Thus, the interconnectivity between different levels of understanding becomes critical to analyze the data and build the novel predictions. But it is observed that the relationship between the data generated from various platforms is nonlinear and dynamic which makes the process of model building a highly challenging effort (Stark et al. 2003). More or less no correlation between the protein expressed and kinetics of the enzymes in cellular system as compared to the single enzyme assay in test tube makes the prediction of target using simple comparative genomics and proteomics rather futile (Eisenthal and Cornish-Bowden 1998).

It is assumed that the cell essentially works as a factory wherein all the enzymes in the cell or organelle work in concert. The orchestrated network of biochemical reactions makes the system complicated *in vivo* compared to a simple one enzyme (with infinite dilution) and one substrate system *in vitro*; this yields to a situation where the results of inhibition data from test tube cannot be linearly translated to the cellular system (Westley and Westley 1996). This can be explained using the simple hypothetical pathway. In Fig. 10.1, S1, S2, S3, and S4 are substrates/products of the reactions of E1, E2, and E3 enzymes catalyzing those reactions. The *in vitro* inhibition of the enzyme E2 will decrease the production of the product S3 and S4, whereas *in vivo* inhibition might not produce the desired reduction (Singh 2009). On the other hand the accumulation of the substrate S2 would possibly (if toxic for the cell) favor the inhibition of the enzyme *in vivo* to kill the cell. Such phenomenon was observed in case of *in vivo* inhibition of orotidine-5′-phosphate (OMP) decarboxylase (catalyzes the reaction: $\mathrm{OMP} \rightarrow \mathrm{CO_2} + \mathrm{UMP}$) which showed a depletion of all intermediates between, and including UMP and dCDP, and accumulation of OMP. But, as the simulation proceeded, OMP reached to a new steady-state concentration, whereas the concentrations of the depleted intermediates rose to their original levels (Duggleby 1998; Duggleby and Christopherson 1984). Thus, the inhibition was counteracted due to the accumulation of the substrate of the inhibited enzyme, and

its effect was nullified consequently. Thus, it was concluded that the *in vitro* assays which are used for preliminary analysis in recent drug discovery process may not portray the correct picture until complemented by cell-based assays. Therefore, the phenotypic and genotypic relationship between the target genes is in need to be validated before choosing them as drug targets (Sams-Dodd 2005). In 2004, Butcher et al. have elaborated that the informatics integration of "*omics*" data sets (a bottom-up approach); computer modeling of pathogen, disease, or organ system physiology (a top-down approach to target selection, clinical indication, and clinical trial design); and the use of complex cell systems themselves will interpret and predict the biological activities of drugs and gene targets (Butcher et al. 2004).

Hence, the systems biology approach will be able to identify the critical pathways of gene products and their time-dependent effective relationships, which plays crucial role in pathogenesis (Chua and Roth 2011). This chapter provided an overview of the *in silico* systems biology techniques that may assist identification of novel drug targets in various pathogenic bacteria. This overview provides an open vista for improvement in various approach adapted till now.

10.2 Infectious Disease Database

Advancements in molecular biology techniques and high-throughput technology have led to the accumulation of large volume of biological data. This agglomerated data was required to be accessed and analyzed in order to answer those questions which have been a mystery to the scientific world for long. In order to solve conundrum related to infectious diseases, bioinformatics has emerged as a new discipline which integrates many core subject areas such as mathematics, statistics, physics, chemistry, computer science, and cell and molecular biology. Revolutions in information technology such as databases have offered more precise and effective storage of biology data, which can be analyzed and accessed simultaneously. Earlier, databases were used to store raw data produced by experimental methods, but the advent of the concept of curated database known as meta-database has made the integrated large volume of data into well-knitted biological information.

Large numbers of infectious diseases continue to affect human civilization. A study revealed that approx 15 million (>25%) of 57 million annual deaths are estimated due to infectious diseases (Morens et al. 2004). Percentages of deaths due to infectious diseases worldwide are depicted as pneumonia ~30%, HIV/AIDS 20%, diarrheal diseases ~17%, tuberculosis ~12% , and malaria ~8% (Fauci 2001). Several initiatives have been taken in order to organize the data pertaining to infectious disease; some of these databases are listed in Table 10.1.

GIDEON is the medical database owned by GIDEON Informatics, which holds information of the diagnosis and reference tool for a number of infectious diseases. Various modules of the database also present the taxonomic identification based on the various phenotypic characteristics on the microbes causing infectious diseases. Biomarkers have been considered as one of the important tools in the analysis of

Table 10.1 List of infectious disease databases existing in the public domain

Name	Type of database	Type of information	URL
GIDEON	Licensed	Reference in the fields of tropical and infectious diseases, epidemiology, microbiology, and antimicrobial chemotherapy	http://www.gideononline.com/
Infectious Disease Biomarker Database (IDBD)	Free	Biomarker information pertaining to infectious diseases	http://biomarker.cdc.go.kr/biomarker/index.jsp
KEGG DISEASE Database	Licensed	Collection of disease entries capturing knowledge on genetic and environmental perturbations	http://www.genome.jp/kegg/disease/
GENI-DB	Free	News events on the topic of over 176 infectious diseases	http://born.nii.ac.jp/

overall progression of the infectious disease. It can be sensitively measured in the human body. Recently a number of initiatives have been taken to identify and develop selective biomarkers that may assist in early prognosis and progression of any infectious disease. These biomarkers also help in determining the essential treatments of the suffering patients. Infectious Disease Biomarker Database (IDBD) is the freely available relational database which provides information on various biomarkers and their potential roles in the early infectious processes (Yang et al. 2008). Also, the database houses various bioinformatics tools that can be used for analysis and visualization of the biomarker data. KEGG DISEASE is one of the databases under the KEGG suite of database which accumulates molecular-level information on the infectious disease along with the genes and genome information of the causing organism. It also provides essential details on the drugs administered and various known biomarkers for the infectious diseases (Kanehisa et al. 2008, 2010). GENI-DB is one unique initiative which covers the various new events regarding 176 infectious diseases affecting human health (Collier and Doan 2012). These events are essentially classified from global news media in ten languages. The database is freely available and is helpful in monitoring the global trends in the spread of various infectious diseases. Apart from being the information resource, it also helps various governments in preparing essential programs in order to check the infectious disease epidemic.

10.3 Drug Target Database

In order to effectively contain the spread of infectious diseases, drug target identification is considered to be an essential step. Over the years, attempts have been made to discover novel drugs that bind to the specific proteins, thereby changing their biochemical and/or biophysical activities which could hamper its biological function.

Table 10.2 List of various drug target databases

Database	URL
TDR Targets Database (TDR targets)	http://tdrtargets.org/
Potential Drug Target Database (PDTD)	http://www.dddc.ac.cn/pdtd/
TB Drug Target Database	http://www.bioinformatics.org/tbdtdb/
DrugBank	http://www.drugbank.ca/
ddTargets	http://www.sciclips.com/sciclips/drug-targets-all.do
Genomic Target Database (GTD)	http://iioab-dgd.webs.com/
SuperTarget	http://insilico.charite.de/supertarget

After several years of research on various infectious diseases, a number of druggable targets have been identified and exploited. The information regarding these targets was kept in various databases in order to provide an easy access to the research community to explore these targets which may help in the development of novel compounds against these targets. Some of the drug target databases are listed in Table 10.2.

The TDR Targets Database is composed of the genomic and bioinformatics data related to organism related to various infectious diseases. The information in the database is essentially extracted both automatically and manually from literature and other databases for each putative drug target (Agüero et al. 2008). Over the decades, molecular biology techniques and high-throughput approaches have elaborated our understanding regarding various infectious diseases. The TDR Target Database largely brings together data and annotation emerging from genome sequencing and functional genomics projects, protein structural data, manual curation of inhibitors and targets, and information on target essentiality and druggability (Agüero et al. 2008). Another database known as Potential Drug Target Database (PDTD) is the database which is known to include information regarding various drug targets along with the knowledge base of potential binding proteins (Gao et al. 2008). The advantage of such approach is that it allows the reverse docking approach in the form of a web server, namely, TarFisDock (Target Fishing Dock), to identify the novel drug targets using the known ligands (Li et al. 2006). The database specifically designed for the drug targets related to the tuberculosis is TB Drug Target Database. It includes the details on the drug targets along with the probable inhibitors of the various target proteins in *Mycobacterium tuberculosis*. The database known as DrugBank also combines the drug information with the comprehensive drug target information, thus allowing effective drug target discovery (Knox et al. 2011). ddTarget is considered as a unique database that provides details of those targets which have been reported in the US patents and other research articles. Genome sequencing of various infectious microbes has given a new vista to drug target approach. The gamut of available data can be used for finding putative genomic drug targets. Genomic Target Database (GTD) is one such initiative in this regard as it provides drug target information based on pathogen-specific unique metabolic pathways, host-pathogen common metabolic pathways, and membrane/surface localized for the pathogenic bacteria (Barh et al. 2009). Recently, Hecker

et al. (2012) has made another attempt to identify novel drug targets in infectious disease. The database known as SuperTarget integrates drug-related information associated with medical indications, adverse drug effects, drug metabolism, pathways, and Gene Ontology (GO) terms for target proteins (Hecker et al. 2012).

10.4 *In Silico* Approach to Target Discovery

Many research using comparative genomics and proteomics has attempted to identify the target protein present in pathogen and absent in host (Starck et al. 2004). The number of bioinformatics toolbox and pipelines such as MBGD Microbial Genome Database for Comparative Analysis (http://mbgd.genome.ad.jp/) has been devised for comparing the protein sequences of the pathogen with that of the host, in order to identify the exclusivity in the protein function (Uchiyama et al. 2010; Himmelreich et al. 1997). Another method which is based on comparing the pathway enzymes of different species and finding their similarity has been used in *M. tuberculosis* to identify essential proteins (Anishetty et al. 2005). The comparative analysis of genomic sequences of a large set of genomes has provided a set of essential genes database for choosing the drug target protein (Zhang et al. 2004). It is observed that proteins essentially function as the part of highly interconnected network referred to as interactome network (Barabasi and Oltvai 2004; Han et al. 2004; Jeong et al. 2001; Vidal 2005). Therefore, the extensive cellular network of proteins can be exploited further to identify novel drug targets against various infectious diseases. Hence, it is necessary to understand the framework of disease and cellular networks (Yildirim et al. 2007) and develop *in silico* approaches and methods in order to identify probable drug target in infectious organisms. Some of these approaches have been discussed briefly in this chapter.

10.4.1 Metabolic Reconstruction

Post-genomic era has been the era of looking at the systems approach of various biological processes, i.e., a view beyond sequences. The number of attempts has been made to construct gene regulatory network or protein interaction network, in order to understand the system from wide perspective which has enabled the identification of various probable drug targets. Metabolic reconstruction is the method of labeling genes in the context of the metabolic pathways. Application and expansion of sophisticated molecular biology technique can provide insight into a possible functional context, but it is also felt that the complete description of the process cannot be understood until the information is projected over relevant metabolic pathways. Since metabolic pathway reconstruction details minute information of any given pathway, therefore, it helps in performing the comparative analysis across various species, which finally leads to the novel drug target identification for

Table 10.3 List of database used to store and analyze various metabolic pathways in various organisms

Database	URL
KEGG	http://www.genome.jp/kegg/pathway.html
BRENDA	http://www.brenda-enzymes.org/
Biochemical Pathway Maps	http://web.expasy.org/pathways/
PathCase	http://nashua.case.edu/PathwaysWeb/
BioCyc	http://biocyc.org/

various infectious diseases. In order to facilitate the metabolic pathway reconstruction, several pathway databases have been accomplished.

Metabolic reconstruction is essentially done in two stages: the first stage leads to the development of the draft reconstruction, while the other stage which is essentially manual performs the refinement of the draft reconstruction. The draft metabolic reconstruction is based on the genome annotation and the biochemical database information of the organism. The genomic annotations and metabolic pathway information can be obtained from the various databases mentioned in Table 10.3. Various software tools such as Pathway Tools and metaSHARK have been proposed for the automated reconstruction procedure (Karp et al. 2002; Pinney et al. 2005). The draft reconstruction obtained by automated means is then subjected to the manual reevaluation and thus refined further. It has been emphasized that the manual reconstruction is important, considering the fact that all the annotation may not have high level of confidence and also various pathway databases are generally organism specific. Therefore, it becomes essential to verify the existence of the respective pathway with the target organism (Thiele and Palsson 2010). Further, the predicted pathway network is verified using various existing experimental data and literature resources. After the manual curation, the final metabolic model is checked for its consistency in terms of physical and chemical nature, dead ends present in the pathways, production of essential metabolites, etc. Constraint-based metabolic model which is based on the reaction stoichiometries is used for verifying the mass balance in the metabolism at the steady states (Durot et al. 2009).

10.4.1.1 Pathway Databases for Pathogenic Organisms

Analysis of pathways has always been the center point of the system-level analysis of various organisms including those involved in various infectious diseases in humans. Over the two decades, "*omics*" approach has been able to correlate the aspects of genes, proteins, and metabolites. Thus, the notion of connecting these building blocks in the form of pathway diagrams or pathway maps has reformed the basic understanding of the biological system. In the past few decades, initiatives have been taken to analyze these pathway data along with its effective storage (Table 10.3).

KEGG (Kyoto Encyclopedia of Genes and Genomes) is one of the comprehensive knowledge bases, which provides the systematic analysis of pathways with the genomic reference in various organisms, presented in graphical representation (Kanehisa and Goto 2000). BRENDA (BRaunschweig ENzyme Database) is an archive that has been reported to have a collection of 3,000 different enzymes from 9,800 different organisms along with their pathways (Schomburg et al. 2004). Another repository known as BioCyc database is the agglomeration of more than 350 organism-specific Pathway along with various pathway analysis tools (Caspi et al. 2008). One of the unique initiatives by PathCase Systems Biology (PathCase-SB) group is the PathCase database which is known to provide metabolic pathway data and systems biology models in the form of one comprehensive database (Cakmak et al. 2011). With an expected increase in experimental data in the future, it is expected that these types of databases and tools will definitely be more precise and comprehensive for the analysis of disease-causing organisms.

10.4.1.2 Comparative Approach to Identify New Drug Targets

The availability of a large number of diverse genome sequences of various organisms has laid the groundwork for better and accurate identification of various drug targets. This is essentially done by the identification of functions of genes and protein from one species to another on the basis of homology. The comparative analysis assists in providing valuable insight to the functional context needed for the detailed reconstruction of the metabolic pathways. One of the earlier attempts has shown the comparative view of identifying potentially important genes in *Mycobacterium* sp. (Cole 2002). Comparative approaches have also identified putative drug targets based on their conservation among *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* species (Payne et al. 2007). Earlier, Lee et al. (2009) concluded the relevance of metabolic reconstruction and *in silico* analyses while identifying potential antibiotic target in *Staphylococcus aureus*. Among other drug targets, polypeptide deformylase (McDevitt and Rosenberg 2001; Brötz-Oesterhelt and Sass 2010; Zhang et al. 2010), aminoacyl-tRNA synthetases (McDevitt and Rosenberg 2001; Gutierrez-Lugo and Bewley 2011), and components of the NAD(P) biosynthetic pathway (Bi et al. 2011) have been identified using comparative genomics approach. Thus, from the present scenario, it is concluded that metabolic reconstruction and *in silico* analyses of various bacterial species can provide a considerable success in potential antibiotic target identification.

10.4.1.3 Kinetic Modeling and Metabolic Control Analysis (MCA)

Essential kinetics of biochemical pathways in steady state are explained and analyzed by Michaelis-Menten kinetics. Considering the reaction model as shown below

$$E + S \rightarrow ES \text{ and } ES \rightarrow E + P$$

the first reaction is assumed to be very fast with the consequence that it would be very close to equilibrium because k_{+1} (rate constant for forward reaction) and k_{-1} (rate constant for reverse reaction) are very high compared to k_{+2} (second reaction); Michaelis and Menten explains why biochemical reactions have a limiting rate of reaction V_f. It is due to the finite amount of enzyme molecules (ET) and the rate at which the enzyme-substrate complex is transformed into enzyme and product (k_p, the turnover number). The other parameter defined by Michaelis constant (K_M) is the dissociation constant of enzyme-substrate complex and so a measure of the affinity of the enzyme for the substrate (or vice versa). High values of K_M correspond to low affinity for the substrate. K_M also happens to be the concentration of substrate for which the initial rate of reaction is half limiting, $V_f/2$. The rates of enzymatic reactions are affected by substances other than the enzyme, the substrates, or the products. These other substances are generically called "*modifiers*" and can often be further classified into "*activators*" or "*inhibitors*," if they increase or decrease the initial rate of reaction for fixed concentrations of all substrates and products, respectively, all in steady state.

However, kinetic modeling offers a great way to capture the dynamics of a biochemical reaction network in mathematical form so as to analyze and simulate its behaviors and ultimately to use the model to answer real physiological questions. Nonlinear ordinary differential equations (ODEs) are by far the most common way to represent the dynamical properties of biochemical reaction networks.

Realizing a reaction network as a system of ODEs is based on two assumptions. First is that the system is "*well stirred*" so that component concentrations (metabolites like substrates and products) do not vary with respect to space. The second basic assumption is that the variables (metabolite concentrations) are continuous functions of time. Given these two simplifying assumptions, ODEs can be effectively used to express mathematically the dynamical consequences of a biochemical network. To get from a reaction network to a set of ODEs, the network should be thought of as a dynamical system whose state is changing from one moment of time to the next. To each metabolite of the network, a single state variable, $X(t)$ is assigned. The collection of values of all these variables ($X_1(t)$, $X_2(t)$, $X_3(t)$,…,$X_n(t)$) at any point in time constitutes the state of the system. Then, for each metabolite, a differential equation is written that describes how its concentration changes over time due to its interactions with other metabolites in the network. The rate of each reaction must be represented by a kinetic rate equation, which would have one or more rate constants associated with it. The set of all rate constants needed to describe the reactions in a molecular interaction network is called the parameter set (p_1, p_2, p_3,…,p_m) of the model (Kanehisa and Goto 2000; Schomburg et al. 2004; Caspi et al. 2008). Biochemical systems are often found in steady states, or at least in quasi-steady states; if not in the wild, at least in laboratory conditions (e.g., cultures of microorganisms in chemostat). From the theoretical point of view, steady states are easier to study because the balance equations for each of the metabolites vanish (at steady state) and so the system can be described by a set of nonlinear algebraic equations. The solution of these equations does not require any integration (and is therefore simpler than in the case of ordinary points). As far as judged by the

experimental precision available, steady states are very frequent in biochemical systems (Kacser and Burns 1973; Heinrich and Rapoport 1974; Fell 1992) and are seen as a property fundamental to life by many.

Metabolic control analysis (MCA) is a phenomenological quantitative first-order sensitivity analysis of variables of the system (usually fluxes and metabolite concentrations) in the vicinity of a stable and structurally stable fixed point (Kacser and Burns 1973; Heinrich and Rapoport 1974). The method essentially based on the kinetics of enzymes, explained by Michaelis-Menton kinetics, in order to parameterize the control coefficients resulting in response coefficients. This control is measured by applying a perturbation on the step being studied and measuring the effect on the variable of interest. The most common control coefficients are those for fluxes and metabolite concentrations. From a systems perspective, MCA relates local behavior, the behavior of a single step considered in isolation, to global behavior, the behavior of the step in the context of a system. MCA was developed in Edinburgh in 1973 by Henrik Kacser and Jim Burns (1973) and independently in Berlin by Reinhart Heinrich and Tom Rapoport (1974). The applicability of the MCA is not only limited to the linear pathways but it was believed that it can also be applied to the branching and cyclic pathways and other enzyme cascades. MCA does not require all the system components such as metabolites or enzymes to be characterized, but the control coefficients can be estimated for different components of the network and for pertinent environmental factors (Fell 1992; Cornish-Bowden and Cárdenas 2000; Cascante et al. 2002). The approach thus investigates the sensitivity of steady-state properties of the network to the small parameter changes (Klipp et al. 2005; LaPorte et al. 1984). Kinetic modeling provides a way for such testing and can be employed to study the dynamical behavior of the pathway without or with modulation in enzyme(s) activity. Application of kinetic modeling to the identification of potential drug targets is exemplified by glycolysis in *Trypanosoma brucei* (Eisenthal and Cornish-Bowden 1998). Kinetic models provide insight into the cellular effect(s) of inhibition of any enzyme in a biochemical pathway.

10.4.2 Flux Balance Analysis

Flux balance analysis (FBA) is one of the widely used approaches in biological systems to understand the complex biological networks. It involves the steady-state analysis of stoichiometric matrix of various pathways (Kauffman et al. 2003). It calculates the flow of the metabolites through the metabolic networks, thus assisting in predicting the overall growth of an organism (Orth et al. 2010). In short, it basically captures the snapshot of the flow of all the metabolites through the various metabolic pathways. Large-scale genome-sequencing initiatives have led to the analysis of the various metabolic reactions annotated with the respective gene information, which has enabled the systemic analysis of various complex and unknown processes (Duarte et al. 2007; Feist et al. 2007; Feist and Palsson 2008; Reed et al. 2003; Oberhardt et al. 2009). An exhaustive set of balanced chemical equations is used

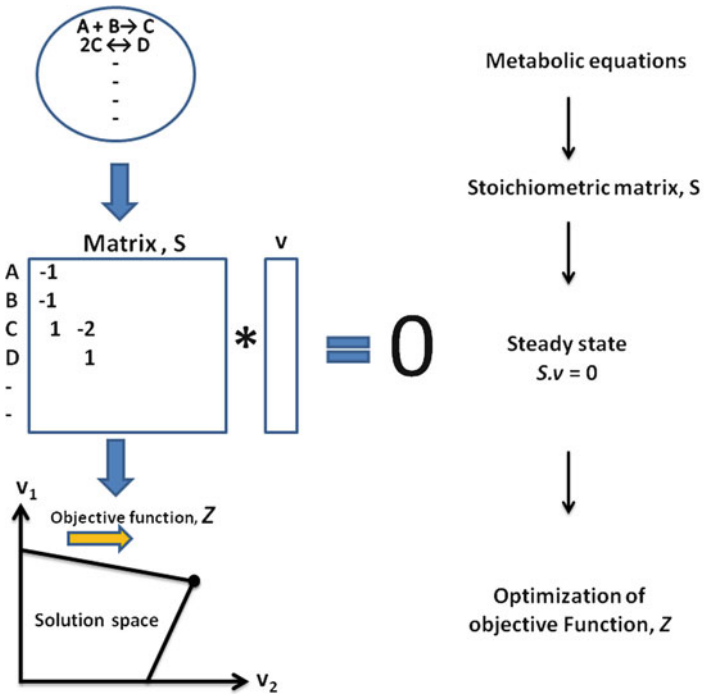


Fig. 10.2 Brief overview of flux balance analysis (*FBA*)

after the genome-scale annotation to frame a mathematical model by forming a matrix as each row represents a metabolite and each column represents a reaction. The overall growth of an organism is thus measured in terms of metabolites consumed during biomass production. The robustness of the method also lies in the fact that it can also incorporate the exchange reactions that are used to calculate the flow of metabolites, such as glucose and oxygen, in and out of the cell. The overall biomass is calculated in terms of a mathematical function known as objective function which is calculated from optimization of the model (Fig. 10.2). Thus, the maximization of the objective function is considered as the parameter of the increased biomass production with the system.

10.4.3 In Silico Knockout Analysis

Flux balance analysis shed light on the major choke points in a metabolic pathway; it is often considered that these choke points may not be an essential reaction in the network. This can readily be tested using *in silico* knockout of the gene or the reaction from the network (Fig. 10.3). The approach is based on the flux balance approach along with the graph-based analysis. In this approach the selected gene is

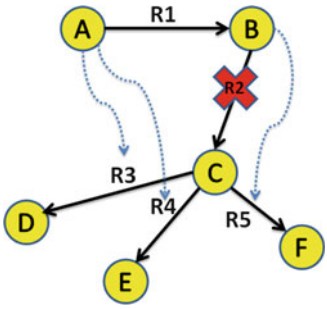


Fig. 10.3 An overview of *in silico* knockout approach. The specific reaction or related gene is knocked out and it is analyzed that the reaction is essential. The *dashed line* (*blue*) shows if the alternate path to the reaction exists in the network

deleted from the network, and using breadth-first search algorithm, it is tested if the neighboring compounds of the knockout can be produced by other reaction and pathways of the reactions. The genes obtained from the analysis can give an overview of the relevance of a specific reaction or gene within the system. A whole-genome metabolic model to compare between natural and persistent *M. tuberculosis* was used to perform *in silico* knockouts by FBA (Singh 2009), and the essential gene set has been prioritized here for guiding experiments. Some of these genes were already reported as potential drug targets; the validity of others can be tested by performing *in vitro* and *in vivo* knockout studies.

10.5 Selective Case Studies

Earlier, it was shown that the analysis and interpretation of metabolic phenotype is feasible using genomic, biochemical, and strain-specific information (Edwards and Palsson 2000). In *E. coli*, stoichiometric model of metabolism has been established in order to describe the balance of metabolic reactions during steady-state growth on glucose and mineral salts (Pramanik and Keasling 1997). Similar *in silico* metabolic genotype models have been used for analysis and interpretation of phenotypic behavior of respective genotypes in *Haemophilus influenzae* and *Helicobacter pylori* (Edwards and Palsson 1999; Schilling et al. 2002).

Two recent extended graph theory-based methods have also been used in systems-based drug target identification, such as choke-point and load-point analysis. Choke points basically are those enzymes which uniquely consume and/or produce a certain metabolite, while load points are considered as the hotspots in the metabolic network based on the ratio of the number of *k*-shortest paths passing through a metabolite/enzyme (in/out) and the number of nearest neighbor links (in/out) attached to it. It has been considered that such analysis may assist in potential drug target identification. The choke-point approach has been used to identify the potential drug targets in the metabolic network of *Bacillus anthracis* Sterne (Mehta and Tagore 2009). In an experiment, drug targets were identified by comparing the fluxes of reactions in pathological and medication state and thus examining the change of reaction fluxes using linear programming models to find the steady optimal fluxes

of reactions and the mass flows of metabolites (Li et al. 2011). Similar analysis has been carried out in the analysis of hyperuricemia-related purine metabolic pathway, thus identifying the drug targets for hyperuricemia and also taking care of the possible side effects of the probable drug administered (Li et al. 2011). Also, various drug targets have been identified using the choke-point and load-point approach in *Helicobacter pylori* (Yeh et al. 2004). In an another attempt, a number of enzymes both single and paired that were essential for growth were identified which were not characterized even with gene deletion experiments, using comparative genome-scale metabolic reconstruction and flux balance analysis in *Staphylococcus aureus* (Lee et al. 2009).

In *M. tuberculosis*, seven new potential drug targets have been identified by analyzing the essential genes from mycolic acid pathway (Raman et al. 2005). As a unique initiative, Beste et al. (2007) created a web-based interface for the implementation of constraint-based flux analysis popularly known as GSMN-TB (http://sysbio3.fhms.surrey.ac.uk/). Using FBA, they have also identified potential drug targets against *M. tuberculosis.* In an attempt, expression-based data has been utilized to perform constraint-based FBA to predict the metabolic capacity of *M. tuberculosis* (Colijn et al. 2009). In 2006, Singh and Ghosh identified novel drug targets using the TCA cycle and simulating the inhibition of flux of branch point enzyme in *M. tuberculosis* (Singh and Ghosh 2006) using kinetic modeling.

In case of *P. falciparum*, genome-scale metabolic network, FBA, and *in silico* gene deletion, 40 essential gene targets were identified that can act as a potential drug target in the organism (Plata et al. 2010). In an analysis, Yeh et al. (2004) identified few potential drug targets in *P. falciparum* using similar approach (Yeh et al. 2004). D-Aminolevulinate dehydratase (ALAD) which has been involved in heme biosynthesis was hypothesized to be valid antimalarial target in *P. falciparum* using choke-point approach (Yeh et al. 2004; Bonday et al. 2000). In an attempt, Fatumo et al. (2009) successfully analyzed the essentiality of various metabolic reactions and also identified 22 new potential drug targets in *P. falciparum* using *in silico* knockout and choke-point analysis (Fatumo et al. 2009). Recently, successful identification of 16 novel drug targets, which contribute to the major metabolic reactions in *P. falciparum*, has shown the relevance of *in silico* approaches in the identification of novel drug target discovery (Huthmacher et al. 2010).

Recently, CD40 signaling has been used as a model to show the role of bimodular assembly of kinases in the imposition of reciprocity to a receptor signaling (Sarma et al. 2012). The finding of the analysis is considered relevant for designing appropriate therapy against the *Leishmania major* infection. In an attempt of similar kind, it was shown that the variation in the interaction designs among the kinases and phosphatases can differentially shape the robustness and signal response behavior of the MAPK cascade. It was thus concluded that the strength of the negative feedback loop is reciprocally related to the strength of phosphatase sequestration (Sarma and Ghosh 2012a, b). The analysis thus was found helpful in uncovering a novel regulatory aspect of MAPK signal processing.

10.6 Advances in New Target Discovery Mechanisms

Drug target discovery is one of the most challenging arenas in the area of modern-day pharmaceutical research. The problem of multidrug resistance (MDR) in various infectious diseases has forced scientist to look beyond the existing drug targets. It was hypothesized that the eradication of slowly metabolizing but the dormant populations of infectious bacterial species, essentially responsible for their relapse and MDR, could help in prevention of such problem. Therefore, drug targets pertaining not only to the growth of infectious organism but the target essential for their persistence can also assist in controlling their spread. Recently the web server UniDrug-Target has been released which combines bacterial biological information and computational methods to stringently identify pathogen-specific proteins as drug targets. The uniqueness of this database lies in the fact that apart from identifying drug targets based on systems approach, it also integrates protein sequences, domains, structures, and metabolic reactions, critical residue information for the prediction of novel drug targets (Chanumolu et al. 2012). Analysis of translationally controlled transcripts during host-pathogen interactions was also hypothesized to reveal some novel potential new drug targets in various bacterial species (Diaz-Guerra et al. 2008). Recently, a review on the significance of the application of genome-scale modeling and analysis in the context of identification of antitubercular drug targets has been highlighted (Lee et al. 2011). The drug-designing approach has almost taken a 180-degree turn due to the influence of bioinformatics, the process used to be starting from chemistry to biological testing, pharmacology, and toxicology tests; now, in post-genomic era, the understanding of disease biology, identification of protein or pathway to eliminate or control the disease via novel drug target(s), and then starting chemistry pipeline are the recent trends in medical sciences.

10.7 Conclusions

Over the years, research in the area of biology largely focus on understanding the crucial aspects of cellular and molecular machinery pertaining to infectious diseases in human and finding the effective drug targets for such diseases. It is well understood now that the infectious diseases or any response to such disease is a collective and coherent mechanism of various cellular and molecular processes and not the role played only on single gene or molecular instance. Therefore, a system-level analysis of infectious disease will not only assist in finding an effective drug target but also assist in developing a molecule that can cease the overall effect and spread of such diseases. The principles and theory of system-level analysis have been described in order to perform such kind of systemic analysis. This chapter focuses on several databases and approaches used in target identification in infectious disease in details for *M. tuberculosis* and *P. falciparum*; however, many such

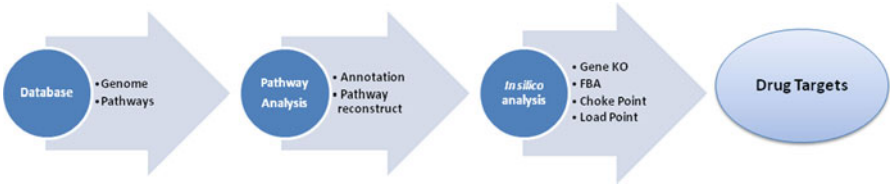


Fig. 10.4 Overview of *in silico* systems biology approach to the novel drug target identification

attempts are cited here for other pathogens. Theoretical approaches such as flux balance analysis, kinetic modeling, metabolic control analysis, *in silico* knockout studies, and load-point and choke-point analysis have been widely accepted and used to find the novel drug targets in various infectious organisms (Fig. 10.4). It has been emphasized that these potential drug targets can act as an effective lead in the identification of drug compounds that may in the future control the proliferation of infectious diseases.

Acknowledgements HRK would like to thank International Center for Genetic Engineering and Biotechnology (ICGEB), New Delhi, for its support and Department of Science and Technology (DST) for providing financial support. IG acknowledges the support from DBT, Government of India, and wish to thank Dr. Vivek Singh and Mr. Ashish Singh Sisodia for sharing the unpublished data from their thesis.

References

Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, Crowther GJ, Doyle MA, Hertz-Fowler C, Hopkins AL, McAllister G, Nwaka S, Overington JP, Pain A, Paolini GV, Pieper U, Ralph SA, Riechers A, Roos DS, Sali A, Shanmugam D, Suzuki T, Van Voorhis WC, Verlinde CL. Genomic-scale prioritization of drug targets: the TDR targets database. Nat Rev Drug Discov. 2008;7(11):900–7.

Anishetty S, Pulani M, Gautam P. *Mycobacterium tuberculosis* through metabolic pathway analysis. Comput Biol Chem. 2005;29:368.

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Barh D, Kumar A, Misra AN. Genomic Target Database (GTD): a database of potential targets in human pathogenic bacteria. Bioinformation. 2009;4(1):50–1.

Beste DJV, Hooper T, Stewart G, et al. GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. Genome Biol. 2007;8:R89.

Bi J, Wang H, Xie J. Comparative genomics of NAD(P) biosynthesis and novel antibiotic drug targets. J Cell Physiol. 2011;226(2):331–40.

Bonday ZQ, Dhanasekaran S, Rangarajan PN, Padmanaban G. Import of host delta-aminolevulinate dehydratase into the malarial parasite: identification of a new drug target. Nat Med. 2000;6: 898–903.

Brötz-Oesterhelt H, Sass P. Postgenomic strategies in antibacterial drug discovery. Future Microbiol. 2010;5(10):1553–79.

Butcher EC, Berg EL, et al. Systems biology in drug discovery. Nat Biotechnol. 2004;22(10): 1253–9.

Cakmak A, Qi X, Coskun SA, Das M, Cheng E, Cicek AE, Lai N, Ozsoyoglu G, Ozsoyoglu ZM. PathCase-SB architecture and database design. BMC Syst Biol. 2011;5:188.

Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee PW. Metabolic control analysis in drug discovery and disease. Nat Biotechnol. 2002;20(3):243–9.

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2008;36:D623–31.

Chanumolu SK, Rout C, Chauhan RS. UniDrug-target: a computational tool to identify unique drug targets in pathogenic bacteria. PLoS One. 2012;7(3):e32833.

Chua HN, Roth FP. Discovering the targets of drugs via computational systems biology. J Biol Chem. 2011;286(27):23653–8.

Cole ST. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. Eur Respir J Suppl. 2002;36:78s–86.

Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, et al. Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. PLoS Comput Biol. 2009;5(8):e1000489.

Collier N, Doan S. GENI-DB: a database of global events for epidemic intelligence. Bioinformatics. 2012;28(8):1186–8.

Cornish-Bowden A, Cárdenas ML. Technological and medical implications of metabolic control analysis. Dordrecht: Kluwer Academic Publishers; 2000.

Diaz-Guerra E, Vernal R, Cantero W, Müllner EW, Garcia-Sanz JA. Translation controlled mRNAs: new drug targets in infectious diseases? Infect Disord Drug Targets. 2008;8(4):252–61.

Drews J. Drug discovery: a historical perspective. Science. 2000;287(5460):1960–4.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA. 2007;104:1777–82.

Duggleby RG. The application of metabolic resistance theory to the selection of preferred target enzymes for therapeutic drugs. Comput Biomed Res. 1998;21(6):579–92.

Duggleby RG, Christopherson RI. Metabolic resistance to tight-binding inhibitors of enzymes involved in the de novo pyrimidine pathway, simulation of time-dependent effects. Eur J Biochem. 1984;143(1):221–6.

Durot M, Bourguignon P-Y, Schachter V. Genome scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev. 2009;33(1):164–90.

Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J Biol Chem. 1999;274(25):17410–16.

Edwards JS, Palsson BO. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A. 2000;97(10):5528–33.

Eisenthal R, Cornish-Bowden A. Prospects for antiparasitic drugs. The case of *Trypanosoma brucei*, the causative agent of African sleeping sickness. J Biol Chem. 1998;273(10):5500–5.

Farkas IJ, Korcsmáros T, Kovács IA, Mihalik Á, Palotai R, Simkó GI, Szalay KZ, Szalay-Beko M, Vellai T, Wang S, Csermely P. Network-based tools for the identification of novel drug targets. Sci Signal. 2011;4(173):pt3.

Fatumo S, Plaimas K, Mallm JP, Schramm G, Adebiyi E, Oswald M, Eils R, König R. Estimating novel potential drug targets of *Plasmodium falciparum* by analysing the metabolic network of knock-out strains *in silico*. Infect Genet Evol. 2009;9:351–8.

Fauci AS. Infectious diseases: considerations for the 21st century. Clin Infect Dis. 2001;32:675–85.

Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol. 2008;26:659–67.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Boil. 2007;3:121.

Fell DA. Metabolic control analysis: a survey of its theoretical and experimental development. Biochem J. 1992;286(Pt 2):313–30.

Gao Z, Li H, Zhang H, Liu X, Kang L, Luo X, Zhu W, Chen K, Wang X, Jiang H. PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics. 2008;9:104.

Gutierrez-Lugo MT, Bewley CA. Susceptibility and mode of binding of the *Mycobacterium tuberculosis* cysteinyl transferase mycothiol ligase to tRNA synthetase inhibitors. Bioorg Med Chem Lett. 2011;21(8):2480–3.

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004;430(6995):88–93.

Hecker N, Ahmed J, von Eichborn J, Dunkel M, Macha K, Eckert A, Gilson MK, Bourne PE, Preissner R. SuperTarget goes quantitative: update on drug-target interactions. Nucleic Acids Res. 2012;40:D1113–17.

Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. Eur J Biochem. 1974;42:89–95.

Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res. 1997;25(4):701–12.

Huthmacher C, Hoppe A, Bulik S, Holzhütter HG. Antimalarial drug targets in *Plasmodium falciparum* predicted by stage-specific metabolic network analysis. BMC Syst Biol. 2010;4:120.

Jeong H, Mason S, Barabasi A-L, Oltvai Z. Lethality and centrality in protein networks. Nature. 2001;411:41–2.

Jiang Z, Zhou Y. Using gene networks to drug target identification. J Integr Bioinformatics. 2005;2(1):14.

Kacser H, Burns JA. Control of enzyme flux. Symp Soc Exp Biol. 1973;27:65–104.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36:D480–4.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38:D355–60.

Karp PD, Paley S, Romero P. The pathway tools software. Bioinformatics. 2002;18:S225–32.

Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Curr Opin Biotechnol. 2003;14:491–6.

Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. Systems biology in practice: concepts, implementation and application. Weinheim: Wiley-VCH; 2005.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011;39:D1035–41.

LaPorte DC, Walsh K, et al. The branch point effect. Ultrasensitivity and subsensitivity to metabolic control. J Biol Chem. 1984;259(22):14068–75.

Lee DS, Burd H, Liu J, Almaas E, Wiest O, Barabási AL, Oltvai ZN, Kapatral V. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets. J Bacteriol. 2009;191(12):4015–24.

Lee D-Y, Chung BKS, Yusufi FNK, Selvarasu S. In silico genome-scale modeling and analysis for identifying anti-tubercular drug targets. Drug Dev Res. 2011;72:121–9.

Li H, Gao Z, Kang L, Zhang H, Yang K, Kunqian Y, Luo X, Zhu W, Chen K, Shen J, Wang X, Jiang H. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 2006;34:W219–24.

Li Z, Wang RS, Zhang XS. Two-stage flux balance analysis of metabolic networks for drug target identification. BMC Syst Biol. 2011;5(1):S11.

McDevitt D, Rosenberg M. Exploiting genomics to discover new antibiotics. Trends Microbiol. 2001;9(12):611–17.

Mehta S, Tagore S. Functional module analysis in metabolomics: chokes. Adv Comput Res. 2009;1:1–4.

Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. Nature. 2004;430:242–9.
Oberhardt MA, Palsson BO, Papin JA. Applications of genome-scale metabolic reconstructions. Mol Syst Biol. 2009;5:320.
Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Biotechnol. 2010;28(3):245–8.
Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov. 2007;6(1):29–40.
Peterson C, Ringner M. Analyzing tumor gene expression profiles. Artif Intell Med. 2003;28:59–74.
Pinney JW, Shirley MW, McConkey GA, Westhead DR. MetaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. Nucleic Acids Res. 2005;33:1399–409.
Plata G, Hsiao TL, Olszewski KL, Linás M, Vitkup D. Reconstruction and flux-balance analysis of the *Plasmodium falciparum* metabolic network. Mol Syst Biol. 2010;6:408.
Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng. 1997;56(4):398–421.
Raman K, Rajagopalan P, Chandra N. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. PLoS Comput Biol. 2005;1:e46.
Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol. 2003;4:R54.51–12.
Sakharkar KR, Sakharkar MK, Chow VT. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. In Silico Biol. 2004;4(3):355–60.
Sams-Dodd F. Target-based drug discovery: is something wrong? Drug Discov Today. 2005;10(2):139–47.
Sarma U, Ghosh I. Different designs of kinase-phosphatase interactions and phosphatase sequestration shapes the robustness and signal flow in the MAPK cascade. BMC Syst Biol. 2012a;6(1):82.
Sarma U, Ghosh I. Oscillations in MAPK cascade triggered by two distinct designs of coupled positive and negative feedback loops. BMC Res Notes. 2012b;5(1):287.
Sarma U, Sareen A, Maiti M, Kamat V, Sudan R, et al. Modeling and experimental analyses reveals signaling plasticity in a bi-modular assembly of CD40 receptor activated kinases. PLoS One. 2012;7(7):e39898.
Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO. Genome-scale metabolic model of *Helicobacter pylori* 26695. J Bacteriol. 2002;184(16):4582–93.
Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 2004;32:D431–3.
Singh V. Metabolic control analysis of biochemical pathways as an approach to *in silico* identification and validation of anti-tuberculosis, anti-malarial and anti-diabetic drug targets. Ph.D. thesis, University of Pune; 2009.
Singh VK, Ghosh I. Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in *Mycobacterium tuberculosis*, and its application to assessment of drug targets. BMC J Theor Biol Med Model. 2006;3:27.
Starck J, Kallenius G, et al. Comparative proteome analysis of *Mycobacterium tuberculosis* grown under aerobic and anaerobic conditions. Microbiology. 2004;150:3821–9.
Stark J, Callard R, Hubank M. From the top down: towards a predictive biology of signaling networks. Trends Biotechnol. 2003;21(7):290–3.
Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5(1):93–121.
Uchiyama I, Higuchi T, Kawai M. MBGD update 2010: towards a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res. 2010;38:D361–5.
Vidal M. Interactome modeling. FEBS Lett. 2005;579:1834–8.
Westley AM, Westley J. Enzyme inhibition in open systems. J Biol Chem. 1996;271(10):5347.
Whittaker PA. What is the relevance of bioinformatics to pharmacology? Trends Pharmacol Sci. 2003;24:434–9.

WHO. Removing obstacles to healthy development. Geneva: World Health Organization; 1999.
Yang IS, Ryu C, Cho KJ, Kim JK, Ong SH, Mitchell WP, Kim BS, Oh HB, Kim KH. IDBD: infectious disease biomarker database. Nucleic Acids Res. 2008;36:D455–60.
Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. Genome Res. 2004;14(5):917–24.
Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M. Drug-target network. Nat Biotechnol. 2007;25(10):1119–26.
Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. Nucleic Acids Res. 2004;32:D271–2.
Zhang D, Jia J, Meng L, Xu W, Tang L, Wang J. Synthesis and preliminary antibacterial evaluation of 2-butyl succinate-based hydroxamate derivatives containing isoxazole rings. Arch Pharm Res. 2010;33(6):831–42.

Hemant Ritturaj Kushwaha, Ph.D., India Hemant Ritturaj Kushwaha received his PhD degree in bioinformatics from School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, and M.Sc. in Biotechnology from University of Allahabad, Uttar Pradesh, India, in 2011 and 2002, respectively. He has been research associate at School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi (2011–2012), India.

He has been recipient of Innovation in Science Pursuit for Inspired Research (INSPIRE) faculty award, 2011, sponsored by Department of Science and Technology (DST), under Ministry of Science and Technology, Government of India. Currently, he is working as DST-INSPIRE faculty at Synthetic Biology and Biofuel group at International Center for Genetic Engineering and Biotechnology (ICGEB), New Delhi. His expertise involves core area of bioinformatics specially, genomics, proteomics, and systems biology.

Indira Ghosh, Ph.D., Professor, India Indira Ghosh is professor and ex-dean in School of Computational and Integrative Sciences of Jawaharlal Nehru University, New Delhi. Before completing her PhD from Indian Institute of Science, Bangalore, she received her MSc from University of Calcutta, West Bengal, India. She did her postdoctoral studies at University of Houston, USA. She has been a Robert Welch Foundation fellow at the University of Houston, Texas (1983–1986). Also, she has been a Fulbright Scholar (senior) awarded by USEFI (1983–1986).

Prof. Ghosh has worked in AstraZeneca PLC, multinational pharmaceutical industry, for 13 years in the area of drug target identification and lead molecule designing. Before coming to Jawaharlal Nehru University as a professor in 2008, she has served in University of Pune as a

professor, Bioinformatics and Biotechnology (during 2003–2008). She has been the member of Indian Biophysical Society, Society of Biological Chemists, India, American Chemical Society, and Cheminformatics and QSAR Society. She has been guiding several projects funded by Ministry of Communication & Information Technology, Government of India, and Department of Biotechnology, India. She has successfully completed projects funded by IBM, India (2004–2008), and also OpenTox (FP7) project funded by European commission (2008–2011). Her major interest includes the development of deeper understanding of diseases, like malaria, tuberculosis, and diabetes using bio- and chemoinformatics. Her research interest recently has been using systems biology, molecular simulation, and chemoinformatics to address the problem of disease.

Chapter 11
Identification of Network Biomarkers for Cancer Diagnosis

Jiajia Chen, Luonan Chen, and Bairong Shen

Abstract Researchers are now routinely identifying cancer biomarkers using proteomic technologies, but the results from independent experiments are highly divergent, and few individual marker candidates have been clinically validated. Cancer is a systems biology disease; hence, the discovery of biomarkers must take account of the complexity and heterogeneity of carcinogenesis. Due to this recognition, there is a growing movement from individual marker discovery to a systems-oriented paradigm. This chapter summarizes recent advances and future perspectives in proteomic study of cancer biomarkers. Of particular interest in this chapter is to describe the emerging network-based biomarker discovery as improved strategies against cancer intervention.

Keywords Cancer • Proteomics • Biomarkers • Network • Personalized biomarker

11.1 Introduction

Cancer is one of the leading causes of human death worldwide. According to GLOBOCAN 2008 (Ferlay et al. 2011), 7.6 million cancer deaths (around 13% of all deaths) are estimated to have occurred in 2008. The global burden of cancer is projected to continue rising as a result of population aging alongside an increasing

J. Chen, Ph.D. • B. Shen, Ph.D. (✉)
Center for Systems Biology, Soochow University, P. O. Box 206, Soochow University
No.1 Shizi street, Suzhou, Jiangsu 215006, China
e-mail: njucjj@126.com; bairong.shen@suda.edu.cn

L. Chen, Ph.D.
Key Laboratory of Systems Biology, Shanghai Institutes for Biological
Sciences Chinese Academy of Sciences,
P. O. Box 5, 500 Cao Bao Road, Shanghai 200233, China
e-mail: lnchen@sibs.ac.cn

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_11,

adoption of cancer-causing lifestyles. A report from the International Agency for Research on Cancer (IARC) predicts that the worldwide cancer deaths will more than double over the next two decades. By 2030, 27 million new cancer cases and 17 million cancer deaths will occur each year worldwide.

It is well recognized that early detection, accurate prognosis, and monitoring of therapy are crucial to the increased overall survival and cure rates of cancer patients. The cancer development usually involves alterations in gene expression. These alterations, which can be monitored quantitatively, are regarded as biomarkers for the disease. As important biological indicators, biomarkers provide valuable information that enables early diagnosis, effective prognosis, and real-time monitoring of the therapy response. Thus, the impetus to discover reliable and accurate cancer biomarkers is compelling.

Great efforts have been made in genomics studies to characterize the differentially expressed genes in cancer, leading to a plethora of gene-based biomarkers. Recently, there is growing realization that it is the proteins not genes that perform most biological functions and directly dictate cancer phenotypes. Thus, many alterations in gene expression might not be reflected at the level of protein expression or function. Proteomics is the most reliable approach in the search for cancer biomarkers in the post-genomic era. However, the protein universe of a cell is highly complex and diverse. Due to the complicate nature of protein functions, proteomic-based biomarker discovery is a research topic with significant challenge. Any such endeavor must take into account the dynamic range of protein expression and protein–protein interactions as well as posttranslation modifications.

11.2 Advances of Proteomic Methods

The past decade has witnessed rapid progress in proteomic instrumentation. A plethora of state-of-the-art techniques have been developed to resolve, identify protein complexes, and detect their interacting partners. These powerful technologies can identify all proteins and their posttranslational modifications in cancer conditions and hence will accelerate progress toward novel tools to track cancer progress and tailor treatments to the patients.

11.2.1 Separation

The standard pipeline for proteomic research begins with the isolation of proteins from the samples, then digestion with site-specific protease, followed by separation steps to resolve protein fractions from the complex sample mixture. Since proteins are physically and chemically diverse, there is no single universally applicable separation method. A multitude of separation techniques are currently available, including gel-based approaches (e.g., 2DE, 2D-DIGE (Marouga et al. 2005))

or gel-free ones (e.g., LC, cICAT (Gygi et al. 1999), iTRAQ (DeSouza et al. 2005), SILAC (Ong et al. 2002)). Two-dimensional electrophoresis (2DE) represents the most common gel-based separation approach. It separates proteins based on size in one dimension and charge in the other. Proteins are stained with fluorescent dyes prior to electrophoresis, and up to 5,000 distinct proteins can be separated simultaneously. Limitations of 2DE are that it cannot resolve hydrophobic proteins or proteins with extreme masses or isoelectric points (Ong and Mann 2005). Additionally, 2DE suffers from limited sensitivity and dynamic range. In this sense, liquid chromatography (LC), a most popular gel-free separation method, provides an alternative to 2DE gels. In LC, the desired spots are isolated and digested into peptides (rather than intact proteins), which are then chromatographically separated by LC. Depending on the sample complexity, the low-molecular weight fractions may be further separated by high-resolution ion-exchange chromatography.

11.2.2 Mass Spectrometry

After separation, the peptides are then subjected to mass spectrometry for qualitative and quantitative analysis. Recent developments in mass spectrometry (MS)-based technologies have largely contributed to the current progress in proteomics. MS is able to convert proteins or peptides to charged pieces that can be separated on the basis of the mass-to-charge ratio (m/z) and their abundances. In the past two decades, MS has been established as the primary method for protein identification at the peptide level. This is done by matching peptide mass fingerprinting against a human protein database. In addition, partial sequence data obtained by tandem mass spectrometry (MS/MS) provides a further basis for protein identification. In MS/MS, an ion of interest is selected and partially fragmented at the peptide bonds. The products then passage through a second mass analyzer, and the resulting fragmentation pattern provides sequence information for the precursor ion. Several MS ionization methods are currently available, including electrospray ionization (ESI), surface-enhanced laser desorption ionization (SELDI), and matrix-assisted laser desorption ionization (MALDI). Known as "soft ionization," these ionization procedures create peptide ions under mild conditions that maintain peptide bonds intact. However, "soft ionization" has yet some reported disadvantages. The resolution is found insufficient to handle complex proteomic samples which contain more than 10^5 components and is not able to distinguish accurately all of the protein variants in the sample. Therefore, analytical techniques that can handle high sample complexity are logically preferable over low-density ones. There is growing use of high-resolution analytical techniques such as Qq-TOF (Lando et al. 2002) and RPLC-FTICR (Shen et al. 2001) mass spectrometry in place of SELDI-TOF/MALDI-TOF mass spectrometry. Qq-TOF and RPLC-FTICR provide resolving power that is roughly 5 times and 100 times higher than SELDI-TOF-MS, respectively, resulting in a dramatic increase in resolution performance.

11.2.3 Protein Microarrays

MS may be applied to broad surveys of the proteome for biomarkers without prior knowledge of the proteins. However, MS-based marker discovery has been the topic of debate. Criticism is directed at the lack of details regarding the putative biomarkers, specifically their identity. An additional problem with the approach is that the samples are usually limited in amount and are difficult to reproduce. Protein microarrays, also known as protein chips, may overcome these difficulties as even a small amount of purified protein is sufficient to print hundreds of arrays for sample screening. Protein microarrays have emerged as a promising approach for a wide variety of applications. The array is a piece of glass on which fluorescently labeled capture probes for the proteins are affixed. Antibodies are most commonly used as capture probes. More recently, there has been a push toward other types of capture molecules such as peptides, aptamers, and affibodies.

There are several broad categories of protein microarrays: analytical microarrays, functional microarrays, and reverse phase microarrays. Analytical microarrays are typically used to monitor differential expression profiles for clinical diagnostics. Functional protein arrays contain proteins or domains with intact function and proper folding, allowing for the determination of protein interactions. In reverse phase protein microarray, arrays compose of cell or tissue lysates or protein fractions isolated from them. Reverse phase protein microarrays are used to study posttranslational modifications that may result from the disease.

11.3 Current Cancer Biomarker Discovered by Proteomics

With recent progress in proteomic technologies, there have been intense interests in proteome-scale discovery of novel biomarkers for cancer, as documented by the numbers of published literatures in the last 10 years (Fig. 11.1).

Through profiling proteomic patterns in cancer patients, candidate lists of protein markers are routinely identified for use in the diagnosis, prognosis, and classification of cancer, as well as the prediction of therapeutic response.

11.3.1 Proteomic Signatures for Early Diagnosis of Cancer

In many cases, cancer is not diagnosed until malignant cells already metastasize throughout the body. Therefore, the diagnosis of early phase malignancies is crucial for its ultimate control and reduced mortality. A diagnostic marker is able to differentiate cancer with other benign abnormalities. Many studies have been published recently showing how proteomics can be applied to the detection of cancer at its earliest stage. In one impressive study by Chaerkady et al. (2008), large-scale tissue proteomic profiling was carried out to identify differentially regulated proteins in HCC (hepatocellular carcinoma) patients by iTRAQ. Using this strategy, some

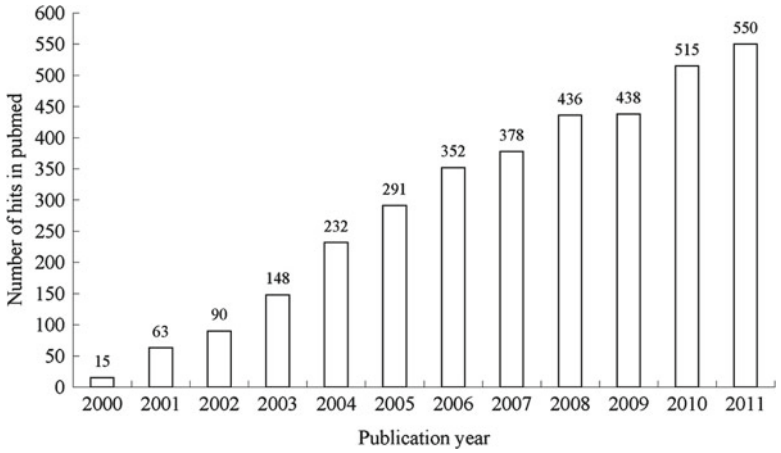


Fig. 11.1 *Number of papers related to proteome-based biomarkers in cancer during the past decade.* Although this query is by no means exhaustive, it provides a meaningful sampling of the available literature. *Bars* represent the number of PubMed hits for the query "(cancer[ti] or carcinoma[ti] or tumor[ti]) AND proteom*[tiab]"

overexpressed proteins with no previous description were discovered including fibroleukin, interferon-induced 56 kDa protein, milk fat globule-EGF factor 8, and myeloid-associated differentiation marker. Using SILAC combined with LTQ-FT-MS/MS, Sun et al. (2008) identified differentially expressed protein profiles between normal liver cell (HL-7702) and HCC cells (HepG2 and SK-HEP-1). TGM2 was suggested to be a novel histological and serologic candidate for the HCC patients with normal serum AFP. Goufman et al. (2006) analyzed the thermostable fractions of serum samples from benign ovarian tumor as well as from patients with ovarian, uterus, and breast cancers, using 2-DE combined with MALDI-TOF/TOF MS. Alpha-1-acid glycoprotein and clusterin were downregulated specifically in breast cancer, whereas transthyretin was decreased in ovarian cancer. More recently (Lee et al. 2009), deregulated proteins in HCC liver tissues were investigated in a cohort of 20 patients using 2D-DIGE coupled with MALDI-TOF-MS. Three remarkably upregulated proteins—aldo-keto reductase 1C2, thioredoxin, and transketolase—were identified and recognized as potential markers that could distinguish between HCC and normal liver tissues. These novel findings might provide important clues to the early detection of various cancer types.

11.3.2 Prognostic Biomarkers for Prediction of Patient Outcome and Therapeutic Response

Prognostication and the variability of therapeutic agents are a topic of major interest in cancer clinical research. Protein markers that reflect therapy response may help predict clinical outcome of the patients and evaluate the therapeutic efficacy. Using

high-throughput proteomic technologies, a variety of studies have recently identified patterns of proteomic expression that are predictive of cancer progression, survival, as well as drug response in cancer patients. As an example, Abdel-Hamid et al. (2011) evaluated the efficacy of ascorbic acid (AA) and diallyl sulfide (DAS) co-therapy against HCC. Polyol pathway (PP) was identified as an early marker for response to chemoprevention. More recently, Li et al. (2011) screened the markers correlated with clinical outcome for transcatheter arterial chemoembolization (TACE) therapy in inoperable HCC. Plasma serine protease inhibitor precursor (protein C inhibitor) was identified as a direct link between peptide marker profiles and TACE therapy. These related works will be instrumental in characterizing drug-resistance mechanisms and thus optimizing anticancer therapies for patients.

11.3.3 Proteomic Signatures Associated with Cancer Classification

Cancer usually develops in association with chronic diseases. Cancer subgroups based on different pathological risk factors usually reflect common pathological features. This poses an important clinical problem for pathologists, because these subgroups could not be easily distinguished from each other. As an alternative, comprehensive proteomic profiling in the past has enabled us to identify a multitude of molecular biomarkers with such discriminative significance. In a pioneering study, Poon et al. (2001) developed classification trees and neural networks that identified serological liver marker profiles and subsequently applied them to discriminate between liver cirrhosis with and without HCC. In a later study by the same research group (Poon et al. 2003), an automated high-throughput, multidimensional SELDI ProteinChip Biomarker System was used for serum proteome profiling. Both cluster analysis and artificial neural network were applied to the identification of liver marker profiles that discriminate HCC subtypes with high sensitivity and specificity. Another interesting study was aimed at discovering novel biomarkers for the classification of ovarian cancer by SELDI-TOF mass spectrometry (Petricoin et al. 2002).

11.4 Databasing of Proteomic Data to Facilitate Systems Interrogations

As the discovery process of biomarkers is being accelerated, there will be an urgent need to store information in proteomic databases. These databases deposit and retrieve proteomic datasets generated from various tissues, fluids, and cell lines. A large number of proteomic databases have been developed, including PRIDE (Martens et al. 2005), UniPep (Zhang et al. 2006), GPM (the Global Proteome Machine) (Craig et al. 2004), Proteopedia (Hodis et al. 2008), PeptideAtlas (Desiere

et al. 2006), CanProVar (Li et al. 2010), and PPD (Plasma Proteome Database) (Muthusamy et al. 2005). Besides annotating the human proteome, these databases will offer a global approach to the study of protein expression in cancer.

11.5 Pathway-Based Approaches to Discovering Novel Cancer Biomarkers

Despite years of intense activity, so far only a handful of the reported biomarkers (e.g., CA-125, CEA, PSA, AFP, DCP) have been validated for routine clinical application in cancer, although some were first reported decades ago. However, the use of these markers is limited due to poor specificity or sensitivity. Thus, the development of new markers that can complement or replace them represents a major goal for cancer biomarker research. The inconsistency of biomarker candidates from different researches remains a major obstacle to their clinical use. For example, in two independent research to identify differentially expressed proteins in HCV-related hepatocellular carcinoma (Yokoyama et al. 2004; Takashima et al. 2003), both groups claimed to have generated protein lists with predictive power, but surprisingly the two lists shared no common protein. Several recent studies have shed light on this problem through meta-analysis across independent datasets (Shi et al. 2005; Choi et al. 2004), but a highly reproducible molecular pattern with universal predictive utility has yet to emerge. One might attribute the diverse outcome to the differences in profiling techniques, specimen manipulations, and analysis methods. Yet this discordance remains even in technical replicate tests using identical samples as in the case of Ein-Dor et al. (2005). Therefore, signature identification at the level of individual proteins has been challenged about its robustness and reliability.

It is now understood that carcinogenesis is a nonlinear dynamic system involving cross talk between pathways and important regulatory circuits. Cancer involves the interaction of many biological components and is not driven by individual causative proteins. While the development of cancer is systems oriented, the current profile-based methodology remains fundamentally reductionism. The conventional expression-alone analysis, which focused on selecting single or a small group of differentially regulated entities over a cutoff value, often ignores the complex interplay within the identified biomarker sets. It is time, one could argue, for a fresh approach to the discovery of cancer biomarkers from the systems perspective. Due to this recognition, many researchers have resorted to a more systems-oriented approach, considering not just individual marker molecules, but dynamic interactions among them as an entire functional network.

Due to the appreciation that mutated proteins often reside in key signaling pathways, an importance of pathway analysis has been emphasized in the study of cancer biomarkers. Pathway analysis typically correlates a given set of deregulated proteins by projecting them onto well-characterized biological pathways that have been previously defined. A number of canonical signaling and metabolic pathways have been collected in curated databases, such as Kyoto Encyclopedia of Genes and

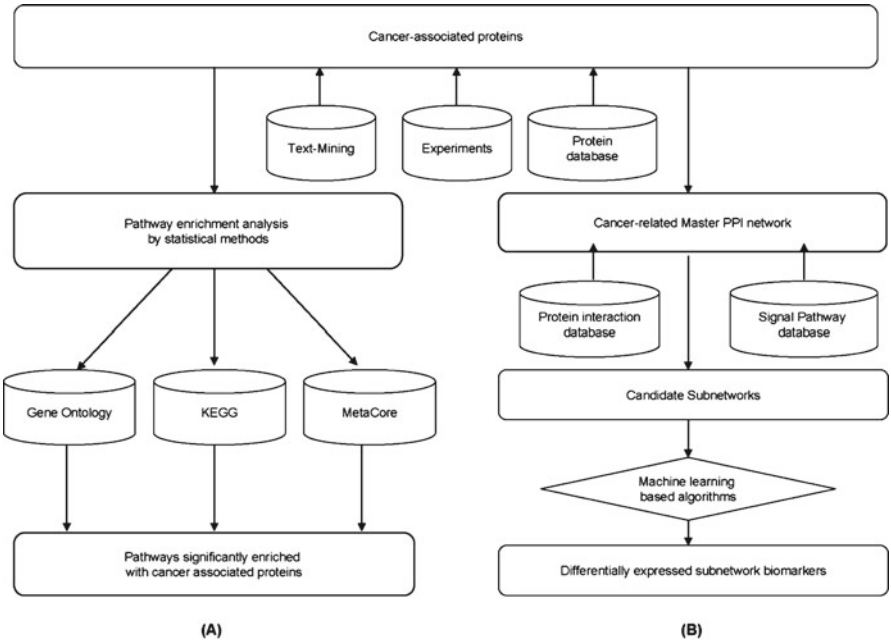


Fig. 11.2 The scheme for (**a**) pathway-based biomarker discovery (**b**) network-based biomarker discovery

Genomes (KEGG, http://www.genome.jp/kegg), Molecular Signatures Database (MSigDB), IngenuityPathway Analysis (IPA, http://www.ingenuity.com), GeneGo by MetaCore™ (http://www.genego.com/matacore.php), and Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea). Enrichment of pathways can be evaluated by overrepresentation statistics. The overall flowchart of the proposed pathway-based biomarker approach is illustrated in Fig. 11.2a.

Using this pipeline, the most enriched pathways are identified and then proposed to be the potential candidate markers. Several pathway-level signatures of cancer have recently been explored and defined. For example, Bi et al. used MALDI-TOF/TOF MS to observe changes in protein expression levels. The change revealed an enhanced glycolytic pathway, a decreased gluconeogenesis, a suppressed glucuronic acid pathway, and an impaired tricarboxylic acid cycle in colorectal cancer (Bi et al. 2006). In another study by Wurmbach et al. (2007), pathway analysis was performed to discriminate early HCC from dysplasia. A significant downregulation of components of the Toll-like receptor pathway, Jak/STAT pathway, TGF pathway, and the insulin-signaling pathway were reported. Components of the Wnt-signaling pathway were found to be upregulated in HCC patients. Using a similar pathway-based approach, Wang et al. found endothelin-1/EDNRA transactivation of the EGFR pathway, a putative novel PCa-related pathway (Wang et al. 2011). The above insight into molecular pathways provides a way to reveal the underlying mechanism of tumorigenesis as well as to design novel targeted therapeutic strategies for cancer.

Pathway-based approaches allow biologists to detect modest expression changes of functionally important genes that would be missed in expression-alone analysis. In addition, this approach enables the incorporation of previously acquired biological knowledge and makes a more biology-driven analysis of proteomic data. The pathway-based approach holds great promise for improved cancer classification and outcome prediction. However, the human proteins that have not been assigned to a definitive pathway undermine the basis for a strictly pathway-centric marker discovery. In addition, the cross talk between different signaling pathways further complicated the pathway-based analysis. Thus, a need for a network-based approach that accounts more extensive protein interaction is obvious.

11.6 Network-Based Approaches to Discovering Novel Cancer Biomarkers

In contrast to "pathway analysis," in which differentially expressed proteins are overlaid onto predefined pathways, networks are more of a global representation of the biological entities. Network-based approach is built on the observation that cancer proteins often function as network hubs that are involved in many signaling pathways. This methodology considers network modularity as a defining feature of tumor phenotype and aims to identify cancer-related network modules. The general strategy for network-based approach involves the following two steps. The first step is to create a master protein–protein interaction (PPI) network of proteins that are related with the cancer phenotype. The second step is to identify significant subnetworks representing relevant functional modules. A subnetwork refers to a smaller or more focused network within a large protein interaction network. The scheme for the network-based biomarker discovery is depicted in Fig. 11.2b.

11.6.1 The Inference of the Cancer-Related Master Protein–Protein Interaction (PPI) Network

Traditional systems biology approaches are focused on a mathematical description of biological systems. However, the flood of high-throughput post-genomic data has prompted great interest in inferring the networks from the data themselves. The inference of the phenotype-related master PPI network is a knowledge-based one, which relies on the established protein knowledge, including protein annotations, protein–protein interactions, as well as curated pathway information. The first step in master network construction is to determine whether a protein is related to the phenotype of interest according to protein annotation Ingenuity Pathway Analysis. However, most of this knowledge hides in published articles or databases. IPA (Ingenuity Pathway Analysis) is a beautiful software that can provide access to such disease information of proteins based on its manually integrated published

results in some big journals. Subsequently, these phenotype-associated proteins are used as seeds to construct a protein–protein interaction network. In this network, nodes represent proteins, whereas edges correspond to regulatory interactions between the nodes. The interactions include direct associations (physical interactions) and indirect associations (functional correlations).

11.6.1.1 PPI Network Construction Based on Physical Interactions

Direct PPIs associated with the network seeds can be extracted via experimentally validated or expert-annotated interactions observed in humans. Over the past few years, high-throughput methods such as yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (AP-MS) have accelerated the generation of protein–protein interaction data on a large scale. A large number of PPIs have been experimentally determined in the human interactome. The primary resources for the PPI data are individual scientific publications. To make this information more readily available, a number of publicly available databases have set out to collect and store various types of information about protein–protein interactions, for example, HPRD (Human Protein Reference Database) (Keshava Prasad et al. 2009), InterPro (Apweiler et al. 2001), OPHID (Online Predicted Human Interaction Database) (Brown and Jurisica 2005), MINT (the Molecular INTeraction database) (Zanzoni et al. 2002), BioGRID (Biological General Repository for Interaction Datasets) (Stark et al. 2010), DIP (Database of Interacting Proteins) (Salwinski et al. 2004), BIND (Biomolecular Interaction Network Database) (Bader et al. 2003), IntAct (Hermjakob et al. 2004), STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) (von Mering et al. 2003), MPPI (MIPS Mammalian Protein–Protein Interaction database) (Pagel et al. 2005), and HPID (Human Protein Interaction Database) (Han et al. 2004). The detailed description of these databases is listed in Table 11.1.

In parallel with these databases, plain text information found in the scientific literature also provides important sources of PPI. Therefore, text mining tools have been developed to automatically extract interrelations between proteins from literatures, for example, Pathway Studio (ariadnegenomics.com/products/pathway-studio/), CBioC (cbioc.eas.asu.edu), Chilibot (www.chilibot.net), GoPubMed (www.gopubmed.org), iHOP (www.ihop-net.org/UniPub/iHOP), iProLINK (pir.georgetown.edu/iprolink), PreBIND (prebind.bind.ca), PubGene (www.pubgene.org), and Whatizit (www.ebi.ac.uk/webservices/whatizit/info.jsf). Such text mining methods also provide valuable information for network inference.

11.6.1.2 Network Expansion

The generation of interaction networks is followed by the network expansion, by which proteins highly connected with seed proteins are also included. The network can be expanded in different ways. According to next neighbor expansion method

Table 11.1 Online database of protein–protein interactions

Database	Website	Description
BIND	http://www.bind.ca	Biomolecular interaction, complex, and pathway information
DIP	http://dip.doe-mbi.ucla.edu	Experimentally determined interactions between proteins
HPRD	http://www.hprd.org	Manually curated interaction networks for each protein in the human proteome a (database)
HPID	http://www.hpid.org	Integration of the protein interactions in BIND, DIP, and HPRD
IntAct	http://www.ebi.ac.uk/intact	Manually curated molecular interaction data from the literature
MINT	http://mint.bio.uniroma2.it/mint	Experimentally verified protein interactions mined from the literature from mammalian organisms
STRING	http://string.embl.de	Known and predicted protein–protein interactions
MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/	High-quality experimental protein interaction data in mammals
BioGRID	http://thebiogrid.org	Protein and genetic interactions from major model organisms
Interpro	http://www.ebi.ac.uk/interpro/	Integration of PROSITE, PRINTS, Pfam for protein families, domains, and functional sites
Orphid	http://ophid.utoronto.ca	Predicted interactions between human proteins

proposed by Chen et al. (2006), if seed A interacts with seed B, then the interactions of the type A-X-B is also included into the network, where X represents a protein out of the seed list. In this manner, all interacting partners of the seeds are extracted from the PPI database, and an expanded protein interaction network is generated. Given the roles of these seed proteins in signal transduction, it is also proposed that the signal proteins that interact with them also contribute to the pathogenesis. So it is also reasonable to expand the network by including the signal partners of seed proteins to get a larger PPI network. The signal proteins appearing in the signal pathways can be detected based on the knowledge from databases of signal transduction pathways. After expansion, a rough PPI network can be constructed that consisted of both cancer-associated proteins and the proteins that are tightly connected with them.

11.6.1.3 PPI Network Construction Based on Functional Associations

A drawback of the networks based on physical interaction is that they are false-negative rich. It is likely that the vast majority of interactions have not yet been described. Another significant issue is that the expression levels of the proteins are not taken into account, and thus irrelevant features which do not play a role in cancer might be included, increasing false-positive rates. Associative networks have

emerged, in part, as an alternate way to circumvent these technical issues. In this approach, networks are inferred based on the functional associations of the network members. Statistical inference methods are used to determine the relationships between components in the system based on shared expression patterns. Pair-wise correlations are used for characterizing co-expression of two proteins. Two proteins exhibiting a high correlation coefficient of their expression profile are considered co-expressed and hence functionally associated. This type of network is very specific to the disease states of interest and potentially includes more relevant biological components than can be currently considered using physical interaction data. A consideration of associative network, however, is that the expression association does not necessarily imply functional relation and further experimental validation is always necessary.

11.6.1.4 Refinement of the Master PPI Network

The resulting rough PPI master network described above can be refined with additional data such as gene functional annotations and transcription factor modules. These information are relevant to study genes and their regulatory interactions and allow further to characterize the interrelation between members of the network. For example, proteins with similar functional annotation in Gene Ontology might be functionally associated. Additionally, a shared transcription factor indicates co-regulation between two proteins. These features could be collected and used to calculate correlation values for each interactive pair of the interaction graph. "Strong" pair-wise interactions with higher correlation values will be focused in subsequent analysis. In doing so, the rough PPI network is refined by reducing the amount of potential false-positive interactions.

11.6.2 The Identification of Cancer-Related Subnetworks

After construction of the cancer-specific PPI master network, the next step in network biomarker discovery is to examine the functional modules for associations with specific disease phenotypes. Subnetwork is defined as small protein interaction units reflecting single functional modules. Subnetwork biomarkers can contain different numbers of proteins, for example, pair biomarker with two proteins and one protein–protein interaction, triple biomarker with three proteins and three protein–protein interactions, and square biomarker with four proteins and four protein–protein interactions. The identification of such subnetworks generally involves several scoring and search steps. The coherent expression patterns of the subnetwork members are used as feature values to train a classifier model based on pattern recognition algorithm. Then, a particular candidate subnetwork is scored based on their classification performances in testing sets. High-scoring subnetworks can then be associated with the disease state. In this manner, each significant subnetwork

with discriminative activities may be viewed as a putative marker. A variety of pattern recognition algorithms have been developed and applied to cancer classification; several representative examples include Bayesian networks, k-nearest neighborhood, logistic models, decision tree learning, partial least square, artificial neural networks, inductive logic programming, support vector machines, and clustering. These methods are in general data driven and depend on a linear or nonlinear discriminant function.

11.6.3 Performance Evaluation of the Network Biomarkers

Finally, the overall performance of the resulting subnetwork classifiers is evaluated by statistical methods, such as cross-validation. Cross-validation is based on splitting the available data into training and test sets. In *k*-fold cross-validation, for example, the dataset is randomly divided into *k* subsets. A single subset is used for the test dataset, and the remaining *k*-1 subsets for training. Training sets are used to train the parameters of classification model, and the test sets are used to evaluate the predictive accuracy of the model, and this procedure is repeated *K* times. The overall predictive accuracy (ACC) can be used to measure the prediction performance. ROC (receiver operating characteristic) curve is used to evaluate the performance by presenting both sensitivity and specificity against different parameters. The ROC analysis is only valid for the binary classification. The area under the ROC curve (AUC) allows to directly compare the inference quality against a random prediction. AUC < 0.7 is considered poor, AUC > 0.8 good. However, the model validation using internal dataset may be insufficient. Classification model should be assessed by additional information, for example, independent dataset that is not used for modeling. Knowledge for such "external validation" is available from experiments, literatures, or databases.

11.6.4 Network-Based Biomarkers for Several Types of Cancers

Network-based approach to biomarker discovery has already proven to aid in the classification and outcome prediction of several types of cancers. For instance, in a pioneering study by Chuang et al. (2007), subnetworks extracted from protein interaction databases were used to distinguish metastatic from nonmetastatic breast cancer. Using expression data from two previous studies, they constructed a network classifier based on logistic regression, which was then evaluated using fivefold cross-validation. For the two studies, the author obtained subnetwork markers with higher accuracy than single markers reported in the original studies. In addition, the resulting subnetworks were reported to be more reproducible between datasets. Subsequently, many other groups have applied network-based approach to prioritizing protein interaction subnetworks that are discriminative of disease signature.

Taylor et al. (2009) used network-level analysis to improve forecasting in breast cancer outcome. Using expression datasets that were obtained from two breast cancer patient cohorts, the author has searched for changes in the global modularity of the human interactome. They argued that the clustering-based classification model displayed favorable performance and suggested altered network modularity to be potential indicator of breast cancer prognosis. Recent study by Jin et al. (2009) further emphasized the importance of network-level perspective in cancer prediction. A prostate cancer-related network was built up by searching the interactions among identified molecules related to prostate cancer. The subnetwork biomarkers were derived from protein–protein interactions and then evaluated using cluster analysis. As a result, subnetworks were achieved with higher accuracy in the classification of prostate cancer. Using 2D-DIGE coupled with MALDI-TOF-MS/MS, Ummanni et al. (2011) identified 79 differentially expressed proteins which were then used to create a master global network. Functional subnetworks from the significant hubs of main network were subsequently built using IPA software. The functional subnetworks revealed several candidate biomarker proteins that were further validated by experiments. Xu et al. (2008) developed a graph-based method to integrate multiple microarray datasets to discover cancer-related co-expression network modules. The method helped discover network modules specific to cancer or its subtypes. More recently, Wang and Chen (2011) developed a computational framework to construct the network-based biomarker for molecular investigation and diagnosis of lung cancer. By network comparison, 40 significant proteins that play potentially important roles in lung carcinogenesis were identified as network-based biomarker. Zhang et al. (2011) developed a diagnostic classifier for hepatocellular carcinoma by combining the differential gene expression with topological features of human protein interaction networks. The classifier had high predictive accuracy and sensitivity to enhance the ability of HCC diagnosis. In a most recent endeavor, Liu et al. (2011) proposed a new approach to detect dysfunctional networks or modules as biomarkers for diseases. The method was then applied to the study of three-stage microarray data for gastric cancer. The result demonstrated the predictive power of the molecular interaction network, which in turn can be used as robust module biomarkers for accurately detecting or diagnosing gastric cancer.

11.7 Personalized Biomarkers for Improved Cancer Diagnosis and Treatment: A Future Perspective

Another key challenge of cancer biomarker development is the biochemical noise associated with interpatient variation. Protein expression profiles of cancer patients are diverse, depending on etiological and genetic factors or personal behaviors. It may be that differences in expression that appear to be related with the disease may in fact represent random genetic variation. This situation will further introduce false discoveries and reduce the overall reproducibility of biomarker detection. This concern was mentioned by Michiels et al. (2005), who investigated the stability of

seven published datasets to predict prognosis of cancer patients. It was observed that the predictive marker lists reported by the various groups were highly unstable and depended strongly on the subset of samples chosen for training. Considering the patient-related diversity, a more practical approach would be to search for common signatures among a cohort of genetically homogeneous people, other than a mixed population. The ongoing challenge for personalized biomarker is to group patients into more meaningful and well-defined subgroups on the basis of each person's unique genetic and environmental information. In this way, the individual difference of cancer mechanism is accounted when we analyze cancer expression data from different resources (Nicholson 2006; Yan 2008).

Personalized proteomic signature shows great potential to screen subsets of patients that will most likely benefit from the therapeutics. Under this scenario, it may ultimately help to customize these strategies to the individual patient. Personalized biomarker research requires patient-specific information via large-scale population proteomic studies, which could only be accomplished by a coordinated effort from funding agencies and research institutions. There have been several international initiatives that integrate studies performed in different groups worldwide to provide detailed information for personalized medicine. An initiative of the US National Cancer Institute, the Early Detection Research Network (EDRN) (Srivastava and Kramer 2000), provides a database that incorporates new molecular diagnostics and biomarkers and coordinates efforts in biomarker research. These efforts would hopefully accelerate the translation of biomarker information into personalized clinical applications.

11.8 Conclusion

Network and pathway analyses have complementary strengths and weaknesses. The two approaches collectively present a more complete picture of the system under study. Although proteomic studies of cancer biomarkers are still in their infancy, the evolving fields of pathway-based and network-based biomarkers offer new insight into cancer mechanisms and allow more robust diagnostics and tailored therapy against cancer. In addition, collaborative studies between academia and pharma are required if further progress is to be made. Hopefully, the pioneering studies reviewed herein might radically reform the clinical practice of cancer in the near future.

References

Abdel-Hamid NM, Nazmy MH, Abdel-Bakey AI. Polyol profile as an early diagnostic and prognostic marker in natural product chemoprevention of hepatocellular carcinoma in diabetic rats. Diabetes Res Clin Pract. 2011;92:228–37.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001;29:37–40.

Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003;31:248–50.

Bi X, Lin Q, Foo TW, Joshi S, You T, Shen HM, Ong CN, Cheah PY, Eu KW, Hew CL. Proteomic analysis of colorectal cancer reveals alterations in metabolic pathways: mechanism of tumorigenesis. Mol Cell Proteomics. 2006;5:1119–30.

Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005;21:2076–82.

Chaerkady R, Harsha HC, Nalli A, Gucek M, Vivekanandan P, Akhtar J, Cole RN, Simmers J, Schulick RD, Singh S, et al. A quantitative proteomic approach for identification of potential biomarkers in hepatocellular carcinoma. J Proteome Res. 2008;7:4289–98.

Chen JY, Shen C, Sivachenko AY. Mining Alzheimer disease relevant proteins from integrated protein interactome data. Pac Symp Biocomput. 2006;11:367–78.

Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S. Integrative analysis of multiple gene expression profiles applied to liver cancer study. FEBS Lett. 2004;565:93–100.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004;3:1234–42.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006;34:D655–8.

DeSouza L, Diehl G, Rodrigues MJ, Guo J, Romaschin AD, Colgan TJ, Siu KW. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. J Proteome Res. 2005;4:377–86.

Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005;21:171–8.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2011;127:2893–917.

Goufman EI, Moshkovskii SA, Tikhonova OV, Lokhov PG, Zgoda VG, Serebryakova MV, Toropygin IY, Vlasova MA, Safarova MR, Makarov OV, Archakov AI. Two-dimensional electrophoretic proteome study of serum thermostable fraction from patients with various tumor conditions. Biochemistry (Mosc). 2006;71:354–60.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999;17:994–9.

Han K, Park B, Kim H, Hong J, Park J. HPID: the Human Protein Interaction Database. Bioinformatics. 2004;20:2466–70.

Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004;32:D452–5.

Hodis E, Prilusky J, Martz E, Silman I, Moult J, Sussman JL. Proteopedia – a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. Genome Biol. 2008;9:R121.

Jin G, Zhou X, Cui K, Zhang XS, Chen L, Wong ST. Cross-platform method for identifying candidate network biomarkers for prostate cancer. IET Syst Biol. 2009;3:505–12.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human Protein Reference Database – 2009 update. Nucleic Acids Res. 2009;37:D767–72.

Lando D, Peet DJ, Whelan DA, Gorman JJ, Whitelaw ML. Asparagine hydroxylation of the HIF transactivation domain a hypoxic switch. Science. 2002;295:858–61.

Lee NP, Chen L, Lin MC, Tsang FH, Yeung C, Poon RT, Peng J, Leng X, Beretta L, Sun S, et al. Proteomic expression signature distinguishes cancerous and nonmalignant tissues in hepatocellular carcinoma. J Proteome Res. 2009;8:1293–303.

Li J, Duncan DT, Zhang B. CanProVar: a human cancer proteome variation database. Hum Mutat. 2010;31:219–28.
Li CY, Wang XL, Wang JH, Yan ZP, Gong GQ, Cheng JM, Chen Y, Liu LX, Li GP, Wang CG, Shi DH. Identifying serum biomarkers for TACE therapy efficiency of hepatocellular carcinoma. Front Biosci (Elite Ed). 2011;3:212–20.
Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc. 2011;19:241–8.
Marouga R, David S, Hawkins E. The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem. 2005;382:669–78.
Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. PRIDE: the proteomics identifications database. Proteomics. 2005;5:3537–45.
Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet. 2005;365:488–92.
Muthusamy B, Hanumanthu G, Suresh S, Rekha B, Srinivas D, Karthick L, Vrushabendra BM, Sharma S, Mishra G, Chatterjee P, et al. Plasma Proteome Database as a resource for proteomics research. Proteomics. 2005;5:3531–6.
Nicholson JK. Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol. 2006;2:52.
Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005;1:252–62.
Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1:376–86.
Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, et al. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005;21:832–4.
Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. Lancet. 2002;359:572–7.
Poon TC, Chan AT, Zee B, Ho SK, Mok TS, Leung TW, Johnson PJ. Application of classification tree and neural network algorithms to the identification of serological liver marker profiles for the diagnosis of hepatocellular carcinoma. Oncology. 2001;61:275–83.
Poon TC, Yip TT, Chan AT, Yip C, Yip V, Mok TS, Lee CC, Leung TW, Ho SK, Johnson PJ. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem. 2003;49:752–60.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004;32:D449–51.
Shen Y, Zhao R, Belov ME, Conrads TP, Anderson GA, Tang K, Pasa-Tolic L, Veenstra TD, Lipton MS, Udseth HR, Smith RD. Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. Anal Chem. 2001;73:1766–75.
Shi J, Zhu L, Liu S, Xie WF. A meta-analysis of case-control studies on the combined effect of hepatitis B and C virus infections in causing hepatocellular carcinoma in China. Br J Cancer. 2005;92:607–12.
Srivastava S, Kramer BS. Early detection cancer research network. Lab Invest. 2000;80:1147–8.
Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2010;39:D698–704.
Sun Y, Mi W, Cai J, Ying W, Liu F, Lu H, Qiao Y, Jia W, Bi X, Lu N, et al. Quantitative proteomic signature of liver cancer cells: tissue transglutaminase 2 could be a novel protein candidate of human hepatocellular carcinoma. J Proteome Res. 2008;7:3847–59.
Takashima M, Kuramitsu Y, Yokoyama Y, Iizuka N, Toda T, Sakaida I, Okita K, Oka M, Nakamura K. Proteomic profiling of heat shock protein 70 family members as biomarkers for hepatitis C virus-related hepatocellular carcinoma. Proteomics. 2003;3:2487–93.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.

Ummanni R, Mundt F, Pospisil H, Venz S, Scharf C, Barett C, Falth M, Kollermann J, Walther R, Schlomm T, et al. Identification of clinically relevant protein targets in prostate cancer with 2D-DIGE coupled mass spectrometry and systems biology network platform. PLoS One. 2011;6:e16833.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003;31:258–61.

Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Med Genomics. 2011;4:2.

Wang Y, Chen J, Li Q, Wang H, Liu G, Jing Q, Shen B. Identifying novel prostate cancer associated pathways based on integrative microarray data analysis. Comput Biol Chem. 2011;35:151–8.

Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, Fiel I, Thung S, Mazzaferro V, Bruix J, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007;45:938–47.

Xu M, Kao MC, Nunez-Iglesias J, Nevins JR, West M, Zhou XJ. An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics. 2008;9 Suppl 1:S12.

Yan Q. The integration of personalized and systems medicine: bioinformatics support for pharmacogenomics and drug discovery. Methods Mol Biol. 2008;448:1–19.

Yokoyama Y, Kuramitsu Y, Takashima M, Iizuka N, Toda T, Terai S, Sakaida I, Oka M, Nakamura K, Okita K. Proteomic profiling of proteins decreased in hepatocellular carcinoma from patients infected with hepatitis C virus. Proteomics. 2004;4:2111–16.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett. 2002;513:135–40.

Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, Bonneau R, Zhang N, Zhou Y, Wollscheid B, et al. UniPep–a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol. 2006;7:R73.

Zhang Y, Wang S, Li D, Zhnag J, Gu D, Zhu Y, He F. A systems biology-based classifier for hepatocellular carcinoma diagnosis. PLoS One. 2011;6:e22426.

Luonan Chen, Ph.D., Professor, China and Japan Luonan Chen received the ME and Ph.D. degrees in the electrical engineering from Tohoku University, Sendai, Japan, in 1988 and 1991, respectively. From 1997, he was an associate professor of the Osaka Sangyo University, Osaka, Japan, and then a full professor. Since 2010, he has been a professor and executive director at Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He was the founding director of Institute of Systems Biology, Shanghai University, and has been a research professor in Institute of Industrial Science, University of Tokyo, Japan.

He serves as chair of Technical Committee of Systems Biology at IEEE SMC Society and as the founding president of Computational Systems Biology Society of ORS China. He serves as editor or editorial board member for major systems biology-related journals, for example, *BMC Systems Biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics, IET Systems Biology, Mathematical Biosciences, Journal of the Royal Society Interface*, and *International Journal of Systems and Synthetic Biology*. His fields of interest are systems

biology, computational biology, and nonlinear dynamics. In recent 5 years, he published over 100 journal papers and two monographs (books) in the area of systems biology.

Bairong Shen, Ph.D., Professor, China Bairong Shen is professor and director in Center for Systems Biology of Soochow University. He received his B.Sc., M.Sc., and Ph.D. degrees in chemistry from Soochow University, East China Normal University, and Fudan University, respectively. Dr. Shen became an associate professor of Physical Chemistry at Fudan University in 1999. Before 1999, Dr. Shen's research is about theoretical and computational surface chemistry; he investigated chemical reactions catalyzed by metals and alloys by ab initio and semiempirical quantum mechanical methods.

After 1999, Dr. Bairong Shen started his research in bioinformatics as a postdoc in University of Tampere, Finland, and then recruited as an assistant professor of Bioinformatics there in the beginning of 2004. From 2006, he worked as guest professor in Tongji University, Shanghai. He joined Soochow University and established a center for systems biology in 2008. Dr. Shen has taught more than ten different courses in bioinformatics and systems biology and published more than 50 peer-reviewed articles in different journals which cover physics, chemistry, biology, and computational science. His recent researches focus on bioinformatics and systems biology of complex diseases.

Jiajia Chen, Ph.D., China Jiajia Chen received her Ph.D. degree from Soochow University and M.Sc. degrees in biology from Nanjing University. Since 2006, she has been a lecturer of Biology at Suzhou University of Science and Technology, Suzhou, China. Dr. Chen's researches focus on systems biology of complex diseases. Her long-term interest is developing systems-level biomarkers for cancer diagnosis and prognosis. In recent 5 years, she has published more than 20 journal papers in the area of systems biology.

Chapter 12
Software Development for Quantitative Proteomics Using Stable Isotope Labeling

Xin Huang and Shi-Jian Ding

Abstract Stable isotope labeling (SIL) coupled with liquid chromatography and high-resolution tandem mass spectrometry (MS) are increasingly useful for elucidation of the proteome-wide differences between multiple biological samples. Developments of more effective programs for the relative peptide/protein abundance measurements are essential for quantitative proteomic analysis. In this chapter, we present a quantification program, termed UNiquant, for analyzing quantitative proteomic data using SIL. The common steps in a quantitative proteomic software, such as MS data preprocessing, peptide identification, peptide quantification, and protein quantification, were dissected in this chapter, using UNiquant as an example. UNiquant was used to analyze the SILAC-labeled proteome mixtures with known heavy/light ratios (H/L = 1:1, 1:5, and 1:10). The pros and cons of the quantification results of UNiquant from two different MS acquisition modes, data-dependent acquisition and data-independent acquisition, were also evaluated and compared.

Keywords Software development • UNiquant • Stable isotope labeling • Mass spectrometry • Quantitative proteomics • Data-dependent acquisition • Data-independent acquisition

X. Huang, Ph.D.
Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

S.-J. Ding, Ph.D. (✉)
Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

Mass Spectrometry and Proteomics Core Facility, University of Nebraska Medical Center, Omaha, NE, USA
e-mail: dings@unmc.edu

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_12, © Springer Science+Business Media Dordrecht 2013

12.1 Introduction

Mass spectrometry-based quantitative proteomics is an emerging field capable of making a unique contribution to the understanding, prevention, and cure of human diseases (Choudhary and Mann 2010; Gstaiger and Aebersold 2009; Koomen et al. 2008). Proteomic analysis now involves larger and more reliable datasets, mostly generated using state-of-the-art mass spectrometry (MS) combined with a bottom-up (or shotgun) profiling of whole protein complements from cells, tissues, and body fluids (Mann and Kelleher 2008). Proteomics has an advantage over genomic-based assays because it offers direct examination of the molecular machinery of cell physiology, including protein expression, cell signaling, and posttranslational modifications (PTMs).

A major hurdle in quantitative proteomics is still identifying and subsequent quantifying of proteins and their expression levels in complex biological systems (Venable et al. 2004). In quantitative shotgun proteomics, proteolysis-derived peptides are commonly measured with LC-MS/MS and are used as surrogates of their parent proteins for relative quantification (Mann and Kelleher 2008; Ong and Mann 2005). In a label-free approach, proteomes under comparison are analyzed separately in standardized LC-MS/MS runs. Peptide intensities, spectra counts, and extracted ion chromatography (XIC) are used to measure the protein abundances (Fang et al. 2006; Finney et al. 2008). Alternatively, by employing stable isotope labeling (SIL), the proteomes under comparison are combined and analyzed together in one LC-MS/MS run. Comparison of the signal intensities of the same peptides and their SIL analogues yields an estimate of protein abundances (Geiger et al. 2010a; Mann 2006). In general, SIL methods minimize variability during sample processing steps and LC-MS/MS analyses and provide results with less systematic error and higher reproducibility compared to the label-free approach (Qian et al. 2010). On the other hand, absolute quantification of proteins can be obtained through the use of a stable isotope-labeled internal standard (Silva et al. 2006a).

Development of software for quantitative proteomics with SIL has made tremendous advances in this field. A number of academically developed software tools, such as ASAPRatio (Li et al. 2003), ProRata (Pan et al. 2006), RelEx (Venable et al. 2004), Xpress (Han et al. 2001), Census (Park et al. 2008), MaxQuant (Cox and Mann 2008), Vista (Bakalarski et al. 2008), WaveletQuant (Mo et al. 2010), UNiquant (Huang et al. 2011a, b), and recently IsoQuant (Liao et al. 2012), have been produced to analyze SIL-based quantitative proteomic datasets. In these experiments, information on peptide abundance is derived either from the intensity of the peptide precursor ion signal at full mass spectra or from the intensity of reporter ions after MS/MS fragmentation. The first catalog includes isotope-coded affinity tagging (ICAT), stable isotope labeling with amino acids in cell culture (SILAC), and $^{16}O/^{18}O$ labeling, while the second catalog includes tandem mass tags (TMT) and isobaric tag for relative and absolute quantitation (iTRAQ). Most of the programs target either precursor ion or reporter ion quantitation. For the precursor ion-based quantitation, low-intensity MS signals present a substantial challenge to

quantify low-abundance proteins by various programs (Bakalarski et al. 2008). Different programs adopt different strategies for distinguishing the peptide signals from the background noise. Using a data-dependent acquisition (DDA) mode, MS/MS fragmentation is performed on the most abundant precursor ions. On hybrid high-resolution mass spectrometry such as LTQ Orbitrap (Thermo Scientific, San Jose, CA), the precursor scan is performed in an Orbitrap analyzer, and the MS/MS fragmentation is usually accomplished in the linear ion trap mass analyzer. In these experiments, an LC-MS/MS data collection cycle starts with a high-accuracy, full MS survey of the all precursor ions and is followed by selecting a number of the intensive precursor ions for MS/MS fragmentation (Wilm 2009). Recently, the data-independent acquisition (DIA) strategy was developed to complement the DDA method for proteomic analysis (Venable et al. 2004). Instead of a serial selection of precursor ions for data-dependent fragmentation, the DIA approach fragments a group of co-eluting precursor ions at each given time, enabling a more unbiased detection of all LC-eluted peptides compared to the DDA method (Ramos et al. 2006; Williams et al. 2003).

Here, we are going to describe the in-house developed UNiquant software for quantitative proteomic MS data analysis with SIL. The major procedures in a quantitative software, including MS data preprocessing, detecting pairs, reading intensity, normalization, and performance and compatibility at different MS platforms, will be introduced by analyzing SILAC-labeled eukaryotic cells with known heavy-versus-light ratios.

12.2 Materials and Methods

12.2.1 Prepare SILAC Protein Mixture with Known Ratios

The human cell lines Jeko-1 and MDA-MB-231 cells were grown in either SILAC "light" (L-arginine and L-lysine) or "heavy" (L-$^{13}C_6$-arginine and L-$^{13}C_6$-lysine for Jeko-1, L-[$^{13}C_6$,$^{15}N_4$]-arginine and L-[$^{13}C_6$,$^{15}N_2$]-lysine for MDA-MB-231) medium for 2 weeks (more than five cell cycles). The heavy and light lysates were harvested mixed in three heavy/light (H/L) ratios: 1:1, 1:5, and 1:10, then followed by sample pretreatments and tryptic digestion (Huang et al. 2011a, b).

12.2.2 LC-MS/MS Analysis with DDA and DIA

In the DDA analysis, the LTQ Orbitrap mass spectrometer automatically switches between MS and MS/MS acquisition modes. In each MS cycle (about 2.5 s), a survey full-scan MS spectra (m/z 375–1,575) were acquired in the Orbitrap with resolution $R = 100{,}000$, then the most five intense ions (depending on signal intensity

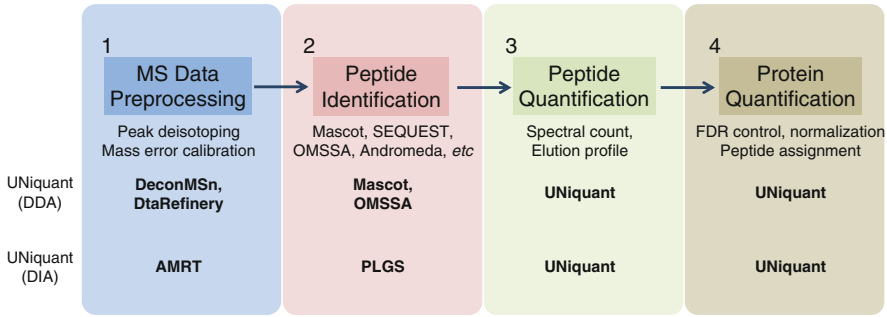


Fig. 12.1 The common steps in a quantitative proteomic software and the components in UNiquant for DDA and DIA

of survey full scan) were sequentially isolated for fragmentation in the linear ion trap using collision-induced dissociation (CID). Former target ions selected for MS/MS were dynamically excluded for 75 s. In the DIA analysis, the SYNAPT G2 mass spectrometer (Waters Co., Milford, MA) equipped with time-of-flight (TOF) analyzer was used. The Tri-Wave ion guides trap and separates precursor ions by ion mobility. Then, the CID cell was operated alternatively with low-energy and elevated energy survey of acquisitions (Bateman et al. 2002). The acquisition time in each mode was 1.0 s with an interscan delay of 0.1 s. In the low-energy mode for the survey full MS scan (*m*/*z* 300–2,000), precursor intensities were collected at constant collision energy (5 eV). In the elevated energy mode for the MS^E scan (MS/MS scan), collision energy was ramped from 15 to 40 eV during each collection cycle.

12.3 UNiquant Software for Quantitative Proteomics

12.3.1 Overview of Quantitative Proteomic Software

Protein identification and quantification are the two major components for a quantitative proteomic program. However, a quantitation software also needs other program components. Figure 12.1 shows four steps that a quantitation program usually involved.

Step 1: MS data preprocessing. Initially, the vender-specific MS raw data need to be converted to the text-formatted peak list files such as dta or mgf files or a common file format using extensible markup language (XML), such as mzXML and mzML.

Step 2: peptide identification. In this step, peptide is identified from the MS/MS peak list through a process of peptide-spectrum match (PSM) using programs such as SEQUEST (Eng et al. 1994), Mascot (Perkins et al. 1999), OMSSA (Geer et al. 2004), and Andromeda (Cox et al. 2011) to compare

the observed peak list to a protein database. Identification by these algorithms is based on a restricted database search in which MS/MS spectra are aligned with protein sequences, probably bearing a few specified PTMs attached to specific amino acids.

Step 3: peptide quantification. In label-free methods, the spectra count or the normalized XIC profiles of a peptide were measured as the intensity of identified peptide. In SIL methods, quantification programs fetch the XIC elution profiles of the heavy and light peptides from the MS raw data (or XML data) in full MS scans (SILAC, ICAT, $^{16}O/^{18}O$, etc.) according to the identified peptide sequences in the same LC-MS/MS runs. For MS/MS scan (iTRAQ, TMT)-based quantification, intensity of reporter ions is measured within the same MS/MS spectrum that a peptide was identified.

Step 4: protein quantification. The quantification results at the protein level are reported by assigning the peptide sequences to different protein IDs. To ensure the confidence of the identification and quantification results, a false discovery rate (FDR) method is used to estimate the false-positive results in the final reports. FDR should be monitored and reported (usually 0.01) at the spectrum, peptide, and proteins levels. Furthermore, peptide intensities are normalized (if needed) to reduce the deviation of quantified ratios if the heavy and light samples are not equally mixed. In a user's view, the result of a quantitative proteomic experiment is a report of thousands of proteins, with their intensities or relative intensity ratios between the heavy and light species.

12.3.2 UNiquant Software for Quantitative Proteomics with DDA

To date, most of the quantitative proteomic data were obtained using the method. In these experiments, an LC-MS/MS data collection cycle starts with a high-accuracy, full MS survey of the all precursor ions and is followed by selecting a number of precursor ions for MS/MS fragmentation (Wilm 2009). A major advantage of DDA is that the fragment ions are derived mostly from a single precursor ion, increasing the specificity of peptide identification. As shown in Fig. 12.1, UNiquant chooses the third-party academic-free softwares DeconMSn and DtaRefinery for MS data preprocessing. The DDA version of UNiquant uses the identification results of Mascot and OMSSA (open source and freely available for academic users) search engines. Finally, UNiquant program is developed for peptide and protein quantification using the outputs from first two steps.

12.3.3 Data Preprocessing and Peptide Identification

In UNiquant, Thermo MS raw data are converted to mgf-formatted MS/MS peak list file before Mascot search. DeconMSn (http://omics.pnl.gov/software/) was used to determine and refine the monoisotopic mass and charge state of parent ions and

to create the peak list files. Next, DtaRefinery (http://omics.pnl.gov/software/) is used to improve mass measurement errors for parent ions by modeling systematic errors based on putative peptide identifications. For the SILAC protein mixture data used in this study, Mascot search engine was used for peptide identification. To ensure the quality of the identification results, usually a "target-decoy" database search strategy (Elias and Gygi 2010) was applied by searching against a concatenated database containing the authentic protein sequences (forward database) and the reverse sequences of all proteins involved (reverse database). Then, an FDR estimator is calculated to access the confidence of the identification results. Previous studies have shown that the PSM score given by the search engines and mass accuracy of the precursors are two important parameters for discriminating the forward and reverse identifications (Ding et al. 2008).

In UNiquant coupled with Mascot as the search engine, quality of peptide identification (QPI) score is calculated by

$$\mathrm{QPI} = s \times e^{-1/2} \tag{12.1}$$

where s is Mascot peptide identification score and e is the mass error (ppm) of the precursor ions which is calculated as

$$e = 1{,}000{,}000 \times \frac{(m_{\text{observed}} - m_{\text{theoretical}})}{m_{\text{theoretical}}} \tag{12.2}$$

where m_{observed} and $m_{\text{theoretical}}$ are the observed accurate mass and theoretical mass of the precursor ions of the peptide, respectively. A Mascot score cutoff of 10 was applied for all identification results. QPI of a peptide was taken as the sum of the QPI for all MS/MS spectra that were matched to this peptide sequence. Identified peptides were sorted by a descending order of QPI values, and a cutoff was applied to ensure a total FDR < 0.01.

12.3.4 Intensity Measurement of Precursor Ions

Precursor ion intensity, measured in the high-resolution full MS, was extracted by UNiquant and used as an abundance measurement for each identified peptide. The input files for quantitation are the Thermo Xcalibur MS data (.raw) and the peptide identification output dat (Mascot) and csv (OMSSA) files. UNiquant also utilizes the search results from other search engines with text-formatted outputs containing the filtered peptide sequence, identification score, scan number, observed m/z, and charge state information. The quantitation algorithm in UNiquant was first developed for hybrid FT-MS instruments (Ding et al. 2008). Briefly, theoretical mass for a peptide (labeled or unlabeled) is calculated according the peptide sequence identified in the MS/MS spectrum and the SIL method. Then, the corresponding high-accuracy, full MS scan which derives the MS/MS spectrum is determined. A search is performed on this MS spectrum within a small range (<20 ppm)

to localize the heavy and light precursor ions. Intensities of both precursor ions are measured with a signal-to-noise (S/N) ratio above 2.0. The output of UNiquant is a tab-delimited text file which includes a list of peptides with refined *m/z*, mass error, S/N ratios, and intensities of light and heavy species.

12.3.5 Peptide and Protein Quantification

A peptide usually appears more than once in the LC-MS/MS data. The spectra count is the number of times that a peptide identified by database search. The relative abundance of each identified peptide was calculated as the sum (based on spectra counts) of peak intensities (PI) for the heavy species of the peptide divided by the sum of intensities for the light species of the peptide:

$$\text{Ratio}_{\text{H/L}} = \frac{\sum_n \text{PI}_{\text{H}}}{\sum_n \text{PI}_{\text{L}}} \tag{12.3}$$

where n is the spectra count for a specific peptide, PI_{H} is the peak intensity of the heavy species, and PI_{L} is the peak intensity of the light species. Similarly, the relative abundance of each identified protein was calculated by dividing the sum of the intensities of all peptide heavy species for the protein by the sum of the intensities of all peptide light species.

12.3.6 Post-measurement Normalization

The post-measurement normalization is needed for correcting the unequal mixing of heavy and light proteins in the quantitative proteomic experiments. In UNiquant, a locally weighted scatterplot smoothing (LOWESS) method was used to correct the H/L ratios of quantified peptides (Cleveland 1979). Briefly, LOWESS method is based on minus-add (M-A) plot of the peptide intensities for the heavy and light species:

$$M = \log_2 \frac{\left(\text{Int}_{\text{heavy}}\right)}{\left(\text{Int}_{\text{light}}\right)} \tag{12.4}$$

$$A = \frac{1}{2}\log_{10}\left(\text{Int}_{\text{heavy}} \times \text{Int}_{\text{light}}\right) \tag{12.5}$$

where $\text{Int}_{\text{heavy}}$ is the intensity of the heavy species from a quantified peptide, while $\text{Int}_{\text{light}}$ is the intensity of the corresponding light species of this peptide. M is the $\log_2$ H/L intensity ratio, and A is half of the $\log_{10}$ H × L intensity product of each quantified peptide. These *M*–*A* points were equally divided into 20 groups, based on their A-values. A linear regression line was obtained from the points in each group, and

then a fitted regression curve was obtained by connecting all the regression lines. Normalization was performed by subtracting the fitted curve from the measured $\log_2$ H/L ratio in the *M–A* plot:

$$M' = \log_2\left(\frac{\text{Int}_{\text{Heavy}}}{\text{Int}_{\text{light}}}\right) - c(A) = \log_2\left(\frac{\text{Int}_{\text{Heavy}}}{k \times \text{Int}_{\text{light}}}\right) \tag{12.6}$$

where $c(A)$ is the fitted LOWESS curve, which is a function of A. M' is the normalized log ratio of quantified peptides, which is obtained by subtracting the value of LOWESS fitting function from the measured log ratio at each value of A.

12.3.7 UNiquant Software for Quantitative Proteomics with DIA

Recent development of the DIA strategy has been introduced as a complement methodology of the DDA strategy for quantitative proteomic experiments. It has been implemented on two MS platforms: Exactive Orbitrap (Thermo Scientific) and SYNAPT G2 (Waters Co). The corresponding DIA method was named as all-ion fragmentation (Geiger et al. 2010b) and LC-MS^E technology (Silva et al. 2006b; Vissers et al. 2009), respectively. UNiquant was recently developed for analyzing the proteomic data on the LC-MS^E platform (Huang et al. 2011b). As shown in Fig. 12.1, UNiquant also covers the last two steps of peptide and protein quantification of the DIA proteomic data, while the MS raw data preprocessing and peptide identification are performed by the ProteinLynx Global Server (PLGS, Waters Co.) software. In the first step, ion detection, clustering, and retention time alignment are processed using an AMRT (accurate mass retention time) method in PLGS (Silva et al. 2005). Next, the AMRT data are searched against the Swiss-Prot protein database using a dual-pass algorithm in PLGS (Li et al. 2009).

Procedures for quantification of the LC-MS^E DIA data are similar to the procedures in the DDA approach. The AMRT files are exported to a local Microsoft Access database. Included in this output are the weight-averaged monoisotopic mass, charge state, ion drift, charge-state-reduced sum intensity, observed apex retention time, and observed start and stop time of the detected ions. Information for the identified peptides is exported to a table file as well. This contains all the theoretical and experimental properties associated with the identified precursor MS spectrum, such as the unique spectrum id, mass over charge (m/z), retention time, peptide sequence, and the identification scores. Theoretical masses of the heavy and light precursors are determined from the peptide sequence and the SIL method. The predicted precursors are used to search the AMRT database for an observed ion that matched the criteria of mass accuracy, elution time, and ion drift. The default settings were mass accuracy <5 ppm, difference in retention time <0.05 min, and difference in ion drift ≤0.5. Intensities of the SIL pair of precursors are extracted, and the heavy/light ratios were sorted and arranged with the peptide sequence and protein entry. Similar to DDA method, the relative H/L ratios of identified peptide are calculated as the

sum of the intensities for the heavy precursors divided by the sum of the intensities for the light precursors.

12.4 Results and Notes

12.4.1 Implementation of UNiquant

UNiquant is an in-house software for the quantitative proteomic data analysis with SIL. The software is developed on the platform of Microsoft .NET Framework (version 2.0). The programming languages are Microsoft VB.NET and C#. It now has two components, for analyzing the DDA and DIA data, respectively. As shown in Fig. 12.2, the DDA version of UNiquant has incorporate the softwares of DeconMSn and DtaRefinery for MS raw data deisotoping and mass error calibration, respectively. Next, the OMSSA database search engine was embedded in UNiquant as the default engine for peptide identification. UNiquant is also compatible with other engines such as Mascot. But the users need to upload the mgf files to the Mascot Daemon Server (Matrix Science) for peptide identification and obtain the output dat files for further quantification. Peptide quantification is performed in the component named "precursor search," fetching the information of precursor intensity from the MS raw data by the Xcalibur Development Kit (XDK) provided by the instrument vendor. Precursor search can be performed for individual LC-MS runs by a user-friendly "drag and drop" process or automatically performed by the UNiquant piper. Finally, intensities of the peptides from different LC-MS runs are merged and

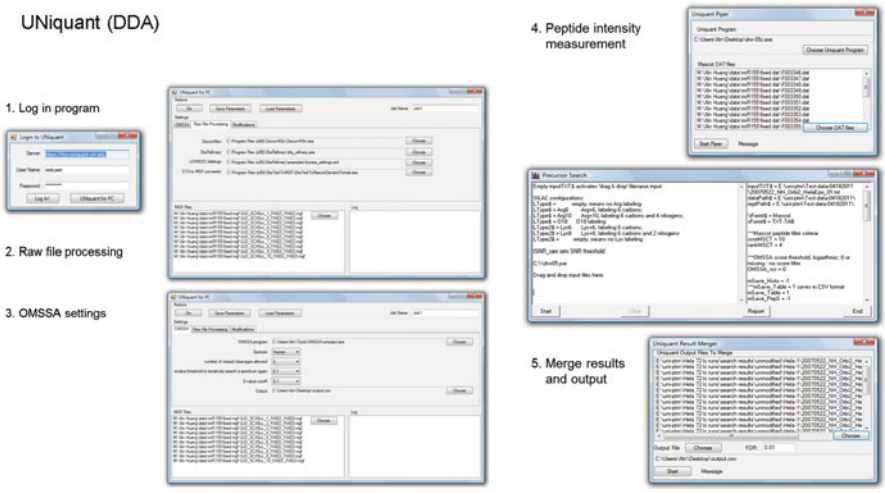


Fig. 12.2 Implementation of UNiquant for *DDA* data analysis. The program components and user interface (*UI*) in different steps are indicated as well

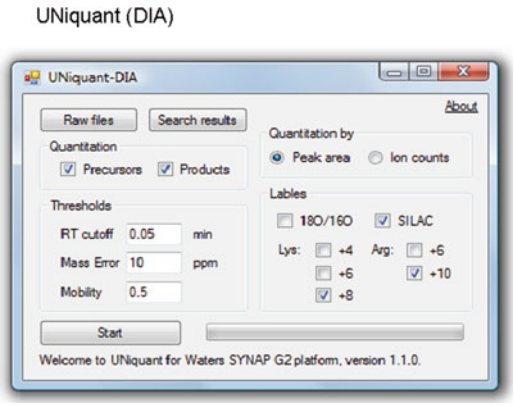


Fig. 12.3 Implementation of UNiquant for DDA data analysis

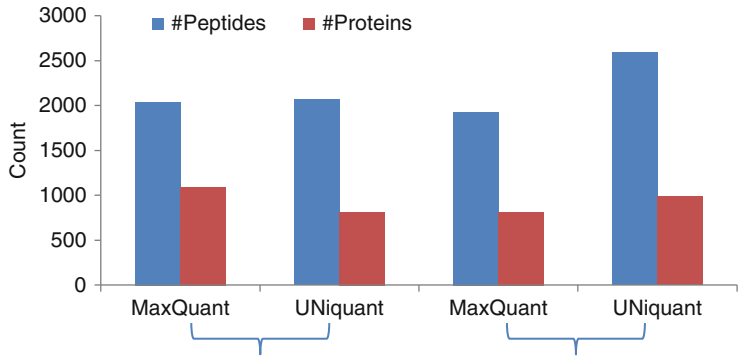
filtered by an FDR threshold and annotated by unique protein ID for protein level outputs of the quantification results.

The user interface of UNiquant for DIA is shown in Fig. 12.3. Here, the UNiquant program reads the deconvoluted MS raw files (AMRT) and peptide identification results (with fixed FDR confidence) from the Waters PLGS software. UNiquant outputs all the peak lists into a local Microsoft Access database, and searching of the heavy and light precursors was performed by the SQL queries with Microsoft Office Development in Visual Studio (ODVS) components. Matching of the SIL heavy and light precursor ions is performed based on the similarity of retention time (default setting, <0.05 min) and ion mobility (<0.5) and the accurate mass (<10 ppm) of difference between the heavy and light precursors. Finally, the DIA version of UNiquant summarizes the quantification results and output the protein level H/L ratios.

12.4.2 Analysis of the SILAC Proteome Mixture with Known H/L Ratios

We analyzed the SILAC-labeled proteome digests with known H/L ratios (H/L = 1:1 and 1:10). For identification of the peptides/proteins, we used the same database search engine (Mascot), with identical searching parameters, and searched the data against the same proteome database (IPI version 3.52). Using the same FDR cutoff of 0.01, the number of peptide pairs and proteins being identified by each program is shown in Fig. 12.4. UNiquant and MaxQuant identified nearly equal numbers of peptide pairs in the H/L = 1:1 mixture data. For the H/L = 1:10 proteome data, UNiquant identified 34 % more peptide pairs and proteins than MaxQuant. However, the number of quantified proteins is similar for these two programs.

Before normalizing of the quantification results, the median log ratio of peptides quantified by UNiquant was generally equal to the true value of the log ratio in each proteome mixtures (Fig. 12.4). In the H/L = 1:1 proteome mixture, the median log ratio of peptides quantified by UNiquant was −0.029, which is closer to the true log



H/L =	1:1 (true log-ratio = 0)			1:10 (truelog-ratio = -1)		
Peptides	N	Median	SD	N	Median	SD
MaxQuant	2038	-0.113	0.096	1929	-1.229	0.189
UNiquant	2063	-0.029	0.176	2591	-0.968	0.393

Fig. 12.4 Peptide and proteins quantified by MaxQuant and UNiquant for the standard SILAC mixtures (H/L = 1:1 and 1:10). The true values of $\log_{10}$ H/L ratios are indicated, and the statistics of quantified peptides for each mixture were tabulated in the following table

ratio = 0, compared to −0.113 obtained by MaxQuant. Similarly, the median ratio log ratios quantified by UNiquant and MaxQuant are −0.968 and −1.229, respectively, for the H/L = 1:10 proteome mixture (true log ratio = −1). The frequency of the log ratios quantified by both UNiquant and MaxQuant is generally Gaussian distributed in all mixture data, but with different variances. The standard deviation of the log ratios quantified by MaxQuant is lower than the log ratios quantified by UNiquant.

Different programs provided complementary results of quantified proteins. UNiquant and MaxQuant chose different strategies for SIL-pair detection. By the scenario of DDA, the selected isotopic peaks for MS/MS fragmentation can be derived from either light or heavy peptides. UNiquant does not detect SIL pair of peptide before identification. After database search, theoretical masses for both the heavy and light peptides were determined, and intensities were calculated based on the confident identifications. This strategy was also applied by other programs such as Vista and IsoQuant. In contrast, MaxQuant uses an alternative strategy for peak pair detection, one which identifies pairs of light and heavy peptides from the MS data prior to peptide identification (Cox and Mann 2008). Advantages of the strategy in MaxQuant are that the resulting peak list is much cleaner than the peak list from the original raw data, the peaks have a high S/N ratio, and co-eluting peptides can be readily identified. However, this strategy may result in some loss of pairs, especially in the case of peptide pairs with low-intensity or high-noise background. Such as the case of H/L = 1:10 data, MaxQuant quantified more

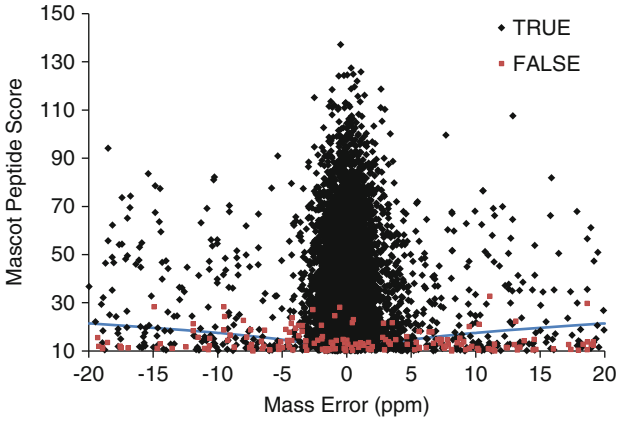


Fig. 12.5 The distribution of Mascot peptide identification score and mass error in the H/L = 1:1 mixture. Peptides from the forward database were labeled as *black points*, and peptides from backward database were labeled as *red points*. The *blue curve* indicates the QPI cutoff. *Points* under the cutoff were removed from the quantitation

peptides than MaxQuant but increases the variance of the quantified results as the compensation.

Furthermore, the way for calculating FDR is slightly different between UNiquant and MaxQuant. MaxQuant corrects the mass precision of precursors and used a posterior error probability based on the peptide *P*-score distribution by different categories of peptide length to set the cutoff of FDR (Cox and Mann 2008; Olsen and Mann 2004). UNiquant used QPI (Eq. 12.1), an estimator involving the peptide identification score and mass error of the precursor ions. Figure 12.5 shows the distribution of the Mascot peptide score versus the mass error of precursor ions for our SILAC datasets. The false peptides have lower peptide score and higher mass errors compared to the true peptides. The blue line in Fig. 12.4 shows the QPI cutoff in this dataset to remove all the low-confidence identifications.

12.4.3 Post-measurement Normalization

We plotted the quantification results of standard SILAC mixture data with DDA in Fig. 12.6. Before normalization, the identified and quantified peptides from all three proteome mixtures with known ratios (H/L = 1:1, 1:5, and 1:10) are plotted by their $\log_2$ (H/L) intensity ratios versus the $\log_{10}$ (H × L) intensity products (Fig. 12.6a–c). The $\log_2$ (H/L) ratios of quantified peptides show a comet-like distribution from the three mixtures. The region of low-abundance peptides generally has higher variance $\log_2$ (H/L) ratios compared to the high-abundance peptides. In H/L = 1:5 and 1:10 mixtures, the data in low-intensity region tends to a log ratio of 0, whereas they should have a value of −2.32 for the 1:5 data and a value of −3.32 for the 1:10 data.

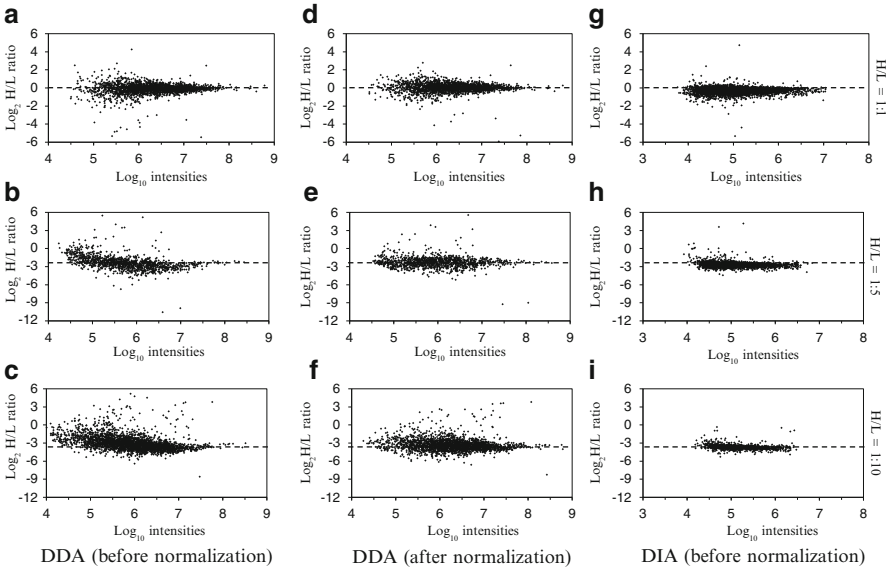


Fig. 12.6 Accuracy of quantitation for SILAC data analysis is compared between the DDA analysis on LTQ Orbitrap platform (**a**–**c**) before and (**d**–**f**) after normalization and the DIA analysis on (**g**–**i**) SYNAPT G2 platform for three proteome mixtures with H/L = 1:1 (**a**, **d**, **g**), 1:5 (**b**, **e**, **h**), and 1:10 (**c**, **f**, **i**). In each scatterplot, the quantified peptides were distributed by their $\log_2$ (H/L) intensity ratios versus $\log_{10}$ (H × L) intensity products. The true $\log_2$ (H/L) ratio is indicated as a *dashed line* for the H/L = 1:1 ($\log_2$ ratio = 0), 1:5 ($\log_2$ ratio = −2.32), and 1:10 ($\log_2$ ratio = −3.32) mixtures, respectively

The LOWESS method corrects and straightens the median ratios of quantified peptides by different categories of intensity and straightens the LOWESS regression curve into a straight line. As shown in Fig. 12.6d–f, the log ratios of peptides in the H/L = 1:1 mixture were similar between and after normalization. But the log ratios of the low-abundance peptides in the H/L = 1:5 and 1:10 mixtures are corrected to the true ratios of −2.32 (H/L = 1:5) and −3.32 (H/L = 1:10), respectively. In the contract, the log ratios of the high-abundance peptides do not change too much in these two mixtures.

In quantitative proteomic data analysis, a normalization approach is usually applied by assuming that the amounts of most proteins in the sample will be unchanged by the variable being tested. The purpose of normalization is to overcome the effects of unequal mixing of the heavy and light species during the sample preparation. Thus, the averaged heavy/light abundances of all the quantified proteins can be adjusted to one. In MaxQuant and UNiquant, the relative peptide/protein abundances before and after normalization are both provided (Cox and Mann 2008). MaxQuant uses the median-center method for normalization, but UNiquant uses the LOWESS method. However, this normalization approach may not be applied in some cases especially if specific portion of proteins are enriched, such as the phosphoproteins (by molecular function) or the nuclear proteins (by cellular components). For instance,

the use of phosphatase inhibitors will affect a broad range of cellular phosphorylation events. Therefore, the assumption of normalization is not valid if only the phosphoproteome was investigated. Furthermore, the assumption that the majority of proteins are unchanged might be incorrect when the specific treatment could affect a broad range of protein concentration, such as transcription factor and microRNA. So a quantitative proteomic solution for accurate relative protein abundance measurement is still necessary.

12.4.4 Comparison of the Quantification Results from DDA and DIA

Peptide quantitation results obtained on the SYNAPT G2 MS with DIA and on the LTQ Orbitrap MS with DDA were compared. In the SYNAPT G2 analysis, $\log_2$ (H/L) ratios of quantified peptides show a more uniform distribution for each of the three mixtures (Fig. 12.6g–i). In the H/L = 1:5 and 1:10 mixtures, the $\log_2$ (H/L) ratios are closer to the expected ratios (−2.32 and −3.32). In the H/L 1:1 mixtures, the dynamic ranges ($\log_{10}$ intensities) of both the LTQ Orbitrap data and the SYNAPT G2 data are about 4 orders of magnitude (Fig. 12.6d, g). In H/L = 1:5 and 1:10 mixtures, the dynamic range of LTQ Orbitrap data is still 4 orders of magnitude (Fig. 12.6e, f), whereas the range of the data from the SYNAPT G2 drops to 3.5 orders of magnitude in the H/L = 1:5 mixture and to 3.0 in the 1:10 mixture (Fig. 12.6h, i).

Currently, the LTQ-FT/Orbitrap MS with DDA is the major MS platform for SIL-based quantitative proteomic applications. With this platform, the MS scans are used for peptide quantitation, while MS/MS scans are used for peptide identification. Because the number of precursor ions selected for MS/MS fragmentation is fixed in the DDA mode, the total number of peptides identified for a given protein from complex proteome mixtures is relatively low due to the limited number of MS/MS spectra that can be generated for peptide identification. The DIA approach is an interesting alternative to complement DDA for SIL-based quantitative proteomic analysis. First, it provides more time for MS/MS fragmentation, making it possible to identify more peptides. Second, the high-resolution MS and MS/MS data makes quantitation possible from both precursor and product ions.

Dynamic range for protein quantitation is one of the key features of quantitative proteomic analysis. Using the DIA method, we identified more proteins in the 1:5 and 1:10 mixtures compared to the 1:1 mixture; however, the number of SIL-peptide pairs and the dynamic range of protein intensities decreased in the 1:5 and 1:10 mixtures. The decrease in peptide pairs and dynamic range of protein intensities is mainly due to the loss of low-intensity heavy peptides and to saturation of high-intensity light peptides. That occurred because the peak detection algorithm used a cutoff for acceptable peaks that was based on peak intensity and signal-to-noise ratio. These excluded small peaks which impacted the number of heavy peptides observed but ensured that the selected peaks were actually peptides. Additionally,

the dynamic range of protein intensities was consistent (about 4 orders of magnitude) in the DDA analysis of the three complex proteome mixtures with different ratios.

Accurate quantitation of protein abundance is an essential task for MS instruments and its associated data analysis tools. Overall, the SYNAPT G2 with DIA approach showed better quantitation accuracy and reliability than the LTQ Orbitrap with DDA analysis presumably due to the fundamental difference between these two mass analyzers (Pan et al. 2006; Bakalarski et al. 2008). In a TOF analyzer such as the SYNAPT G2, the signal intensity comes from direct ion counting, and many spectra are accumulated up to 10,000 specs per second. Each TOF spectrum usually has a small dynamic range, and a collection of multiple spectra can increase it. If the TOF analyzer is optimized for high sensitivity such as in the case of this study, the SYNAPT G2 instrument gives correct intensity measurements for low-intensity ions but saturated readings for high-intensity ions. Thus, the very high-intensity ions are discriminated against in the final results because the saturated ions increase internal errors of both measured intensity and mass accuracy. In the Orbitrap analyzer, signal intensities are obtained by Fourier transformation of an ion signal induced on the detection electrodes of the Orbitrap cell (Hu et al. 2005). Just one ion signal spectrum is sufficient to obtain a full m/z spectrum with a high dynamic range in ion intensities. In addition to these signal detection differences, the front part ion optics for the two instruments is distinct. The SYNAPT G2 uses a stack ring ion guide, while the LTQ Orbitrap uses a linear quadrupole. These differences produce different ion intensity scales. Together, these differences may explain the difference in the quantification results obtained by the SYNAPT G2 and LTQ Orbitrap platforms.

References

Bakalarski CE, Elias JE, Villen J, Haas W, Gerber SA, Everley PA, Gygi SP. The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. J Proteome Res. 2008;7:4756–65.

Bateman RH, Carruthers R, Hoyes JB, Jones C, Langridge JI, Millar A, Vissers JP. A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. J Am Soc Mass Spectrom. 2002;13:792–803.

Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11:427–39.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Statist Assoc. 1979;74:829–36.

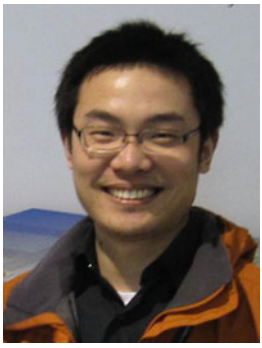
Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10:1794–805.

Ding SJ, Wang Y, Jacobs JM, Qian WJ, Yang F, Tolmachev AV, Du X, Wang W, Moore RJ, Monroe ME, Purvine SO, Waters K, Heibeck TH, Adkins JN, Camp 2nd DG, Klemke RL, Smith RD. Quantitative phosphoproteome analysis of lysophosphatidic acid induced chemotaxis applying

dual-step (18)O labeling coupled with immobilized metal-ion affinity chromatography. J Proteome Res. 2008;7:4215–24.
Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol. 2010;604:55–71.
Eng JK, Mccormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectr. 1994;5:976–89.
Fang R, Elias DA, Monroe ME, Shen Y, McIntosh M, Wang P, Goddard CD, Callister SJ, Moore RJ, Gorby YA, Adkins JN, Fredrickson JK, Lipton MS, Smith RD. Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. Mol Cell Proteomics. 2006;5:714–25.
Finney GL, Blackler AR, Hoopmann MR, Canterbury JD, Wu CC, MacCoss MJ. Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. Anal Chem. 2008;80:961–71.
Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004;3:958–64.
Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods. 2010a;7:383–5.
Geiger T, Cox J, Mann M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics. 2010b;9:2252–61.
Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet. 2009;10:617–27.
Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol. 2001;19:946–51.
Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. J Mass Spectrom. 2005;40:430–43.
Huang X, Tolmachev AV, Shen Y, Liu M, Huang L, Zhang Z, Anderson GA, Smith RD, Chan WC, Hinrichs SH, Fu K, Ding SJ. UNiquant, a program for quantitative proteomics analysis using stable isotope labeling. J Proteome Res. 2011a;10:1228–37.
Huang X, Liu M, Nold MJ, Tian C, Fu K, Zheng J, Geromanos SJ, Ding SJ. Software for quantitative proteomic analysis using stable isotope labeling and data independent acquisition. Anal Chem. 2011b;83:6971–9.
Koomen JM, Haura EB, Bepler G, Sutphen R, Remily-Wood ER, Benson K, Hussein M, Hazlehurst LA, Yeatman TJ, Hildreth LT, Sellers TA, Jacobsen PB, Fenstermacher DA, Dalton WS. Proteomic contributions to personalized cancer care. Mol Cell Proteomics. 2008;7:1780–94.
Li XJ, Zhang H, Ranish JA, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem. 2003;75:6648–57.
Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics. 2009;9:1696–719.
Liao Z, Wan Y, Thomas SN, Yang AJ. IsoQuant: a software tool for stable isotope labeling by amino acids in cell culture-based mass spectrometry quantitation. Anal Chem. 2012;84: 4535–43.
Mann M. Functional and quantitative proteomics using SILAC. Nat Rev Mol Cell Biol. 2006;7:952–8.
Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci U S A. 2008;105:18132–8.
Mo F, Mo Q, Chen Y, Goodlett DR, Hood L, Omenn GS, Li S, Lin B. WaveletQuant, an improved quantification software based on wavelet signal threshold de-noising for labeled quantitative proteomic analysis. BMC Bioinformatics. 2010;11:219.

Olsen JV, Mann M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc Natl Acad Sci U S A. 2004;101:13417–22.
Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005;1:252–62.
Pan C, Kora G, McDonald WH, Tabb DL, VerBerkmoes NC, Hurst GB, Pelletier DA, Samatova NF, Hettich RL. ProRata: a quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. Anal Chem. 2006;78:7121–31.
Park SK, Venable JD, Xu T, Yates 3rd JR. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods. 2008;5:319–22.
Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.
Qian WJ, Petritis BO, Kaushal A, Finnerty CC, Jeschke MG, Monroe ME, Moore RJ, Schepmoes AA, Xiao W, Moldawer LL, Davis RW, Tompkins RG, Herndon DN, Camp DG, Smith RD. Plasma proteome response to severe burn injury revealed by (18)O-labeled "universal" reference-based quantitative proteomics. J Proteome Res. 2010;9:4779–89.
Ramos AA, Yang H, Rosen LE, Yao X. Tandem parallel fragmentation of peptides for mass spectrometry. Anal Chem. 2006;78:6391–7.
Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S. Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem. 2005;77:2187–200.
Silva JC, Denny R, Dorschel C, Gorenstein MV, Li GZ, Richardson K, Wall D, Geromanos SJ. Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome: a sweet tale. Mol Cell Proteomics. 2006a;5:589–607.
Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics. 2006b;5:144–56.
Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004;1:39–45.
Vissers JP, Pons S, Hulin A, Tissier R, Berdeaux A, Connolly JB, Langridge JI, Geromanos SJ, Ghaleh B. The use of proteome similarity for the qualitative and quantitative profiling of rep-erfused myocardium. J Chromatogr B Analyt Technol Biomed Life Sci. 2009;877:1317–26.
Williams JD, Flanagan M, Lopez L, Fischer S, Miller LA. Using accurate mass electrospray ionization-time-of-flight mass spectrometry with in-source collision-induced dissociation to sequence peptide mixtures. J Chromatogr A. 2003;1020:11–26.
Wilm M. Quantitative proteomics in biological research. Proteomics. 2009;9:4590–605.



Xin Huang, Ph.D., Nebraska, USA Xin Huang is the graduate student at the Department of Pathology and Microbiology at University of Nebraska Medical Center. He earned a BE and MS from Zhejiang University, and a PhD from University of Nebraska Medical Center.

Shi-Jian Ding, Ph.D., Director, Assistant Professor, USA
Shi-Jian Ding is the technical director of the Mass Spectrometry and Proteomics Core Facility at University of Nebraska Medical Center and an assistant professor at the Department of Pathology and Microbiology at University of Nebraska Medical Center. Prior to moving to Nebraska, Dr. Ding was a postdoctoral fellow at Pacific Northwest National Laboratory (2004–2007). His research interests center on development and application of mass spectrometry-based proteomic approaches for answering biological questions.

He has over 20 publications including *Proceedings of the National Academy of Sciences*, *Molecular and Cellular Proteomics*, *Analytical Chemistry*, *Journal of Proteome Research*, and *Proteomics*. He earned a BS from Lanzhou University and a Ph.D. from Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences.

Chapter 13
Clinical Translation of Protein Biomarkers Integrated with Bioinformatics

Xu Yang, Juanjuan Zhou, and Chaoqin Du

Abstract Proteomics technology offers a conceptually attractive platform for disease biomarker discovery. In protein biomarker discovery phase, some maturely quantitative proteomics methods such as stable isotope labeling by amino acids in cell culture (SILAC), isobaric peptide tags for relative and absolute quantification (iTRAQ), and label-free can be chosen. A common endpoint for a biomarker discovery experiment is a list of putative marker proteins. A reasonable next step is to perform target quantitative measurements of these proteins in an expanded patient population to verify their statistical significance. In addition, developments in quantitative MS, such as multiple reaction monitoring (MRM), have greatly enhanced both the specificity and sensitivity of MS-based assays to the point that they can rival immunoassay for some analytes. It is powerful for a lot of candidates verified. But for clinical diagnosis or monitoring, in general, the target biomarker is less than five proteins, so an immunoassay such as ELISA, Western blot, and antibody microarray is more suitable for the detection.

Many of the experimental techniques are developed in proteomics, which allow lots of proteins can be detected or quantified simultaneously. At the same time, bioinformatics solutions being developed make us get the maximum information derived from proteome data sets. There are some protein identified pipelines: peptide mass fingerprinting (PMF), tandem MS, or MS/MS search with a theoretical digestion database sequences. The use of searching against a decoy database that comprises

X. Yang, M.D., Ph.D. (✉) • J. Zhou
Department of Human Complex Disease Research of Research and Cooperation Division of BGI-Shenzhen, Beishan Industrial Zone, Main Building, Yantian District, Shenzhen, China
e-mail: yangxu@genomics.cn

C. Du
Proteomics Division, BGI-Shenzhen, Beishan Industrial Zone, Main building, Yantian District, Shenzhen, China

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_13, © Springer Science+Business Media Dordrecht 2013

sequences known to be incorrect allows the false discovery rate (FDR) to be estimated; other methods such as spectral library approach and de novo peptide sequencing method can also be used for peptide identification. Quantitative proteomics can be separated into two major approaches: (1) the use of stable isotope labeling, use isobaric tags information in MS/MS or comparative 2D gel approach for quantitative, and (2) label-free techniques, use area under the curve (AUC) or signal intensity measurement based on precursor ion spectra or spectral counting, which is based on counting the number of peptides assigned to a protein in an MS/MS experiment. Databases and data processing methods help researchers discover and validate their interesting protein or biomarker much simpler.

With the bioinformatics and mass spectrometry technique developing, the clinical translation of protein biomarker is much easier than before. But effective biomarkers which can improve diagnosis, guide molecularly targeted therapy, and monitor activity and therapeutic responses across a wide spectrum of diseases are still difficult to get. The effective and quick translation of protein biomarkers needs the accelerated mass spectrometry-based biomarker researches and bioinformatics techniques exploring, as well as accelerated the development of such research applied to routine clinical tests, in order to achieve commercialization ultimately.

Keywords Stable isotope labeling • Label-free • MRM • PMF • Quantification • Biomarker

13.1 Protein Biomarker Research

Biomarkers are any biological measurement that provides actionable information regarding a specific biological state, particularly one relevant to the risk of contraction, the presence, or the stage of disease. Though historically often a physical trait or physiological metric, the term is now typically shorthand for a molecular biomarker.

The protein domain is likely the most ubiquitously affected in disease, response and recovery, however, and proteomics holds special promise for biomarker discovery. Proteomics technology offers a conceptually attractive platform for disease biomarker discovery.

Advances in methods and technology now enable construction of a comprehensive biomarker pipeline from six essential process components: candidate discovery, qualification, verification, research assay optimization, biomarker validation, and commercialization.

Because the ultimate goal of biomarker discovery is usually the development of a blood test, blood is a logical fluid to use for biomarker discovery. Human blood, in the form of plasma or serum, is one of the most valuable specimens for protein biomarker discovery (Samir et al. 2008) because it is routinely collected, collection is minimally invasive, and it contains thousands of proteins, including those secreted or shed into the blood by tumors. Human plasma has been described as the most

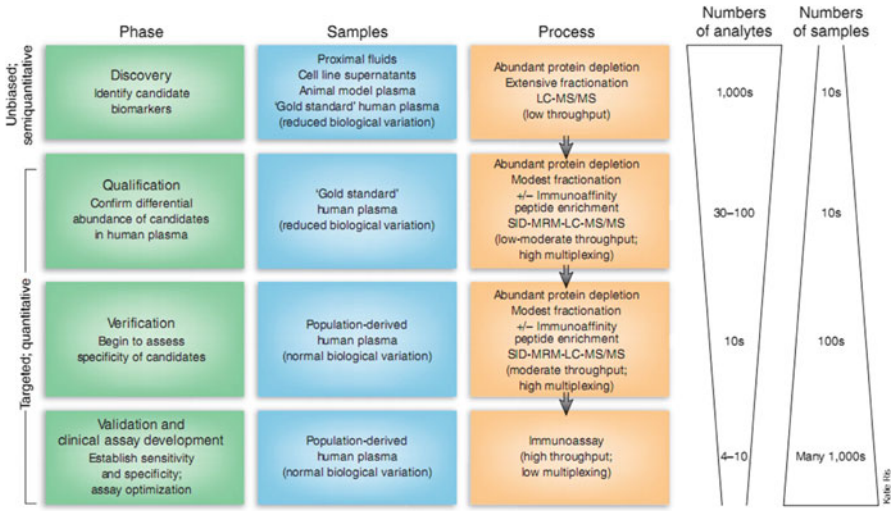


Fig. 13.1 Process flow for the development of novel protein biomarker candidates

comprehensive human proteome, a circulating representation of all body tissues and of both physiological and pathological processes. This comprehensiveness, together with the accessibility of human plasma and the vast medical laboratory infrastructure already in place for its analysis, ensures that it will remain the preferred diagnostic material for the foreseeable future. Despite its long study and immense clinical importance, human plasma remains very incompletely characterized, however, in part because it is thought to have tens of thousands of core proteins and to span 10–11 orders of magnitude in protein abundance, from albumin to the cytokines. Cell line homogenates, tissue lysates, and alternative biofluids, such as urine and cerebrospinal fluid, may also be used for discovery. Although less is known about the protein-abundance range and overall complexity of these samples, it is clear that they pose similar problems for proteomic analysis.

13.1.1 Quantitative Proteomics

Quantification on a large scale has been a main theme of dynamical proteome studies. For protein biomarker study, there are four phases (Rifai et al. 2006) (Fig. 13.1). The first "discovery" is the unbiased, quantitative process by which the differential expression of specific proteins between states is first defined. Discovery can employ model systems (e.g., mouse models or cell lines) or a variety of human biological materials and usually comprises a simplified, binary comparison between diseased and normal tissues, avoiding "contamination" by other diseases or confounding conditions. The "products" of the discovery phase are lists of proteins found to be

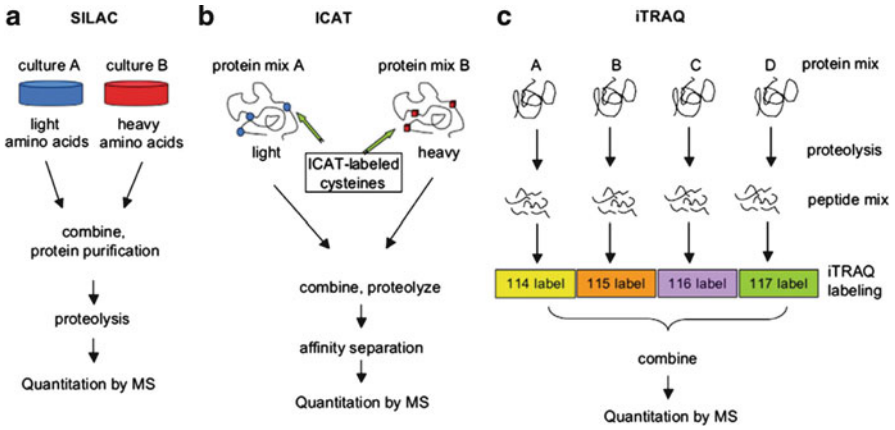


Fig. 13.2 Stable isotope labeling for protein quantification currently used in disease research

differentially expressed between the normal and diseased states based on semiquantitative assessment of relative peptide abundance in the MS data or the use of exogenous isotopic labeling or label-free.

The original quantitative "shotgun" technique involved digesting proteins, chemically or enzymatically labeling them with isotope tags (the two most common methods being isotope-coded affinity tags, or iCAT, and isotope tags for relative and absolute quantification, or iTRAQ), injecting them into a mass analyzer, and identifying and quantifying them by matching the resulting fragmentation spectrum to known protein spectra held on public databases (Goo and Goodlett 2010). As show in Fig. 13.2, A is SILAC (metabolic labeling). Cells grow in isotopically enriched culture media contain amino acids (15N or 13C) and metabolically incorporated during cell culture. Labeled cells are combined, purified, and proteolyzed prior to MS and MS/MS to determine relative protein abundance and for protein identification. It is applicable only to cultured cells and cannot be used for tissues and other body fluids. B is ICAT (chemical labeling). Heavy and light affinity tags are labeled to cysteine residues of protein mixture. Labeled proteins are combined, proteolyzed, and affinity separated, and relative protein abundance is determined. C is iTRAQ (chemical labeling). Labeling occurs at the peptide level after protein digestion. N-terminal and lysine residues of all peptides are labeled though acylation with reactive chemical compounds. It can multiplex up to eight different samples (four labels shown here).

More recently, chemical labeling has been replaced by metabolic labeling—the SILAC method (stable isotope labeling with amino acids in cell culture) popularized by Mann's group. Most cell lines, including those derived from animals, can be labeled with a heavy stable isotope, allowing very good quantitative studies.

The use of primary cell lines is challenged by the fact that they can only be cultured for a few passages to preserve their primary phenotype. This precludes the use of metabolic labeling of proteins such as N14/15 labeling and SILAC for quantitation as a certain number of passages are needed to obtain full incorporation. Instead,

alternative stable isotopic labeling methods such as iTRAQ are useful for labeling of primary cells. Using this method, the labels are added at the peptide level, and therefore, no demands are placed on the number of cell culture passages (Mitchell 2010).

Another trend in MS methods, according to Yates, is a shift from stable isotope labeling to label-free methods. Many labs now prefer label-free methods because they are much cheaper, and easier to perform, than SILAC. The two main methods produce proxy data that correlate well with protein abundances in complex samples. One measures the peak intensities of peptide ions, the limitation here being the purity of the peak. "Getting a clean peak and aligning the peaks can be difficult," says Yates. The other method uses spectral counting, which counts the number of tandem MS spectra assigned to each protein and the number of spectra for each peptide or protein being proportional to the amount of protein in the sample, that is, the frequency with which the peptide of interest has been sequenced by the MS. The main drawback of this method is the difficulty of measuring small changes in the quantity of low-abundance proteins, which is often masked by sampling error. However, the method has an excellent linear dynamic range of about three orders of magnitude, which isotopic labeling such as SILAC does not approach.

The next phase, "qualification," serves two somewhat disparate roles. First, qualification confirms that the differential candidate expression observed in discovery is seen using alternative, targeted methods (see Sect. 13.1.2). Second, qualification confirms differential expression of candidate biomarkers in simplified comparisons of diseased and normal human plasma samples, if discovery was not initially performed in plasma. Discovery and qualification are principally concerned with the consistency of association between marker and disease, emphasizing marker sensitivity (the likelihood that a diseased sample will test positive) over specificity (the likelihood that an unaffected sample will test negative).

13.1.2 Targeted Proteomics

In "verification," the analysis is extended to a larger number (hundreds) of human plasma samples, now incorporating a broader range of cases and controls, which begins to capture the environmental, genetic, biological, and stochastic variation in the population to be tested. Thus, in verification, biomarker candidate sensitivity is affirmed, and specificity begins to be assessed.

The conventional pipeline for biomarker verification and validation typically is conducted by immunoassay. While this approach is suitable for the development of single biomarkers, unlike in the discovery phase, hundreds of samples need to be run for statistical verification, so the process becomes inefficient and costly. Consequently, the emphasis is now shifting toward performing full biomarker discovery, qualification, and quantification on the same technology platform. The ease of multiplexing and ability to determine protein modifications makes MS an attractive alternative to antibody-based technologies. In addition, developments in

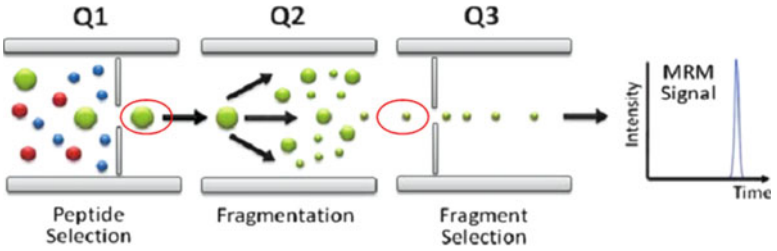


Fig. 13.3 MRM quantitative principle

quantitative MS, such as multiple reaction monitoring (MRM), have greatly enhanced both the specificity and sensitivity of MS-based assays to the point that they can rival immunoassay for some analytes (Kitteringham et al. 2009).

In MRM, only specific *m*/*z* values are selected for fragmentation, and only specific fragment ions are measured and reported, as shown in Fig. 13.3; the selected intact *m*/*z* and fragment *m*/*z* together are called a "transition." The intensity of this ion over time, specifically the area under its ion curve, is proportional to that ion's abundance in the sample (Hüttenhain et al. 2009). Because of this, careful selection of transitions is crucial; the precursor and fragment masses together must be unique among all possible peptides contained in a biological sample (McIntosh et al. 2009). Recently, more and more researchers have used MRM method to verify protein candidate biomarkers (Addona et al. 2011; Whiteaker et al. 2011).

The few candidate biomarkers that perform well in verification then move through "assay optimization" to formal "validation," in which a research grade of the final assay (a contemporary immunoassay) is tested against many thousands of samples that precisely reflect the full variation of the targeted population. Notably, the adoption of a well-characterized immunoaffinity reagent, whether it occurs in validation or at an earlier phase, represents the second important state change in biomarker development, as it can dramatically enhance signal-to-noise, streamline processing and enhance throughput.

Validated biomarkers may be selected for "commercialization," in which the research immunoassay is refined to meet the rigorous standards required for clinical tests. Although there is latitude in the nomenclature and partitioning, a coherent biomarker development pipeline must embrace all aspects of this process.

13.1.3 Modified Proteomics

To find useful markers of disease, the researchers focused on a vital biochemical event the addition of phosphate groups to serines and threonines in cellular proteins. Cells use this simple covalent modification over and over again to regulate protein-protein binding and activity of key enzymes. Measurement of this modification in specific proteins reveals their activation.

Andersen and colleagues (2010) have identified phosphoprotein biomarkers for a pathway often altered in cancer (the phosphatidylinositol 3-kinase (PI3K) pathway) and have shown that one of these predicts the sensitivity of cancer cells to a promising class of cancer drugs: inhibitors of AKT, a kinase that promotes growth and inhibits cell death.

Glycoproteomics represents an attractive approach for conducting peripheral blood-based cancer biomarker discovery due to the well-known altered pattern of protein glycosylation in cancer and the reduced complexity of the resultant glycoproteome.

Xuemei Zeng et al. (2010) apply it to a set of pooled non-small-cell lung cancer (NSCLC) case sera (nine adenocarcinoma and six squamous cell carcinoma pools from 54 patients) and matched control pools (eight matched healthy control pools from 106 cancer-free subjects). The goal of the study is to discover biomarkers that may enable improved early detection and diagnosis of lung cancer. A total of 38 glycopeptides from 22 different proteins were significantly differentially abundant across the case/control pools, and their abundances led to a near complete separation of case and control pools based on hierarchical clustering. The differential abundances of three of these candidate proteins were verified by commercially available ELISAs applied in the pools. Strong positive correlations between glycopeptide mass chromatograms and ELISA-measured protein abundance were observed for all of the selected glycoproteins.

13.2 Proteome Bioinformatics Approaches and Challenges

Genome project for medical and pharmaceutical industry is brought about a revolution, but should also see, the simple genetic analysis is very difficult to diagnosis the complex diseases. The complex interaction between genes in cells, activities, and the influence of the environment will influence the expression of genes and proteins of translation process after. So, reliable diagnosis and treatment should be based on the body of the development process of gradual regulation and disorders and must be given to the environment factors influence. Proteome is a powerful weapon in exploring this field. Combining different genomic and proteomic results obtained from the same biological system will substantially increase our understanding of complex biological processes.

Along with the development of the genome project, more and more protein sequence database was established. Also the improvement of MS technology has been a basic resource for proteomics research. So, it asks for deeper proteome bioinformatics analysis systems.

Mass spectrometry (MS)-based proteomics has undergone breathtaking advances in the past decade. Thanks to numerous advances, biological mass spectrometry technology development, for large-scale proteomics research, provides a powerful weapon. In protein identification by mass spectrometry in all aspects of the liquid chromatography mass spectrometry (LC-MS) platform, it is now possible to detect

and sometimes quantify thousands of proteins simultaneously from a complex sample. Modern mass spectrometers can churn out tens of thousands of MS/MS spectra every hour, so probably millions of MS/MS spectra are collected around the world every day.

Therefore, users lacking the knowledge or resources to analyze data could simply rely on the identifications returned, but those who wished to dig deeper on their own would get better spectra to work with.

13.2.1 Protein Identification

In the past decade and a half, mass spectrometry-based proteomics has witnessed breathtaking advances. Thousands of proteins can be indentified for one spectrum experiment. Among the many proposed experimental workflows, the most widely practiced method is probably the "bottom-up" approach of shotgun proteomics. The key steps of this approach are as follows: (1) proteins are digested into shorter peptides that are more amenable to LC-MS analysis, (2) peptides are further fragmented by tandem mass spectrometry (MS/MS) to yield characteristic fragmentation patterns, and (3) the MS/MS spectra are assigned to their originating peptides by various computational methods.

This last step of assigning MS/MS spectra to their peptide identifications is often the rate-limiting step of the whole proteomics experiment and has received well-deserved attention over the past decade. These computational methods can be generally classified into three groups, in terms of the "search space," the set of candidate sequences to consider as possible answers (Lam 2011a). On one end of the search space scale are de novo sequencing methods, which made no initial assumption on what peptides might be present in the sample. In the middle of the search space scale are sequence database searching methods, which rely on available sequence databases to limit the search space to only peptides that are derivable from known protein sequences. Aided by the rapid advances in genome sequencing and gene prediction that produces protein sequence databases for many model organisms, sequence database searching has become the method of choice for most proteomics researchers, despite its great demand for computational power. Toward the other end of the search three space scale is spectral library searching. In spectral library searching, the search space is further restricted to only those peptides which have been previously detected and identified and for which their fragmentation patterns have been experimentally recorded and compiled into spectral libraries.

13.2.2 Phosphorylation Analysis

Data analysis and interpretation remain major logistical challenges when attempting to identify large numbers of protein phosphorylation sites by nanoscale reverse-phase liquid chromatography/tandem mass spectrometry.

Beausoleil et al. (2006) address challenges that are often only addressable by laborious manual validation, including data set error, data set sensitivity, and phosphorylation site localization. As core, that measures the probability of correct phosphorylation site localization based on the presence and intensity of site-determining ions in MS/MS spectra. This method may spread for other posttranslational modification analyses.

13.2.3 Protein Quantification

Finding the protein differential expression between the case and control group is a convention method for finding biomarkers. There are many different quantification experiment methods. So, different bioinformatics methods are needed.

Isotope labeling combined with LC-MS/MS provides a robust platform for quantitative proteomics. Protein quantitation based on mass spectral data falls into two categories: one determined by MS/MS scans, for example, iTRAQ-labeling quantitation, and the other by MS scans, for example, quantitation using SILAC, ICAT, or 18O labeling.

The height or area of a peak at a particular mass-to-charge ratio (m/z) from a mass spectrum reflects the number of ions for that m/z detected by the mass spectrometer at any given time. This is typically known as the ion abundance. Although the ion abundance cannot be used to directly infer absolute protein or peptide concentration (due to different ionization efficiency for each peptide), comparing the ratio of ion abundances between identical peptides obtained in different experiment runs can be used to estimate differential expression.

The spectral count for a protein refers to the number of MS/MS spectra acquired from proteolytic peptide ions for that protein during a LC-MS/MS run. The premise of the method is that the more abundant the peptide, the more likely it will be selected for MS/MS analysis. In controlled experiments, it was found that the correlation of protein abundance with spectral count is superior to that of protein sequence coverage or peptide count (Wong and Cagney 2010).

While comparative quantification is adequate when the aim of the experiment is to find differences in protein expression between samples, absolute quantification would be useful for comparing protein levels between data generated at different times by different laboratories using different MS-based proteomics setups. Recently, reaction monitoring techniques that incorporate labeled peptide standards of known concentration have been adapted for proteomics work.

Selected reaction monitoring (SRM) (Prakash et al. 2009) is a powerful tandem mass spectrometry method that can be used to monitor target peptides within a complex protein digest. The specificity and sensitivity of the approach, as well as its capability to multiplex the measurement of many analytes in parallel, has made it a technology of particular promise for hypothesis-driven proteomics. It needs to create MRM transition lists from downloaded or custom-built spectral libraries, restricts output to specified proteins or peptides, and filters based on precursor peptide and product ion properties.

13.2.4 New Techniques

13.2.4.1 Spectral Archives

Tandem mass spectrometry (MS/MS) experiments yield multiple, nearly identical spectra of the same peptide in various laboratories, but proteomics researchers typically do not leverage the unidentified spectra produced in other labs to decode spectra they generate. Ari M Frank et al. (2011; Lam 2011b) propose a spectral archives approach that clusters MS/MS data sets, representing similar spectra by a single consensus spectrum. Spectral archives extend spectral libraries by analyzing both identified and unidentified spectra in the same way and maintaining information about peptide spectra that are common across species and conditions. Thus, archives offer both traditional library spectrum similarity-based search capabilities along with new ways to analyze the data. The method for clustering billions of unidentified tandem mass spectra from shotgun proteomics experiments offers new ways of storing, organizing, and analyzing proteomics data, with potential benefits to the entire proteomics community (Yen 2011).

13.2.4.2 Spectrum-to-Spectrum Searching

Spectrum-to-spectrum searching (Yen et al. 2009) using a proteome-wide spectral library can help researchers indent identifying more peptide. The proteome-wide spectral library can be built by combing spectral form public library and the simulated spectra.

13.2.4.3 Combining Engines Searching

Combining different search engines' result can contribute for more protein identification. Existing methods pool statistical significance scores such as p values, fdr_score, or posterior probabilities of peptide-spectrum matches (PSMs) from multiple search engines after high-scoring peptides have been assigned to spectra. Then researchers can give an agreed method to control identification error rates. Using a composite target-decoy database search strategy, we effectively estimated the error rates of applied score filter criteria, allowing comparisons of multiple data sets with similar error rates. Target-decoy search strategies are used for estimating statistical significance of MS/MS assignments and have seen recent use in evaluating false discovery rates in database search, spectral library search (Lam et al. 2010), multiple search, and other identification algorithms. This method has been proven to be effective for all of these protein identification methods.

Lukas Käll et al. (2008) describe a simple FDR inference method and then describe how estimating and taking into account the percentage of incorrectly identified spectra in the entire data set can lead to increased statistical power.

13.2.4.4 SWATCH Protein Identification

For this technology, data were acquired a fast, high-resolution quadrupole time-of-flight (TOF) instrument by repeatedly cycling through 32 consecutive 25-Da precursor isolation windows (swaths) (Gillet et al. 2012). Targeted data extraction enables ad libitum quantification refinement and dynamic extension of protein probing by iterative remaining of the once-and-forever acquired data sets. This combination of unbiased, broad range precursor ion fragmentation and targeted data extraction alleviates most constraints of present proteomic methods and should be equally applicable to the comprehensive analysis of other classes of analytes, beyond proteomics.

Therefore, the development of transparent tools for the analysis of proteomic data using statistical principles is a key challenge (Aebersold and Mann 2003). Only once such tools are tested, validated, and widely accepted will it become feasible to apply quality standards for protein identification, quantification, and other measurements and to compare complementary proteomic data sets generated in different laboratories. These comparisons will also depend critically on transparent file structures for data storage, communication, and visualization.

Different results were gotten from different analysis methods or different softwares. How to evaluate the different results is another problem. The uses of the new method need more standardization, simple and easily used software.

13.3 Challenges of Biomarker Clinical Application

The utility and importance of biomarkers has been recognized by substantial public and private funding, and biomarker discovery efforts are now commonplace in both academic and industrial settings. Yet despite intensified interest and investment, few novel biomarkers are used in clinical practice, and their rate of introduction is falling. Indeed, since 1998, the rate of introduction of new protein analytes approved by the US Food and Drug Administration has fallen to an average of one per year (Leigh Anderson, Plasma Proteome Institute, personal communication). The reasons for this disjunction are manifold and reflect the long and difficult path from candidate discovery to clinical assay and the lack of coherent and comprehensive processes (pipelines) for biomarker development (Rifai et al. 2006).

Countless millions of dollars have been thrown at the problem of looking for biomarkers; those discovered by proteomics researchers have turned out to be so nonspecific as to be next to useless, far from the “holy grail” envisaged some 10–15 years ago. “Biomarkers have been the biggest disappointment of the decade, probably because proteomics’ role in their discovery was overhyped,” says John Yates, director of the Proteomic Mass Spectrometry Lab at the Scripps Institute (La Jolla, CA, USA).

One difficulty has been the large dynamic range: the fact that protein abundance in biological fluids—particularly plasma, a favorite specimen for early biomarker

discovery work—spans some ten orders of magnitude (Fig. 13.1). “The serum proteomics debacle led to the realization that you can’t discover markers that are low abundance by doing discovery in serum or plasma,” says Daniel Liebler of Vanderbilt University (Nashville, TN, USA).

Another reason for the billion-dollar biomarker fiasco is the lack of validation, suggests Bernhard Kuster, chair of Proteomics and Bioanalytics at Technische Universitaet Muenchen, Freising, Germany. “I am sick of seeing papers proving that a known biomarker is a marker for yet another disease,” he says. “All it means is that the biomarkers discovered so far are mainly the same proteins that pop up in all kinds of diseases, indicating that the organism is under some kind of stress but not distinguishing between diseases. Various calgranulin proteins, for example, have been identified as serum biomarkers for everything from inflammatory arthritis to squamous cell carcinoma.”

Another stopper has been the problem of diversity among study participants diagnosed with a given disease and the lack of clear methods for defining clinical phenotypes so that samples can be classified consistently, which is vital to correlating the expression level of a protein with the presence of disease.

If we can discover and confirm the disease earlier and then give early treatment, it will greatly improve the survival rate. Therefore, we should take an effective use of known biomarkers and in-depth study to explore the new biomarkers in order to effectively achieve early detection of disease, early diagnosis, early treatment and good prognosis assessment, and prevention of recurrence.

Hence, future therapies will be tailored to the specific deranged molecular circuitry of an individual patient’s disease. The successful transition of these groundbreaking proteomic technologies from research tools to integrated clinical diagnostic platforms will require ongoing continued development and optimization with rigorous standardization development and quality control procedures (Calvo et al. 2005).

References

Addona TA, et al. A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nat Biotechnol. 2011;29:635–43.

Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422:198–207.

Andersen JN, et al. Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors. Sci Transl Med. 2010;43(2):43–55.

Beausoleil SA, et al. A probability-based approach f or high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006;24(10):1285–92.

Calvo KR, et al. Clinical proteomics: from biomarker discovery and cell signaling profiles to individualized personal therapy. Biosci Rep. 2005;25:107–26.

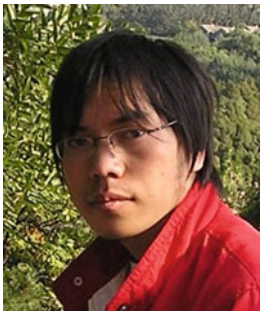
Frank AM, et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat Methods. 2011;8:587–91.

Gillet LC, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;18. doi:O111.016717.

Goo YA, Goodlett DR. Advances in proteomic prostate cancer biomarker discovery. J Proteomics. 2010;73:1839–50.
Hüttenhain R, et al. Perspectives of targeted mass spectrometry for protein biomarker verification. Curr Opin Chem Biol. 2009;13(5–6):518–25.
Käll L, et al. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res. 2008;7(01):29–34.
Kitteringham NR, et al. Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics. J Chromatogr B. 2009;877(13):1229–39.
Lam H. Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics. 2011a;10(12):R111.008565.
Lam H. Spectral archives: a vision for future proteomics data repositories. Nat Methods. 2011b;8:546–8.
Lam H, et al. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J Proteome Res. 2010;9:605–10.
McIntosh M, et al. Biomarker validation by targeted mass spectrometry. Nat Biotechnol. 2009;27:622–3.
Mitchell P. Proteomics retrenches. Nat Biotechnol. 2010;28:665–70.
Prakash A, et al. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J Proteome Res. 2009;8(6):2733–9.
Rifai N, et al. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol. 2006;24:971–83.
Samir M, et al. Mining the plasma proteome for cancer biomarkers. Nature. 2008;452:571–9.
Whiteaker JR, et al. A targeted proteomics–based pipeline for verification of biomarkers in plasma. Nat Biotechnol. 2011;29:625–34.
Wong JWH, Cagney G. An overview of label-free quantitation methods in proteomics by mass spectrometry. Methods Mol Biol. 2010;604:273–83.
Yen CY. Spectrum-to-spectrum searching using a proteome-wide spectral library. Mol Cell Proteomics. 2011;10. doi:M111.007666.
Yen CY, et al. A simulated MS/MS library for spectrum-to-spectrum searching in large-scale identification of proteins. Mol Cell Proteomics. 2009;8:857–69.
Zeng X, et al. Lung cancer serum biomarker discovery using glycoprotein capture and liquid chromatography mass spectrometry. J Proteome Res. 2010;9(12):6440–9.

Xu Yang, Ph.D., Director, China Dr. Xu Yang received his Ph.D. degree from Cardiovascular Institute and Fu Wai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, at 2008. He was a technical expert of Healthcare Division of BGI-Shenzhen from 2008 to 2009 and then the director of Department of Human Complex Disease Research of Research & Cooperation Division of BGI-Shenzhen from 2009 to now. His major area of research is related to human genetics, genomics, and human complex disease research using emerging technologies such as next-generation sequencing. He has participated in a lot of international projects such as 1,000 Genomes and LuCAMP (Diabetes-associated Genes and Variations Study).

Juanjuan Zhou, Project head, China Miss Juanjuan Zhou received her master's degree from Life Science College of Beijing Institute of Technology, Beijing, China, at 2010. She was a researcher of Proteomics Division of BGI-Shenzhen from 2008 to 2010 and then the proteomics expert of Human Medical Research & Cooperation Division of BGI-Shenzhen from 2011 to now. Her interests of proteomics are related to qualitative, quantitative, and modified proteomics by mass spectrometry and bioinformatics tools, protein biomarkers screening and validation using MS techniques, also early diagnosis methods, etc. She was involved in many researches such as cancer developed mechanism and antiaging drug investigation.



Chaoqin Du, China Mr. Chaoqin Du received her bachelor's degree from Applied Sciences College of University of Science and Technology Beijing, Beijing, China, at 2010. He was a researcher of Proteomics Division of BGI-Shenzhen from 2010 to now. His interests of proteomics are related to software exploit of protein identification-based mass spectrometry.

Chapter 14
Proteomic Approaches for Urine Biomarker Discovery in Bladder Cancer

Ming-Hui Yang and Yu-Chang Tyan

Abstract Bladder cancer is the most common urological cancer with higher incidence rate in the endemic areas of blackfoot disease (BFD) in southern Taiwan. Urine, a blood filtrate produced by the urinary system, is readily collected and is an important source of information for bladder cancers because it is directly exposed to bladder epithelium. Global analysis of the human urinary proteome is important for understanding urinary tract diseases. The aim of this chapter was to utilize the proteomic approach to establish urinary protein patterns of bladder cancer. The experimental results showed that most patients with bladder cancer had proteinuria or albuminuria. In the proteomic analysis, the urinary proteome was identified by nano-high-performance liquid chromatography electrospray ionization tandem mass spectrometry (nano-HPLC-ESI-MS/MS) followed by peptide fragmentation pattern analysis. ADAM28, identified by proteomic approaches and confirmed by ELISA, showed significant differences compared with normal individuals. The upregulation of urinary ADAM28 in bladder cancer was revealed, so it may be a biomarker of bladder cancer.

Keywords Proteomics • Bladder cancer • Urinary protein • Protein identification • Biomarker

M.-H. Yang
Department of Chemical and Materials Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan

Y.-C. Tyan, Ph.D. (✉)
Department of Medical Imaging and Radiological Sciences/Center of Excellence for Environmental Medicine/National Sun Yat-Sen University-Kaohsiung Medical University Joint Research Center, Kaohsiung Medical University,
100,Shi-Chuan 1st Road, Kaohsiung 807, Taiwan
e-mail: yctyan@kmu.edu.tw

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_14,

14.1 Human Body Fluids

Human body fluids play a significant part in proteome research. Body fluids such as serum, CSF, urine, and serous fluids/effusions (pleural, pericardial, and peritoneal) have proven to be a rich source of biomarkers for the detection of disease.

Among them, urine samples are the most easily obtained and are one of the most common samples in clinical analysis (Zerefos et al. 2006). Although human urine samples are ideal bio-samples and have tremendous potential as sources of biomarkers (Ru et al. 2006), they are also considered one of the most difficult proteomic samples with which to work. Samples originating from biological sources often contain a complex mixture of inorganic salts, buffers, chemotropic agents, surfactants/detergents, preservatives, and other solubilizing agents. High concentrations of salts are often present in urine samples to solubilize or stabilize analytes such as proteins. Due to inorganic salts or contaminants in urine, the utility of the proteomics approach in identification of urinary proteins has been poorly defined. Also, buffer concentrations employed by biochemists in urine samples often are in unacceptable ranges for mass spectrometric analysis (Oh et al. 2004).

14.2 Analytical Techniques in Proteomics: An Overview

"Proteome" and "proteomics" are relatively new words, coined by Wilkins et al. in (1996). Proteome is the PROTEins expressed by the genOME. Proteomic analysis means a comprehensive analysis of protein, and proteomics is the science by which proteins are comprehensively investigated with regard to their roles as functional elements. Use of proteomic techniques to identify disease-specific protein biomarkers is a powerful tool for defining prognosis of disease and gaining deep insights into disease mechanisms in which proteins play a major role. These techniques can also investigate structures and functions of protein complexes working as molecular micromachines expressed spatiotemporally through protein-protein interactions and signal transduction by PTM (Tyers and Mann 2003). Proteomic technologies such as immunoblotting, IEF, 2-DE, and MS have been proven to be useful for deciphering this unique proteome. Protein profiling has often been performed by the "classical" 1- or 2-DE based on the densitometric quantification of proteins visualized using dyes (silver, CBB, SYPRO Ruby, or DIGE) on gels. After in-gel enzymatic digestion of the subject protein spots, the resulting peptides are subjected to MALDI or ESI-MS. This review highlights some of the promising areas of effusion proteomic research and their clinical applications. Historically, 2-DE has been the primary method of separation and comparison for complex protein mixtures. This method is critical in developing our understanding of the complexity and variety of proteins contained in samples.

Although 2-DE is a powerful technique for the separation of proteins, there are some fundamental problems and limitations. The technique is laborious and time consuming (Fujii et al. 2004). In clinical and diagnostic proteomics, it is essential to

develop a comprehensive and robust system for proteome analysis (Hamler et al. 2004). Although impressive improvements in 2-DE technologies have occurred in recent years, identification of low-abundance proteins is still challenging. In contrast, 2D-LC separation of peptides following trypsin digestion of a complex mixture is the most sensitive protein identification technique developed yet and has become a central tool in linking the proteome to the genome. The use of MS and automatic database searching to identify the original proteins has been applied to overcome the shortcomings of the 2-DE.

MALDI-TOF-MS is the preferred method of MS. It is applicable to a relatively pure protein, for example, a single spot on a 2-DE gel. The protein is enzymatically cleaved into a peptide mixture, which undergoes MS and PMF for unambiguous protein identification using a large database (Poliness et al. 2004).

Recent advances in LC, MS, and data analysis software enable the direct analysis of extremely complex peptide mixtures, often referred to as shotgun proteomics or multidimensional protein identification technology (MudPIT) (Wolters et al. 2001). Although multidimensional liquid chromatography/tandem mass spectrometry (MD-LC/MS/MS) systems have been recently developed as powerful tools especially for the identification of protein complexes, these systems still have some drawbacks in their applications to clinical researches that require an analysis of a large number of human samples (Fujii et al. 2004).

A liquid-phase separation scheme coupled with MS is presented for the proteomic analysis of body fluid samples. Nano-HPLC-ESI-MS/MS is a highly sensitive and selective analytical technique that has become a powerful method for the identification of proteins present in complex mixtures. In particular, urine serves as a rich source of putative biomarkers that are not solely limited to cancer or infection. It is useful for the analysis of human samples in a clinical research environment (Lafitte et al. 2002).

14.3 Bladder Transitional Cell Carcinomas

Bladder cancer now represents the fourth most common malignancy in men and the tenth most common in women with a 20% death rate each year. It includes a wide spectrum of histological heterogeneous tumor types that arise predominantly in the transitional epithelium (urothelium) lining of the urinary bladder and the ureters (Celis et al. 2004). Tumor types of the urothelium include transitional cell carcinoma (TCC), squamous cell carcinoma (SCC), and adenocarcinoma, as well as other infrequent lesions (Bane and Rao 1996). At diagnosis, TCC of the urothelium represents over 90% of all bladder cancers, approximately 75% of which were papillary tumors localized to the urothelium or lamina propria. Only 5–8% of cases were SCC that had potential to invade deeper layers, and only 2% were adenocarcinoma, which were very likely to invade deeper layers (Tolson et al. 2006). At the present time, the two most reliable means of diagnosis and surveillance of bladder cancer are cystoscopic examination and bladder biopsy for histological confirmation. Such invasive bladder cancers may spread outside the bladder and affect other

Table 14.1 Summary of the sensitivity and specificity of bladder cancer detection methods

Methods of bladder cancer detection	Sensitivity[a] (%)	Specificity[a] (%)
NMP22	>68	>61
FDP	>82	>86
BTA stat/BTA TRAK	>80	>72
Telomerase activity	>85	>80
Cell-surface antigen with urine cytology[b]	>86	>90

[a]The percentages of sensitivity and specificity were accorded to Lee et al. (2005)
[b]The cell-surface antigen test and urine cytology can be carried out on the same slide

organs. However, when the disease is diagnosed and treated in an early stage, the survival rate is high. Early diagnosis is extremely important in this disease, and this fact underscores the need for development of a noninvasive, reliable, and simple examination to increase the detection rate of bladder cancer.

Although the traditional diagnostic methods of voided urine cytology and cystoscopy can achieve a specificity of 90–100% and are considered as the "gold standard" in bladder cancer diagnosis, they suffer from a lower sensitivity (27%), particularly in the detection of low-grade lesions and high-grade tumors (Konety and Williams 2004; Grossman 1998; Ramakumar et al. 1999; Zhang et al. 2004; Lwaki et al. 2004). To date, several urine-based markers for bladder cancer have been identified and investigated (Kageyama et al. 2004). Bladder tumor antigen tests (BTA tests), including the BTA stat, BTA TRAK, nuclear matrix protein-22 (NMP22), fibrin/fibrinogen degradation products (FDP), the urinary bladder cancer test (UBC), and other related tests, for example, telomerase activity, are accessible in the laboratory. These markers appear to have an advantage over urine cytology in terms of sensitivity, especially for detecting low-grade tumors. However, they are not likely to replace cytology methods because of the false-positive rates (Table 14.1). In general, all of the biomarkers mentioned above have higher sensitivity and lower specificity than urine cytology. Thus, no tumor marker with high specificity and sensitivity has been found as a routine diagnostic or screening tool for bladder carcinoma (Zhang et al. 2004; Lwaki et al. 2004; Kageyama et al. 2004; Celis et al. 2002; Lee et al. 2005). Therefore, the development of a noninvasive, reliable, and simple examination to increase the detection rate of bladder cancer would be advantageous.

14.4 Urine Proteome and Sample Preparation

Urine has long been known as a rich source of diagnostic information because of its physical properties and chemical composition. It is a suitable specimen for peptidomic studies, due to the fact that many low-molecular-weight proteins and peptides can pass through the glomerulus membrane, are catabolized within the proximal tubules, and are finally secreted in the urine. Therefore, the urinary

proteome is central to clinical and pharmaceutical research. Numerous efforts are being made to discover, identify, and validate biological markers for the diagnosis of renal function or kidney and urinary tract diseases (Hampel et al. 2001; Jürgens et al. 2005; Pieper et al. 2004; Marshall and William 1998; Delanghe 1997; Celis et al. 1996; Tantipaiboonwong et al. 2005). Recently, a more thorough investigation of the total protein composition of human urine is therefore required. Such a global analysis is important and can enhance our understanding of urogenital tract diseases and pathogenesis.

Human midstream urine specimens (first urine in the morning) were collected from 30 bladder cancer patients (ages in the range of 50–80) and 30 normal individuals with no evidence of disease (control group, ten males and ten females, aged 25–75, who did not consume aspirin or other nonsteroidal anti-inflammatory drugs for at least 2 weeks previous and have no history or evidence of urological cancers). The diagnosis of bladder cancer was confirmed by pathological examination (the primary tumor of TCC, stages T1 to T2, grades II to III). No female was menstruating at the time of collection. In cases of recurrence, none of the patients had received transurethral resection of the bladder tumor or chemotherapy before specimen collection. Urine samples were collected in polypropylene centrifuge tubes (DB Falcon, 50 mL, sterilized by gamma irradiation). After urine collection, protease inhibitor tables (Complete™ Mini; Roche) were added immediately to avoid proteolysis.

The urine samples were placed on ice prior to centrifugation at $2{,}000 \times g$ for 10 min at 4°C for the removal of cellular material and were frozen at −80°C to prevent bacterial growth. Protein concentrations were measured by Bio-Rad Bradford total protein assay kit (Bio-Rad Laboratories, Inc.). The total protein and urinary albumin concentrations in the urine samples of bladder cancer patients were significantly higher than normal individuals and displayed as the phenomenon of proteinuria and albuminuria (total protein >150 mg/L; bladder cancer patients, 178.2 ± 21.4 mg/L; normal individuals, 101.9 ± 17.5 mg/L; albumin concentration >20 mg/L).

14.5 Two-Dimensional Gel Electrophoresis Analysis for Urine

14.5.1 IEF and SDS-PAGE

The protein in each urine samples from bladder cancer patients and normal individuals was adjusted to 1 mg/mL by 25 mM ammonium bicarbonate. The proteins in the urine were precipitated with 10% w/v trichloroacetic acid (TCA) in acetone. The precipitates were washed with acetone three times and dissolved in isoelectric focusing (IEF) rehydration buffer for 2-DE analysis. IEF strips (pH 4–7, IPGphor, Amersham Biosciences, Uppsala, Sweden) were developed through a stepwise incremental voltage program: 30 V for 16 h, 500 V for 1 h, 1,000 V for 1 h, and 8,000 V for 4 h, with a total power of 34 kV-h. Then the strips were subjected to a two-step equilibration in

buffers containing 6 M urea, 30% glycerol, 2% SDS, and 50 mM Tris-HCl (pH 8.8) with 1% w/v dithiothreitol (DTT, Affymetrix/USB, 15395) for the first step and with 2.5% w/v iodoacetamide (IAA, Amersham Biosciences, RPN6302V) for the second step. The strips were then transferred onto the second-dimensional SDS-PAGE equipment and developed on 1.0-mm-thick gradient (8–16%) polyacrylamide gels at 10 °C. Proteins were identified using silver staining.

14.5.2 Silver Staining

The gels were fixed in 40% v/v ethanol and 10% v/v acetic acid in water overnight and then incubated in a buffer solution containing 30% v/v ethanol, 6.8% w/v sodium acetate, and 0.312% w/v sodium thiosulfate for 30 min. After washing three times in water for 5 min each, the gels were stained in 0.25% w/v silver nitrate solution containing 0.02% w/v formaldehyde for 30 min. Development was performed for 10 min in a solution consisting of 2.5% w/v sodium carbonate and 0.01% w/v formaldehyde. Acetic acid solution (5% v/v) was used to stop the development, and the stained gels were then rinsed three times in water for 5 min each.

14.5.3 2-DE Image Acquisition

The stained gels were scanned using an ImageScanner and LabScan 3.00 software (Amersham Biosciences). Image analysis was carried out using the ImageMaster 2D, Version 2002.1 (Amersham Biosciences). In this study, we performed 2-DE on six repeats of each sample from bladder cancer patients and normal individuals, respectively. However, conventional 2-DE analysis was not able to analyze the urine samples directly. Due to the differences in protein compositions of urine samples, the 2-DE images cannot be compared (Fig. 14.1). Therefore, we proposed to use non-gel-based proteomic approaches to analyze proteome in urine samples.

14.6 Urine Sample Cleanup, Buffer Exchange, and Digestion by MACS Separator System

Human urine samples were cleanup by magnetic nanoparticles and MACS® Separation column system with Milli-Q grade water (Millipore Co., Inc.). Magnetic nanoparticles, which surfaces were modified with –COOH group, were immersed in the coupling agent: 75 mM *N*-ethyl-*N*′-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC, E-6383, Sigma) and 15 mM *N*-hydroxysuccinimide (NHS, H-7377, Sigma) at 4 °C for 30 min (Delden et al. 1997; Kuijpers et al. 2000). Water-soluble EDC and NHS were used for activating O═C–OH (Kang et al. 1993; Tyan et al. 2002), and then the EDC-NHS buffer was removed and replaced by urine

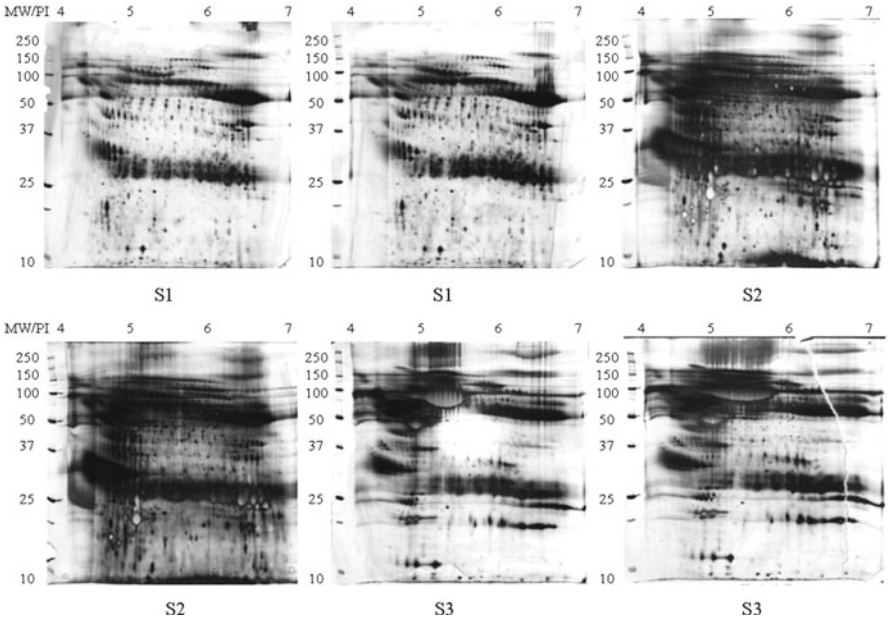


Fig. 14.1 Two-dimensional gel electrophoresis map of three different human urine samples. Each 2D-PAGE was loaded 150 μg proteins, and each urine sample has two replicates. S1, S2, and S3 were samples from three bladder cancer patients. Samples from the same column were run in the same 2D-PAGE tank

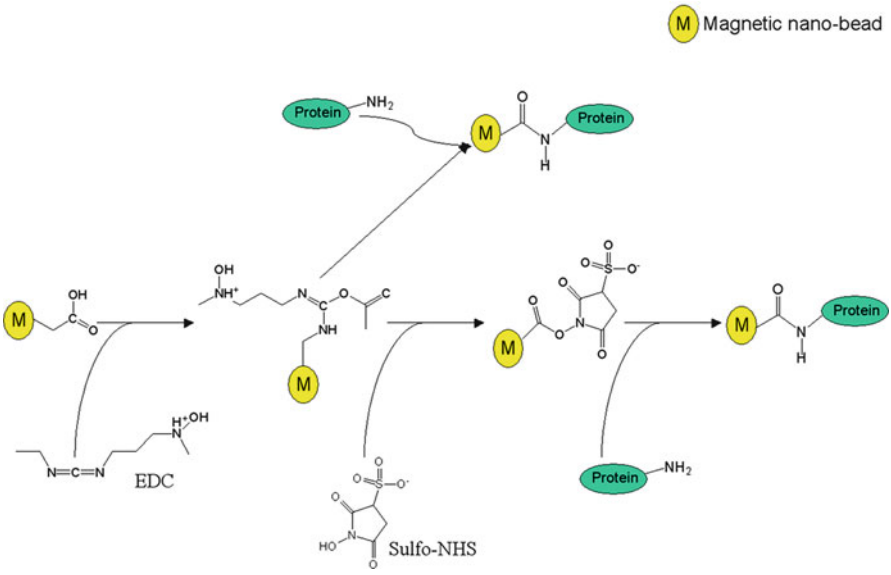


Fig. 14.2 Theory of proteins binding with magnetic nano-beads in urine

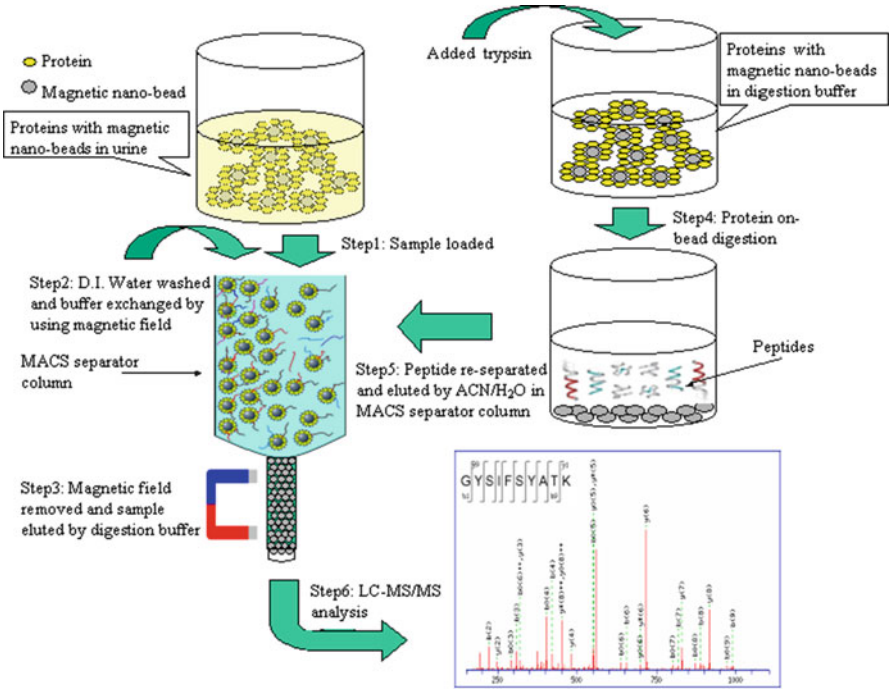


Fig. 14.3 Scheme of the overall experiment of urine sample analysis

samples. The protein-binding reaction was showed in Fig. 14.2. The protein-binding magnetic nanoparticles were loaded into the MACS® Separation column and washed twice with D.I. water. Then, protein-binding magnetic nanoparticles were eluted by digestion buffer and reduced, alkylated, and then digested with trypsin (Promega, V5111) to generate the constituent peptides. Peptides and magnetic nanoparticles were separated using the MACS® Separation column with magnetic field, and the peptides were eluted with 100 μL of 50% acetonitrile/H_2O for subsequent nano-HPLC-ESI-MS/MS analysis. Scheme of the overall experiment of urine sample analysis was showed in Fig. 14.3.

14.7 Shotgun Proteomic Analysis for Urine

14.7.1 Proteome Analysis by HPLC-MS/MS

The complex peptide mixtures were separated by RP-nano-HPLC-ESI-MS/MS. The protein tryptic digests were fractionated using a flow rate of 400 nL/min with a nano-HPLC system (nanoACQUITY UPLC, Waters, Milford, MA) coupled to an ion trap mass spectrometer (LTQ Orbitrap Discovery Hybrid

FTMS, Thermo, San Jose, CA) equipped with an electrospray ionization source. For RP-nano-HPLC-ESI-MS/MS, a sample (2 μl) of the desired peptide digest was loaded into the reverse phase column (Symmetry C18, 5 μm, 180 μm × 20 mm) by autosampler. The RP separation was performed using a linear acetonitrile gradient from 99% buffer A (100% D.I. water/0.1% formic acid) to 85% buffer B (100% acetonitrile/0.1% formic acid) in 100 min using the micropump at a flow rate of approximately 400 nL/min. The separation is performed on a C18 microcapillary column (BEH C18, 1.7 μm, 75 μm × 100 mm) using the nano-separation system. As peptides eluted from the microcapillary column, they were electrosprayed into the ESI-MS/MS with the application of a distal 2.1-kV spraying voltage with heated capillary temperature of 200°C. Each cycle of one full scan mass spectrum (*m*/*z* 400–2,000) was followed by three data-dependent tandem mass spectra with collision energy set at 35%.

14.7.2 Database Search

For protein identification, Mascot software (Version 2.2.1, Matrix Science, London, UK) was used to search the Swiss-Prot human protein sequence database. For proteolytic cleavages, only tryptic cleavage was allowed, and the number of maximal internal (missed) cleavage sites was set to 2. Variable modifications of cysteine with carboxyamidomethylation, methionine with oxidation, and asparagine/glutamine with deamidation were allowed. Mass tolerances of the precursor peptide ion and fragment ion were set to 3 ppm and 0.8 Da, respectively. Positive protein identifications were defined when Mowse scores greater than 100 were considered significant ($p < 0.05$). Proteins were initially annotated by similar searches using UniProtKB/Swiss-Prot databases.

The fragmentation spectra obtained by the RP-nano-HPLC-ESI-MS/MS analysis in gradient detection mode were compared with a nonredundant protein database using Mascot software. All Mascot results were visually confirmed. In addition, the criterion requires a readily observable series of at least four *y* ions for an identified peptide (Jaffe et al. 2004). When a protein was identified by three or more unique peptides, no visual assessment of spectra was conducted, and the protein was considered to be present in the sample. Proteins were initially annotated by similarity searches using NCBI PubMed (http://www.ncbi.nlm.nih.gov/), Swiss-Prot/TrEMBL (http://www.expasy.org/), and Bioinformatic Harvester EMBL (http://harvester.embl.de/) databases. If a protein was reported as a candidate protein in the literatures, it was selected as a potential biomarker and validated by Western blotting or ELISA. Scheme of the overall experiment design developed to use proteomic approaches for biomarker discovery was showed in Fig. 14.4.

Each RP-nano-HPLC-ESI-MS/MS analysis typically generated about 3,750 MS/MS spectra. A total around 45,000 (3,750 scans × 2 sample preparations × 6 repeat) MS/MS spectra were analyzed by the database search software Mascot. Only a small fraction (21.6%, 9,726) of searches produced significant matches according to the inclusion

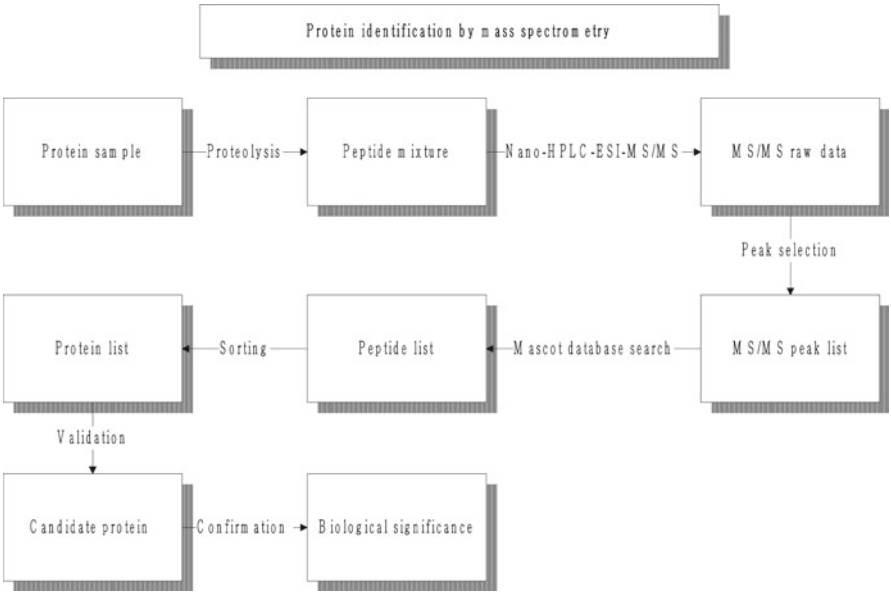


Fig. 14.4 Scheme of the overall experiment design developed to use proteomic approaches for biomarker discovery

criteria that we used for this study. The inclusion criteria were based on XCorr values, indices calculated by the search algorithms to reflect the similarity between product ion data and computer-simulated peptide fragmentation patterns. The peptide identification for the resulting MS/MS spectra was carried out using the Mascot database search software. These 9,726 significant matches represented 3,783 peptides. The total number of unique peptides identified in the urine sample was 661. The tryptic peptides produced were often mapped to protein sequence entries in the Swiss-Prot database. The 661 unique matched peptide sequences belonged to 504 protein sequence entries in the Swiss-Prot database. The database search resulted in 504 proteins, and most of these were identified at minimal confidence level, which was only one unique peptide sequence matched. Experimental results reported a total of 143 protein identifications with higher confidence levels (at least three unique peptide sequences matched).

Figure 14.5 showed the distribution of cellular locations of protein identification by using RP-nano-HPLC-ESI-MS/MS approach in this study. Among 504 proteins identified, 24.7% were known to be nuclear/nucleus proteins, 18.0% were known to be secreted proteins, 16.9% were known to be cytoplasmic proteins, and 11.2% were as membrane proteins. A few Golgi apparatus, cytoskeleton, mitochondrial, and microsome proteins were also identified (13.5%). A considerable portion of the identified proteins (20.2%) has not been reported for their cellular locations.

We used the ExPASy Molecular Biology Server (Expert Protein Analysis System) of the Swiss Institute of Bioinformatics (SIB) to explore what known functions of the identified proteins had been reported in the literature. The Swiss-Prot identifiers could

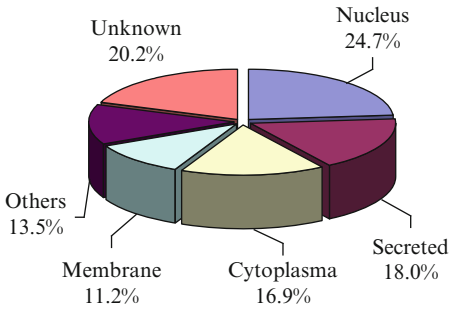


Fig. 14.5 Distribution of the identified human urine proteins according to their location. Assignments were made on the basis of information provided on the Swiss-Prot database at the ExPASy Molecular Biology Server. Some proteins were described of the different subcellular location, which explains the total sum being substantially larger than 100%

be employed for linkage of proteins to defined vocabulary of terms describing the biological processes, cellular components, and molecular functions of known Gene Ontology (GO). Gene Ontology Consortium provides annotation of each protein and structure allowed us to organize selected proteins into biologically relevant groups. These groupings can serve as the basis for identifying those areas of biology showing correlated protein changes (O'Donovan et al. 2002; Guo et al. 2005). The proteins identified in this study were presented into function categories based on their annotations in the GO database. Figure 14.6 shows the number and percentage of proteins with certain reportedly known biological processes and molecular functions.

Proteins identified in the category also include immunoglobulins, defenses, and urinary enzymes such as lysozyme and peroxidase, which form the urinary defense system to protect the urinary tract. Other groups were included some typical urinary proteins exhibiting binding, transport, metabolism, and catalytic activity. Among 504 proteins, 207 proteins were binding proteins. Eighty-six proteins were about catalytic activity. Seventy-four proteins have been known to be associated with metabolism, and 66 proteins have been known to be associated with structural molecule activity. Protein functions related to defense/immunity protein activity, transporter activity, and cell communication were also surveyed, and these functions were linked to considerable portions of the identified proteins in this study. Some proteins still had no prior functional information reported. It is not surprising that the largest groups of the identified proteins were functionally unknown proteins (including the hypothetical proteins).

14.8 Biomarker Discovery in Body Fluids

Body fluids, like urine, from very complex matrixes were contained a large number of potential biomarkers. It was well known that urine contains protein originating from blood plasma, the kidneys, and the urogenital tract and amounts of body filtrates

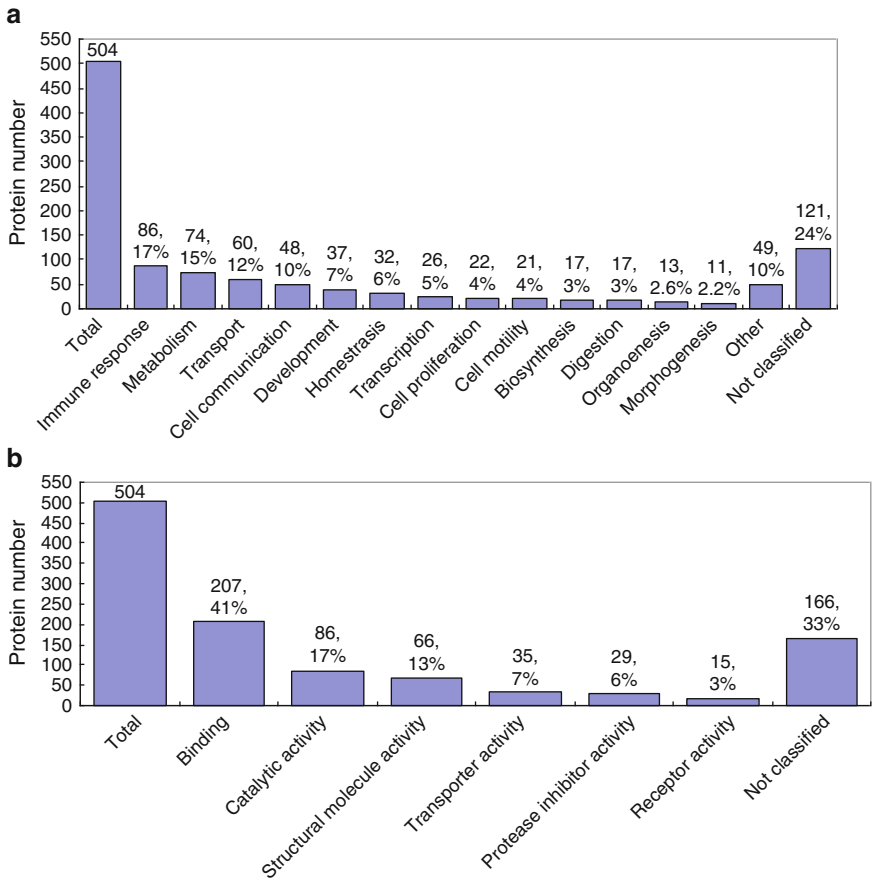


Fig. 14.6 The number and percentage of proteins with certain reportedly known biological processes (**a**) and molecular functions (**b**). Proteins involved in biological process and molecular functions were selectively presented and illustrated, the number of protein represents the percentage out of total GO term selected, and some proteins were described of the different biological process or molecular functions, which explains the total sum being substantially larger than 100%

such as water, salts, electrolytes, and nitrogenous waste products (Thongboonkerd et al. 2002; Beetham and Cattell 1993). Approximately 30% of urinary proteins were plasma proteins, whereas the other 70% were produced in the kidney (Tyan et al. 2006; Thongboonkerd and Malasit 2005). Therefore, urinary proteins were a mixture of plasma and renal proteins. Normal urinary proteins generally reflect normal tubular physiology. Information on changes in urinary protein excretion by various interventions is essential for a better understanding of tubular and glomerular responses to physiological stimuli (Thongboonkerd et al. 2002). The present study shows that the elimination of high-abundance proteins, salts, and interfering small molecules from urine enhances low-abundance protein profiles. A global analysis of urine proteins is crucial to a better understanding of the biology and physiology of

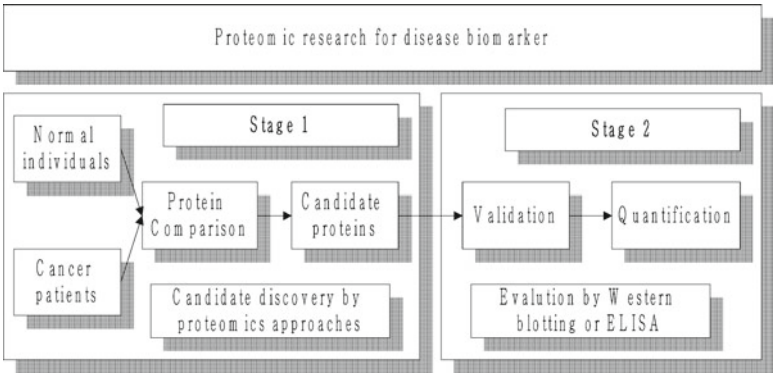


Fig. 14.7 Scheme of the overall experiment for disease biomarker discovery by proteomic approaches

the urinary tract (Smith et al. 2005; Thongboonkerd et al. 2003; Park et al. 2006; Castagna et al. 2005; Hong and Kwon 2005). Determination of protein composition of urine may lead to an increased understanding of renal physiology and will build a database for comparison to urine from patients with various urinary diseases. This chapter describes a fast and sensitive method of screening proteins in human urine. With the use of proteomics techniques, individual urinary proteins and/or urinary proteome map can be developed for diagnosis of bladder cancers.

The biomarker was defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacology responses to a therapeutic intervention." The search for protein biomarkers has been a highly pursued topic in the proteomics community in the last decade. In a basic proteomic research for disease biomarker, it contained several steps as below: candidate discovery, validation, quantification, and discussion (Fig. 14.7). In this study, each step was followed as candidate discovery by shotgun proteomics, validation and quantification by ELISA, and discussion by supportive literatures. The scheme of the protein analysis in this study was as below. When compared with the protein list from normal urine samples and previous literatures, some of these proteins were present in disease specimens, which may be associated with bladder cancer. If a protein was reported as a candidate protein in the literatures, it was selected as a potential biomarker and validated by ELISA or Western blotting.

14.9 Potential Biomarker for Bladder Cancer

A total of 143 proteins were identified with high levels of confidence, and 14 of these were significantly differentially expressed between the urine samples from the bladder cancer patients and those from the normal individuals (Table 14.2). Of these 14 proteins, ADAM28 is of special interest since it may play a role in the regulation of cell proliferation and differentiation.

Table 14.2 The 14 unique proteins identified by the higher confidence level (at least three unique peptide sequences matched) with significant difference between bladder cancer patient and control group urine samples in this study

Protein No.	Accession[a]	Protein name/peptide sequence[b]	Match peptide[c]	Cellular location	Molecular function	Biological process
P01	P06727	*Apolipoprotein A-IV precursor* K.LGEVNTYAGDLQK.K K.SELTQQLNALFQDK.L K.SLAELGGHLDQQVEEFR.R	3	Secreted	Lipid binding	Circulation, lipoprotein metabolism
P02	P05090	*Apolipoprotein D precursor* R.NPNLPPETVDSLK.N R.WYEIEK.I K.NILTSNNIDVK.K R.NPNLPPETVDSLK.N R.WYEIEK.I	5	Secreted	Lipid/protein binding	Transport, lipid metabolism
P03	P02647	*Apolipoprotein A-I precursor* K.LLDNWDSVTSTFSK.L R.DYVSQFEGSALGK.Q R.EQLGPVTQEFWDNLEK.E K.LREQ@LGPVTQEFWDNLEK.E	4	Secreted	Lipid binding and transporter activity	Circulation, lipoprotein metabolism
P04	Q9NQ38	*Serine protease inhibitor, Kazal type 5* R.VLPRIGYLCPKDLK.P K.CAM#CASVFK.L K.CEESSTPGTTAASM#PPSDE.-	3	Secreted	Serine-type endopeptidase inhibitor activity	Anti-inflammatory response
P05	Q9UKQ2	*ADAM metallopeptidase domain 28 isoform 1 preproprotein* K.C*GDNKVC*INAECVDIEK.A K.CGDNKVC*INAECVDIEK.A K.CGDNKVC*IN@AECVDIEK.A	3	Membrane	Metalloendo-peptidase activity	Spermatogenesis

P06	Q4VXC7	*Dynein, axonemal, heavy polypeptide 8* -.M#MKLYIDNAAPDKLK.G K.KISDLCEMHIDTVLK.E K.LQFYQRQFNEIIR.G K.VQAKFDAAMNEK.M	4	Unknown	ATP binding, ATPase a ctivity	Unknown
P07	P18031	*Protein tyrosine phosphatase, non-receptor type 1* K.C*STYQIKGSPNLTLPK.E K.KKIWASSM#DLLC*TADR.D K.SYHDLSQASLYPHRK.N	3	Cytoplasm	Protein tyrosine phosphatase activity	Signal transduction
P08	Q14008	*Colonic and hepatic tumor overexpressed protein isoform b* K.KGKPAAPGGAGNTGTK.N K.SGPIFIVVPNGKEQR.M R.DEYIEQLKTQMSSCVAK.W	3	Spindle bolt body	Microtubules binding	Unknown
P09	Q5T794	*MDN1, midasin homolog* K.GQEKDKEDPDSK.S K.SLLKQPIPEPKGGR.L R.MKLRDLMEAAFK.F	3	Nucleus	ATP/nucleotide binding	Protein complex assembly
P10	P08120	*Collagen, type XXVII, alpha 1* K.AGAPGRRGVQGLQGLPGPR.G K.GQKGDPGLSPGK.A R.GHLGSRGFPGIPGPSGPPGTK.G	3	Cytoplasm	Extracellular matrix structural constituent	Cell adhesion, phosphate transport
P11	P00739	*Haptoglobin-related protein* K.AVGDKLPEC*EAVC*GKPK.N K.DIAPTLTLYVGK.K K.GSFPWQAK.M K.SC*AVAEYGVYVK.V R.TEGDGVYTLNDKK.Q	5	Secreted	Hemoglobin binding, trypsin activity	Proteolysis and peptidolysis

(continued)

Table 14.2 (continued)

Protein No.	Accession[a]	Protein name/peptide sequence[b]	Match peptide[c]	Cellular location	Molecular function	Biological process
P12	Q8N4B3	*Zinc finger, MYND domain containing 11 isoform b* K.AVANM#QGEMDRKC*K.Q R.VYHSKC*LSDEFRLR.D R.STQTTNDGVCQSMC*HDK.Y	3	Nucleus	DNA binding	Regulation of transcription, DNA dependent
P13	51460893	*Cardiomyopathy associated 3* K.AKWLFETQPLEK.I K.DSDKKGK.E K.NQEDKLKM#VPR.K R.EYAVHIAMENNLEK.V	4	Unknown	Binding	Unknown
P14	P49750	*YLP motif containing 1* R.DYVPDRM#DWER.E R.GHEEFPLDGRNAPMER.E K.DAEEEESELGYIPK.S	3	Nucleus	Unknown	Unknown

[a]Swiss-Prot/TrEMBL accession number was given from http://us.expasy.org/ch2d/publi/inside1995, and NCBI accession number was given as a part of output from SEQUEST database search. Swiss-Prot entries start with a letter, and NCBI entries are all numbers

[b]Amino acid modifications: #: oxidation on methionine, average mass change: 15.9994; *: carboxyamidomethylation on cysteine, average mass change: 57.0215; @: deamidation on asparagine or glutamine, average mass change: 0.9840

[c]Number of unique peptide sequences generated from SEQUEST database search result for this protein

The ADAM28 was significantly higher in bladder cancer patient urine than in control urine. In the 30 urine samples of bladder cancer patients, ADAM28 was detected in all urine samples. To prevent the complication of age-related disease, urine samples from 30 healthy individuals, aged 25–75, were collected, and the ELISA analysis was performed on those samples. ADAM28 was not detected in control urine samples of similar or younger ages. Thus, the factors of a different metabolic profile and lifestyle due to age differences can be eliminated. The average ADAM28 concentration of bladder cancer patients was 0.0282 ± 0.0076 μg/μL in the urine specimens. For the control group, the ELISA results were lower than the detection limit value. Due to the fact that ADAM28 was not detected in the urine samples from control individuals, the p value between bladder cancer patients and control individuals should be less than 0.05.

ADAM28 belongs to a family of secreted, membrane-anchored, cell-surface glycoproteins that possess both proteolytic and adhesive properties. ADAMs are structurally related to snake venom disintegrins and have been implicated in a variety of biological processes involving cell-cell and cell-matrix interactions, including fertilization, muscle development, and neurogenesis (Mochizuki and Okada 2009). ADAM family members are expressed in several kinds of tumor types, such as lung cancer (ADAM8, 9, 12, 15, and 17), brain tumors (ADAM8, 17, 22, and 23), prostate cancer (ADAM1, 4, 5, 9, 10, 11, 15, and 17), liver carcinoma (ADAM9, 12, and 17), breast cancer (ADAM9, 12, 15, and 17), colon carcinoma (ADAM9, 10, 12, 15, and 17), pancreatic carcinoma (ADAM9, 10, 15, and 17), and kidney and bladder carcinoma (ADAM12 and 17), which suggests that ADAMs are involved in cancer progression including carcinogenesis (Rocks et al. 2008). Ohtsuka et al. confirmed the significantly higher expression levels of ADAM28 in lung carcinoma with lymph node metastasis by RT-PCR and tissue immunoblotting (Ohtsuka et al. 2006). ADAM28 is highly expressed in breast and non-small-cell lung carcinoma cells with expression levels correlating to tumor size, cell proliferation status, invasion, and metastasis activity (Mochizuki and Okada 2009). It was also indicated that SOX4 induced ADAM28 expression in bladder carcinoma cells (Aaboe et al. 2006). ADAM28 is likely involved in cell fusion, adhesion, membrane protein shedding, and proteolysis and may serve as a diagnostic tool for human cancers (Mochizuki and Okada 2007; Okada 2007; Kuroda et al. 2010; Mitsui et al. 2006). The correlation of ADAM28 in the urine with bladder cancer was indicated, and the upregulation of urinary ADAM28 in bladder transitional cell carcinoma was revealed in this study. Our data support the hypothesis that ADAM28 is involved in carcinogenesis and may be a valuable screening biomarker in human bladder transitional cell carcinoma.

14.10 Concluding Remarks

Many instrumental strategies have been developed and were based on efficient separation followed by MS identification. The method that biologists prefer is 2-DE. However, this 2-DE procedure suffers from many limitations, such as

sensitivity. Several separation methods have been reported in the scientific literature employing a single high-resolution procedure as well as an approach of HPLC with a mass spectrometer for protein identification. In this study, we examined a new preparation method of pooled urine samples for analysis of urinary proteins.

Urine, a blood filtrate produced by the urinary system, is readily collected and is an important source of information for bladder cancers because it has direct contact to bladder epithelium. The search for new biomarkers for early detection, monitoring, and prognosis of bladder cancer is an active area of interest. For urine proteome analysis, a useful cleanup method combined with LC-based separation was reported. The large-scale identification of urinary proteomes using HPLC-ESI-MS/MS may serve as an ideal and efficient method for the establishment of a panel of potential biomarkers and may help elucidate the mechanisms involved in bladder cancer. Our data demonstrated that the development of strategy to isolate urinary proteins can expand the urinary proteomic map. With the bioinformatic analysis, the urinary proteins may help discern the origin of proteins in the urinary tract.

Using proteomic approaches, we have profiled protein expression from bladder cancer and identified differentially expressed proteins. The upregulation of urinary ADAM28 in bladder cancer was revealed, and the correlation of ADAM28 in the urine with bladder cancer was indicated in this study. The database we generated provides information both on the identities of proteins present in human urine and a potential diagnostic biomarker for bladder cancer. Our results support its applicability in the characterization of the human urinary proteome in health and bladder cancer. This approach is a potentially powerful tool to discover new biomarkers and/or causative factors of disease-related proteins in urinary clinical studies. However, the results of these studies must be verified by larger clinical studies.

Acknowledgements We are thankful to S. Sheldon MT (ASCP) of Oklahoma University Medical Center Edmond (USA) for fruitful discussions. This work was supported by research grants NSC-099-2811-E-22-002 and NSC-100-2320-B-037-007-MY3 from the National Science Council and NSYSUKMU 101-015 from NSYSU-KMU Joint Research Project, Taiwan, Republic of China.

References

Aaboe M, et al. SOX4 expression in bladder carcinoma: clinical aspects and in vitro functional characterization. Cancer Res. 2006;66(7):3434–42.

Bane BL, Rao JY. Pathology and staging of bladder cancer. Semin Oncol. 1996;23(5):549–70.

Beetham R, Cattell WR. Proteinuria: pathophysiology, significance and recommendations for measurement in clinical practice. Ann Clin Biochem. 1993;30(Pt 5):425–34.

Castagna A, et al. Exploring the hidden human urinary proteome via ligand library beads. J Proteome Res. 2005;4(6):1917–30.

Celis JE, et al. Bladder squamous cell carcinomas express psoriasin and externalize it to the urine. J Urol. 1996;155(6):2105–12.

Celis JE, et al. Proteomic strategies to reveal tumor heterogeneity among urothelial papillomas. Mol Cell Proteomics. 2002;1(4):269–79.

Celis JE, et al. Impact of proteomics on bladder cancer research. Pharmacogenomics. 2004;5(4):381–94.
Delanghe J. Use of specific urinary proteins as diagnostic markers for renal disease. Acta Clin Belg. 1997;52(3):148–53.
Delden CJ, et al. Heparinization of gas plasma-modified polystyrene surfaces and the interactions of these surfaces with proteins studied with surface plasmon resonance. Biomaterials. 1997;18(12):845–52.
Fujii K, et al. Multidimensional protein profiling technology and its application to human plasma proteome. J Proteome Res. 2004;3(4):712–18.
Grossman HB. New methods for detection of bladder cancer. Semin Urol Oncol. 1998;16(1):17–22.
Guo Y, et al. 1-DE MS and 2-D LC-MS analysis of the mouse bronchoalveolar lavage proteome. Proteomics. 2005;5(17):4608–24.
Hamler RL, et al. A two-dimensional liquid-phase separation method coupled with mass spectrometry for proteomic studies of breast cancer and biomarker identification. Proteomics. 2004;4(3):562–77.
Hampel DJ, et al. Toward proteomics in uroscopy: urinary protein profiles after radiocontrast medium administration. J Am Soc Nephrol. 2001;12(5):1026–35.
Hong SS, Kwon SW. Profiling of urinary proteins by nano-high performance liquid chromatography/tandem mass spectrometry. J Liq Chromatogr Relat Technol. 2005;28(6):805–22.
Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics. 2004;4(1):59–77.
Jürgens M, et al. Towards characterization of the human urinary peptidome. Comb Chem High Throughput Screen. 2005;8(8):757–65.
Kageyama S, et al. Identification by proteomic analysis of calreticulin as a marker for bladder cancer and evaluation of the diagnostic accuracy of its detection in urine. Clin Chem. 2004;50(5):857–66.
Kang IK, et al. Immobilization of proteins on poly(methyl methacrylate) films. Biomaterials. 1993;14(10):787–92.
Konety BR, Williams RD. Superficial transitional (Ta/T1/CIS) cell carcinoma of the bladder. BJU Int. 2004;94(1):18–21.
Kuijpers AJ, et al. In vitro and in vivo evaluation of gelatin-chondroitin sulphate hydrogels for controlled release of antibacterial proteins. Biomaterials. 2000;21(17):1763–72.
Kuroda H, et al. ADAM28 is a serological and histochemical marker for non-small-cell lung cancers. Int J Cancer. 2010;127(8):1844–56.
Lafitte D, et al. Optimized preparation of urine samples for two-dimensional electrophoresis and initial application to patient samples. Clin Biochem. 2002;35(8):581–9.
Lee SW, Lee KI, Kim JY. Revealing urologic diseases by proteomic techniques. J Chromatogr B Analyt Technol Biomed Life Sci. 2005;815(1–2):203–13.
Lwaki H, et al. Diagnostic potential in bladder cancer of a panel of tumor markers (calreticulin, gamma -synuclein, and catechol-o-methyltransferase) identified by proteomic analysis. Cancer Sci. 2004;95(12):955–61.
Marshall T, William KM. Clinical analysis of human urinary proteins using high resolution electrophoretic methods. Electrophoresis. 1998;19(10):1752–70.
Mitsui Y, et al. ADAM28 is overexpressed in human breast carcinomas: implications for carcinoma cell proliferation through cleavage of insulin-like growth factor binding protein-3. Cancer Res. 2006;66(20):9913–20.
Mochizuki S, Okada Y. ADAMs in cancer cell proliferation and progression. Cancer Sci. 2007;98(5):621–8.
Mochizuki S, Okada Y. ADAM28 as a target for human cancers. Curr Pharm Des. 2009;15(20):2349–58.
O'Donovan C, et al. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 2002;3(3):275–84.
Oh J, et al. Establishment of a near-standard two-dimensional human urine proteomic map. Proteomics. 2004;4(11):3485–97.

Ohtsuka T, et al. ADAM28 is overexpressed in human non-small cell lung carcinomas and correlates with cell proliferation and lymph node metastasis. Int J Cancer. 2006;118(2):263–73.
Okada Y. Modulation of the microenvironment and adhesion of cancer cells by ADAMs (a disintegrin and metalloproteinase). Verh Dtsch Ges Pathol. 2007;91:29–38.
Park MR, et al. Establishment of a 2-D human urinary proteomic map in IgA nephropathy. Proteomics. 2006;6(3):1066–76.
Pieper R, et al. Characterization of the human urinary proteome: a method for high-resolution display of urinary proteins on two-dimensional electrophoresis gels with a yield of nearly 1400 distinct protein spots. Proteomics. 2004;4(4):1159–74.
Poliness AE, et al. Proteomic approaches in endometriosis research. Proteomics. 2004;4(7):1897–902.
Ramakumar S, et al. Comparison of screening methods in the detection of bladder cancer. J Urol. 1999;161(2):388–94.
Rocks N, et al. Emerging roles of ADAM and ADAMTS metalloproteinases in cancer. Biochimie. 2008;90(2):369–79.
Ru OC, et al. Proteomic profiling of human urine using multidimensional protein identification technology. J Chromatogr A. 2006;1111(2):166–74.
Smith G, et al. Development of a high-throughput method for preparing human urine for two-dimensional electrophoresis. Proteomics. 2005;5(9):2315–18.
Tantipaiboonwong P, et al. Different techniques for urinary protein analysis of normal and lung cancer patients. Proteomics. 2005;5(4):1140–9.
Thongboonkerd V, Malasit P. Renal and urinary proteomics: current applications and challenges. Proteomics. 2005;5(4):1033–42.
Thongboonkerd V, et al. Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation. Kidney Int. 2002;62(4):1461–9.
Thongboonkerd V, et al. Sodium loading changes urinary protein excretion: a proteomic analysis. Am J Physiol Renal Physiol. 2003;284(6):F1155–63.
Tolson JP, et al. Differential detection of S100A8 in transitional cell carcinoma of the bladder by pair wise tissue proteomic and immunohistochemical analysis. Proteomics. 2006;6(2):697–708.
Tyan YC, et al. Assessment and characterization of degradation effect for the varied degrees of ultra-violet radiation onto the collagen-bonded polypropylene non-woven fabric surfaces. Biomaterials. 2002;23(1):65–76.
Tyan YC, et al. Proteomic profiling of human urinary proteome using nano-high performance liquid chromatography/electrospray ionization tandem mass spectrometry. Anal Chim Acta. 2006;579(2):158–76.
Tyers M, Mann M. From genomics to proteomics. Nature. 2003;422(6928):193–7.
Wilkins MR, et al. Current challenges and future applications for protein maps and post-translational vector maps in proteome projects. Electrophoresis. 1996;17(5):830–8.
Wolters DA, Washburn MP, Yates JR. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. 2001;73(23):5683–90.
Zerefos PG, et al. Characterization of the human urine proteome by preparative electrophoresis in combination with 2-DE. Proteomics. 2006;6(15):4346–55.
Zhang YF, et al. Tree analysis of mass spectral urine profiles discriminates transitional cell carcinoma of the bladder from noncancer patient. Clin Biochem. 2004;37(9):772–9.

Yu-Chang Tyan, Ph.D., Associate Professor, Taiwan Yu-Chang Tyan received his B.S. degree in Medical Technology from Kaohsiung Medical University (Taiwan, 1996), and Ph.D. degree in Biomedical Engineering from Chung-Yuan Christian University (Taiwan, 2002). During the past 10 years, he has been a postdoctoral research fellow in the Department of Environmental & Occupational Health, National Cheng Kung University (Taiwan, 2003–2007), and an Assistant Professor in the Department of Medical Imaging and Radiological Sciences, Kaohsiung Medical University (Taiwan, 2007–2010).

He is currently an Associate Professor in the Department of Medical Imaging and Radiological Sciences, Kaohsiung Medical University (Taiwan, 2010~). His expertise includes proteomics, biomaterials and biomedical engineering. Award: Who's Who in Medicine and Healthcare 2011–2012. International health professional of the year 2012, International Biographical Centre, Cambridge, England.

Ming-Hui Yang, Ph.D., Taiwan Ming-Hui Yang received her B.S. degree in Medical Technology from Kaohsiung Medical University (Taiwan, 1996), B.S. degree in Chemistry from the University of Central Oklahoma (USA, 2000), M.S. degree in Chemistry from the University of Texas at Dallas (USA, 2003), and Ph.D. degree in Chemistry from Texas Christian University (USA, 2007). During the past 5 years, she has been a postdoctoral research fellow in the Graduate Institute of Medicine, Kaohsiung Medical University (Taiwan, 2007–2008), and Department of Chemistry, National Sun Yat-sen University (Taiwan, 2008–2010).

She is currently a postdoctoral research fellow in the Department of Chemical and Materials Engineering, National Yunlin University of Science & Technology (Taiwan, 2010~). Her expertise includes proteomics, biomaterials and biochemistry.

Chapter 15
Antibody Microarrays and Multiplexing

Jerry Zhou, Larissa Belov, Nicola Armstrong, and Richard I. Christopherson

Abstract This chapter presents a range of statistical methods for antibody microarray normalization and data analysis. Commonly used techniques for cluster generation, differential analysis, and classification are covered. The focus is on the implementation of each technique to the technology and its suitability in relation to sample types and experiment design.

Keywords Antibody microarray • Bioinformatics • Data variability • Normalization • Unsupervised clustering techniques • Supervised differential analysis • Multiple testing • Classification

15.1 Introduction

Antibody microarrays are one of several high-throughput proteomic technologies with great promise in biomedicine. One of the advantages of this technology over other proteomic approaches, such as liquid chromatrography mass spectrometry, is that it employs tiny quantities of antibodies to identify and quantify proteins that may be involved in development of a disease, or a protein to be

J. Zhou (✉) • L. Belov • R.I. Christopherson, Ph.D.
School of Molecular Bioscience, University of Sydney,
NSW, 2006, Australia
e-mail: jzho7551@uni.sydney.edu.au; larissa.belov@sydney.edu.au;
richard.christopherson@sydney.edu.au

N. Armstrong
Cancer Research Program, Garvan Institute of Medical Research,
Darlinghurst, NSW, 2120 Australia

School of Mathematics and Statistics and Prince of Wales Clinical School,
University of New South Wales, Rensington, 2052 Australia
e-mail: n.armstrong@garvan.org.au

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_15,

modulated by a treatment. The protein-binding patterns obtained can then be used as signatures to identify diseases or responses to treatment. Antibody microarray may also be used for biomarker discovery, drug target discovery, and insights into disease biology.

The field of antibody microarrays has developed with a major focus on the technology platform (e.g., surface material, antibodies, detection methods, quantification), rather than subsequent processes such as data handling, bioinformatics and data reporting. By contrast, numerous bioinformatic techniques have been developed for cDNA/DNA microarrays, and data reporting has been standardized in the form of MIAME (minimum information about a microarray experiment). While several of the statistical methods for normalization and data analysis developed for DNA arrays can be applied to antibody microarrays with minor changes. Researchers need to determine which statistical approach to use on their data to obtain meaningful and reliable results. This chapter presents a range of statistical methods for antibody microarray normalization and data analysis. The focus is on the implementation of each aspect of data analysis for antibody microarray and its suitability in relation to sample types and experimental design.

15.2 Antibody Microarray Formats and Detection Methods

A number of variants of antibody microarrays have been developed (Fig. 15.1). The simplest approach for detection of protein binding on an antibody array employs direct labeling (Fig. 15.1a), where proteins are tagged with fluorescent molecules. An alternative approach employs indirect labeling (Fig. 15.1b), where the proteins are pre-tagged with biotin or small haptens (like digoxigenin or dinitrophenol). Once bound on the array, these proteins are detected by incubating with a secondary labeled binder, such as fluorescently labeled avidin or hapten-specific antibody. The main advantage of the direct labeling approach over indirect labeling is that hundreds of proteins can be analyzed on the same slide. Fluorescence multiplexing can be used to compare two or more samples labeled with different fluorophores, on the same array. The main disadvantage, is that covalent labeling may alter the tertiary structure of the proteins and interfere with antibody recognition. The sensitivity of these methods is often dependent on the selection of high-affinity antibodies, or antibodies recognizing epitopes unaffected by labeling.

For the sandwich ELISA (Fig. 15.1c), unmodified proteins bind to the array; then a second antibody recognizing a different epitope on the same protein is applied. This second antibody is often tagged for detection (e.g., with a fluorophore, an enzyme such as horse radish peroxidase, or biotin). Alternatively, an untagged second antibody can be used, followed by a third tagged antibody (e.g., anti-mouse). The sandwich ELISA can be used for quantitative analysis, but only a small number of proteins can be detected simultaneously, mainly due to potential cross-reactivity between antibodies. In addition, the availability of antibody pairs recognizing different epitopes on the same protein is limited. A novel application of antibody microarrays

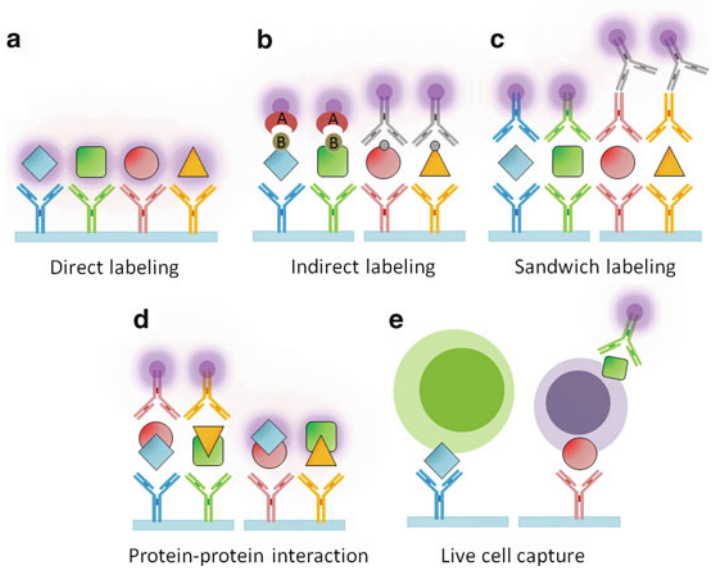


Fig. 15.1 Antibody microarray assay formats. (**a**) Direct labeling: proteins are labeled for detection before introduced to microarray. (**b**) Indirect labeling: proteins are tagged with biotin or small haptens. Proteins bound to microarray are then detected with labeled avidin or hapten-specific antibody. (**c**) Sandwich labeling: unmodified proteins are captured and detected with a labeled secondary antibody against the same protein. (**d**) Protein-protein interaction: protein interaction is detected with a labeled antibody against the binding partner of the protein of interest. Purified target protein can also be labeled to detect for interaction with other proteins on the array. (**e**) Whole cell capture: cell surface antigens are profiled on whole cells with antibodies specific for cell surface molecules. Unique cell populations can be profiled with labeled secondary antibodies specific for cell type (e.g., CD3 for T cells)

is to analyze protein/protein interactions (Fig. 15.1d). Cell protein extract is applied to the array, and a labeled antibody against the proposed binding partner of the protein of interest is applied to detect protein-protein interactions (Yang et al. 2006; Wang et al. 2000). One variant of this application requires purified fluorescently-labeled proteins of interest. After cell protein extracts are applied and bound to the array, the labeled purified proteins are introduced. Positive fluorescence indicates interaction of the labeled purified protein with binding partners in the cell extracts. Another application of the technology is the profiling of surface antigens on live whole cells (Fig. 15.1e). Cells are captured directly by antibodies against expressed surface antigens selected for the array (Belov et al. 2001, 2006). Cell-binding densities on antibody dots can be measured directly with an optical scanner. Alternatively, fluorescence multiplexing can be employed to discriminate between different cell types with fluorescently labeled antibodies against specific sub-populations bound on the array (Zhou et al. 2011).

A range of different detection methods is available for visualization of bound proteins. Fluorescent dyes such as Cyanine, Alexa, or Oyster are most commonly

used for detection. The continuing development of more sensitive and pH-stable dyes with narrow excitation and emission spectra makes fluorescence detection the method of choice. Most scanners enable multiplexing with up to 4 fluorophores, enabling direct multiplex comparison and relative quantification of four different samples, proteins, or cell types. Radioactive ^{125}I or ^{3}H provides greater sensitivity, but carries a risk of contamination, and is incompatible with high-throughput screening methods. Alternatives to fluorescence detection, such as rolling circle amplification (RCA) (Schweitzer et al. 2000), improve sensitivity, while quantum dot crystals potentially offer superior photostability compared to chemical dyes (Wu et al. 2003).

15.3 Data Variability

Systematic biases arising before analysis can disrupt expression profiling of clinical samples. Careful statistical examination of results can detect biases, but problems cannot be corrected within a set of samples collected under the same conditions. Systematic errors between samples should be minimized or eliminated, through good experimental design, careful analysis, and quality-control protocols.

Biological variability is an issue when protein profiling is used for *de novo* discovery. An adequate sample size is essential, but defining "adequate" can be challenging, especially when no prior knowledge exists about the target proteins (White et al. 2004). Regardless of the number of samples, the consistency of bound protein level measurements should be evaluated within disease groups. Proteins determined as significant in an analysis should be checked for consistency of levels across samples via bootstrapping, visual inspection, and/or outlier detection methods to ensure that the results are not influenced by a few samples. For example, bootstrap cluster analysis has been employed in DNA microarray studies to evaluate data consistency (Kerr and Churchill 2001).

One of the major sources of inconsistencies in antibody microarray data is the variation in affinity and performance of different antibodies on the array. Heterogeneity in antibody affinity can span many orders of magnitude. The performance of each antibody can depend on the chemistry of the solid surface to which it is applied, and on the buffer, storage conditions, and storage time. The shelf lives of antibodies vary in solution and in the solid phase. Although there are thousands of antibodies commercially available, incorporation of multiple antibodies with redundant antigen specificities in an array may be wasteful. Although several attempts have been made to replace expensive antibody production in hybridomas by the production of scFvs or Fabs from phage display (Hallborn and Carlsson 2002) or ribosomal display (Hanes et al. 2000), this approach is not usually applied to microarrays.

Antibodies immobilized on the microarray, can display varying performance, ranging from no activity to reduced affinity or specificity (Angenendt et al. 2002). Antibody affinity can be increased by optimizing the surface to which antibodies are

applied, applying indirect immobilization strategies (Kusnezow et al. 2003) [13] or selecting a MAb from a different hybridoma clone. Another source of variation comes from protein tagging. In addition to possible epitope damage, the diversity and differing quantities of available amino acids, acting as targets for coupling chemistries, may prevent absolute quantification of proteins in complex samples. The measurement of recombinantly produced proteins, engineered to allow indirect labeling by affinity tags or fusions with fluorescent reporter molecules (Kukar et al. 2002), has been suggested, but both approaches can cause changes to protein tertiary structure or folding conformation. Taken together, these limitations rule out reliable quantification, preventing direct comparisons between different proteins on the same array. Fortunately, since these issues remain relatively consistent between batches of arrays and apply equally to all samples tested under identical conditions, it can be assumed that expression differences for each protein can be accurately calculated between samples.

15.4 Normalization

To accurately determine differential expression between samples, it is important that systematic biases in the measured protein levels are eliminated. In antibody microarrays, systematic differences can arise from differences in labeling and detection efficiencies for the fluorescent labels, differences in the quantity of initial protein, and combinations of these effects. These problems can cause systematic differences in protein intensities between arrays, and the intensities must be adjusted before data analysis. In general, intensities are log-transformed using log base 2 providing data that is homoscedastic, that is, has constant variance, an important assumption for many statistical methods.

15.4.1 Internal Ratio Normalization

This normalization procedure borrows directly from two-dye cDNA array methods, where the sample of interest is labeled with Cy3 and a reference sample is labeled with Cy5. These dye pairs have nearly identical chemistries, but different fluorescence wavelengths, one in the red range and the other in the green. If there is more of either Cy3 or Cy5, the spot will be either green or red, respectively. If the samples are present in equal amounts, the mixture will show as yellow. A reference sample or internal standard is routinely included to give an exact value to the “yield” or “efficiency” of the assay. However, in the case of antibody microarrays, as the number of analytes increases, it becomes more difficult to assemble specific standards. For studies on cellular proteomics, reference samples can be prepared by pooling material from cell lines, tissues, or extracts, as long as a large amount is made and stored properly to avoid freeze-thaw cycles (Pollard

et al. 2005, 2007). Serum studies can employ commercial samples of pooled human serum that has been well categorized proteomically by conventional mass spectrometry (Srivastava et al. 2006).

This approach carries with it the limitation that the data are always ratios, not absolute intensity levels. Although ratios reveal some patterns in the profile, they remove information about individual protein levels. Various parameters depend on the measured intensity, including the confidence limits that are placed on any microarray measurement. For two-dye systems, within slide normalization is necessary to adjust for the differences in intensity levels between dyes. Red Cy5 intensities are generally lower than green Cy3 intensities. A MA plot showing the distribution of the red/green intensity ratio (M) plotted by the average intensity (A) is used to identify intensity-dependent biases in microarray data. If the data cloud is approximately horizontal, this suggests that no biases exist and a simple ANOVA-like normalization method would suffice. However, if intensity-dependent biases do exist, represented as curves in the data cloud, a nonparametric local regression procedure such as Loess (Yang et al. 2002) may be used.

Olle et al. (2005) have produced a variant of the two-dye microarray, where the antibody is used as an internal control, with one color quantifying the antigen and the second quantifying the antibody. Primary antibodies on the microarray are introduced to biotin-labeled protein samples (whose binding is detected with labeled streptavidin), followed by labeled universal antibodies against the constant region of the antibodies. Normalization of antigen concentration is determined as a ratio of the median antigen fluorescence intensity divided by the median spotted antibody fluorescence intensity. Because the kinetics of antibody-antigen interactions depends on both antibody and antigen concentration, quantifying antibodies bound to the slide as an internal control allows for more accurate assessment of proteins, comparable to that of Western blotting (Olle et al. 2005).

15.4.2 Mean/Median Normalization

Mean normalization is the simplest and most widely used global normalization technique for antibody microarray data. The log-transformed intensities are adjusted such that the mean of each array is set to the same value. In cases where a large number of outliers in the data may provide an inaccurate estimate of the true center of distribution, median centering can be employed. The advantage of both mean and median normalization methods is their mathematical simplicity, enabling quick and easy implementation. Mean centering and its variations have proven to be a reliable method of minimizing technical variance in large-scale high-throughput data (Mestdagh et al. 2009; Wylie et al. 2011). It also performs well for reproducibility and accuracy when applied to antibody microarray data, provided the average concentrations of the measured proteins are constant between samples (Hamelinck et al. 2005). However, because average concentrations may vary, normalization by mean centering may occasionally produce results that inaccurately reflect the

trends in the data, especially if only a small number of proteins are measured. Discrepancies between mean and median values may lead to data being normalized in opposite directions depending on which summary measure is chosen. The main issue with this normalization method is that it implicitly assumes differences between arrays and/or samples are of a linear nature. Despite these shortcomings, mean and median centering remain the preferred normalization approaches in antibody microarray analysis. It is hoped that a normalization method capable of handling nonlinear signals may be developed in the future, leading to standardization of antibody microarray normalization.

15.4.3 Loess/Cyclic Loess Normalization

Nonlinear normalization techniques, such as Loess, apply intensity-based corrections to account for biases in the data that may arise from nonlinear relationships between two samples. This technique is based upon the idea of the MA plot generated by two color channel arrays. An extension of this, called cyclic Loess, can be applied to probe intensities from two samples at a time and extended to look at all pairwise combinations of samples when dealing with more than two arrays (Bolstad et al. 2003). This approach is necessary for single-color antibody microarrays, where data tend to be nonlinear and require more adjustments for mismatches in the global intensity across each array. Cyclic loess normalization is applied by generating a loess regression line of fit through the scatter plot of two arrays (Yang et al. 2002). Loess smoothing performs locally weighted least squares regression on a section of the data, before applying this in continuity to the rest of the data set, using a moving window of local data points to derive a fit line. Locally weighted regression can be performed for every spot on the array or on a range of spots predetermined by a grid. The latter reduces computational cost, but in practice, the differences are minimal on most modern desktops. The normalizations are carried out in a pairwise manner, recording an adjustment for both arrays in each pair. After looking at all possible pairs of arrays, the set of adjustment are applied to the set of arrays. This is repeated, and typically, 1 or 2 complete iterations through all pair-wise combinations are performed. Hence, this method is more time consuming than mean/median centering.

Direct implementation of Loess normalization to antibody microarray data has been shown to lower reproducibility and reduce correlation between data (Hamelinck et al. 2005). The main reason for this is that the Loess method was developed for DNA microarray data and relies on having a large number of data points to produce an accurate picture of intensity-based biases between two samples. Generally, antibody microarrays have many fewer data points, often less than a thousand different antibodies, compared to well over 50,000 different oligonucleotides on its DNA counterpart. Therefore, without compensation for fewer data points, direct translation of this method may introduce added noise to the data and reduce true biological variation.

15.4.4 Normalizing Against House-keeping Proteins

Sets of house-keeping genes are often used to normalize DNA microarray data, but similar "house-keeping" proteins are much more difficult to find for antibody microarrays. We showed that CD44 could be used as a "house-keeping" protein for normalizing antibody microarray data when investigating changes induced by several differentiation-inducing agents in surface antigen expression profiles of human HL60 leukemia cells (White et al. 2005). The same normalization approach could not be used for surface proteome profiling of a range of clinical leukemia and lymphoma samples (Belov et al. 2001), because CD44 expression varied. Instead, data were normalized on maximum dot intensity for each sample, resulting in correct disease classification for 97.6% of the patient samples.

It may be difficult to find an effective single "house-keeping" reference protein for complex biological samples (e.g., serum, tissue). Serum IgM levels have been used with some success, for normalizing serum-protein profile data (Hamelinck et al. 2005). Also, normalization against a spiked-in standard, such as 2,4-dinitrophenol (DNP)-labeled BSA, has provided accurate results independent of the size of the array (Hamelinck et al. 2005; Miller et al. 2003). However, the reliability of this approach depends on the quality of the microarray measurements, and on the accuracy of values for the spiked-in protein standard, and it would not correct for sources of bias that occurred before the standard was introduced. It has been suggested that a more "global" normalization approach against a given set of protein analytes, anticipated to display very little or no inter-sample variation, may yield improved results for analysis of complex biological samples.

15.4.5 Summary

The normalization of antibody microarray data is complex and should be incorporated into the experimental design. A reference sample or internal standard protein/s may be used for normalization, but may not be feasible for some antibody microarray experiments. Global linear transformations using mean/median centering can correct for systematic errors that affect all intensities equally across samples. Loess approaches allow for nonlinear trends to be removed, but current methods developed for cDNA microarrays may not translate to antibody microarrays due to the fewer data points.

15.5 Comparing Expression Groups

Statistical analysis of antibody microarray data can be separated into unsupervised and supervised methods. Figure 15.2 illustrates the workflow for data analysis. In unsupervised methods, classes or categories are unknown and need to be discovered from the data. This approach is useful for hypothesis generation, where

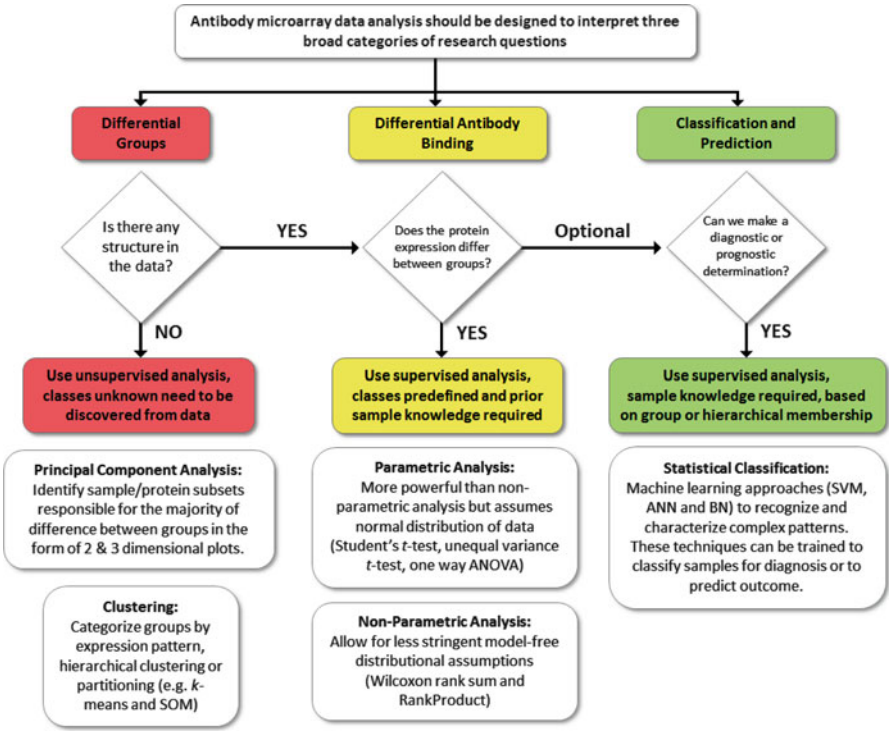


Fig. 15.2 Workflow for antibody microarray data analysis. *SOM* self-organizing map, *ANOVA* analysis of variance, *EB* empirical Bayes, *SAM* significance analysis of microarray, *MMM* mixture model method, *SVM* support vector machine, *ANN* artificial neural network, and *BN* Bayesian network

proteins are categorized exclusively by expression patterns, using methods including hierarchical clustering, *k*-means clustering, self-organizing maps (SOMs), and principal component analysis (PCA) (Quackenbush 2001). On the other hand, in supervised methods, classes are predefined by prior knowledge, and a training set of well-characterized objects (e.g., samples) can be used to form a classifier (e.g., a disease signature and a normal signature) for classification of subsequent observations (Ringner et al. 2002). The approaches discussed below are summarized in Table 15.1.

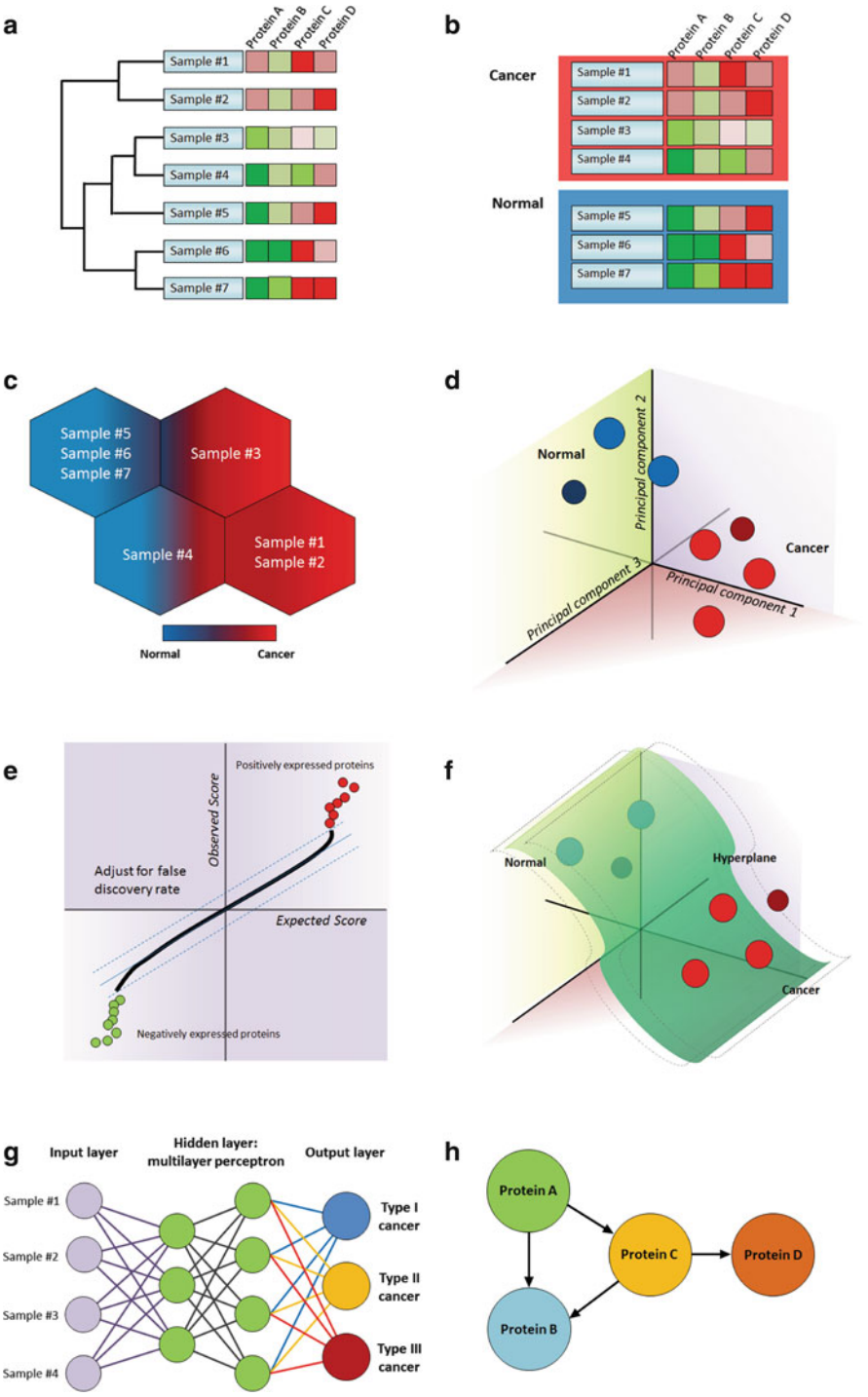
15.5.1 Hierarchical Clustering

Unsupervised clustering involves organizing microarray data from multiple samples and/or proteins into groups that have similar patterns. The most common method is hierarchical clustering (Fig. 15.3a). To construct a hierarchy, the distance or

Table 15.1 Summary of the strength and limitation of statistical approaches used in antibody microarray data analysis

Technique	Strength	Limitation
Unsupervised cluster analysis		
Hierarchical clustering	Powerful hypothesis generation tool that allows for simple data visualization (heat maps) and easy identification of biologically significant clusters	Clusters will always be produced, even in unrelated data. Without any measurements for the significance of difference between clusters, the division of clusters is determined subjectively by external criteria. The robustness and stability of clusters is linked to choice of parameters. Optimal combination of distance/linkage measure can be difficult to determine
k-means clustering	Prior knowledge is used to partition data into fixed groups (*k*) before clustering to reduce unintentional clustering of unrelated data	Results are dependent on the number of clusters set, but it can be difficult to determine the optimal number of clusters. Only applicable for low-dimensional data as increasing *k* also increases risk of overfitting. Starting "seeds" for *k*-means often lead to different results
Self-organizing maps (SOMs)	SOM provides a simple-to-understand representation of the relationships between data. Can also be trained to recognize and classify complex patterns	Value is required for each dimension of each member of samples in order to generate map, and this can be difficult to achieve in high-dimensional data. Ever SOM generated is slightly different, and similar samples are not always near each other. A lot of maps need to be constructed to get a reliable final map
Principal component analysis (PCA)	Simplifies dimensionality of data without significant loss of information. Allow for visualization of trends and the influence of particular variables to data distribution. Can be used to identify technical variables (e.g., batch effect)	Difficult to set precise boundaries of distinct clusters or to define proteins belonging to each cluster. Not applicable to data with nonlinear relationships between variables
Supervised techniques		
t-Test	Simple to implement and easy to interpret results. More powerful than nonparametric approaches. Univariate *t*-test suited for data with unequal variance	Strong assumption of normal distribution. Can only test one factor at a time between two groups/treatments

Analysis of variance (ANOVA)	Test for interactions between several factors in two or more groups at a time. The overall type I error rate can be controlled	Assumptions of normal distribution and homogeneity of variance. ANOVA only indicates significance in the data, and additional post hoc tests are required to identify where the significant differences are
Wilcoxon rank sum	Reliable alternative for the *t*-test in the case of non-normally distributed data (sample distribution with differences in location)	Loss of power, compared to the *t*-test, for normally distributed data. Reliable analysis requires sample distribution to have the same shape (homogeneity of variance)
Rank product (RP)	Specifically designed for detection of differential expression in microarray experiments. Good at handling noisy data sets with a very small number of replicates. Results from various experimental platforms can be combined in one analysis as long as they are expressible as rank lists	Sensitive to variations in protein-specific variance, especially in data with high variance of weakly expressed proteins
Shrinkage methods	Borrow information across proteins and conditions to provide better estimates of variance of expression levels. Allows user to control the average number of false positives and ability to correlate many different types of parameters. Repeat permutations of microarray replicates make it possible to distinguish genuine results from noise, even in a small number of samples	Relies on estimating the measurement of variance for each single protein, thus can become unreliable with a low number of replicates. A large number of proteins are required to estimate the null distribution and obtain meaningful results
Supervised machine learning		
Support vector machine (SVM)	Superior at handling high-dimensional data. SVM outperforms ANN as training data decreases	Can only classify samples into two factors at a time. Noisy data can cause overfitting
Artificial neural network (ANN)	More straightforward probability interpretation. Can be adopted to predict continuous values instead of dichotomous classes	Requires preprocessing of high-dimensional data to avoid overfitting
Bayesian network (BN)	Able to handle incomplete and noisy data and to express causal relationships among variables	Quality of the BN is dependent on reliable prior knowledge used in the Bayesian inference processing (the quality and extent of proteomic information may be lacking)


a
Protein A
Protein B
Protein C
Protein D
Sample #1
Sample #2
Sample #3
Sample #4
Sample #5
Sample #6
Sample #7
b
Cancer
Normal
c
Sample #5
Sample #6
Sample #7
Sample #3
Sample #4
Sample #1
Sample #2
Normal
Cancer
d
Normal
Principal component 2
Cancer
Principal component 3
Principal component 1
e
Positively expressed proteins
Observed Score
Adjust for false discovery rate
Expected Score
Negatively expressed proteins
f
Normal
Hyperplane
Cancer
g
Input layer
Hidden layer: multilayer perceptron
Output layer
Sample #1
Sample #2
Sample #3
Sample #4
Type I cancer
Type II cancer
Type III cancer
h
Protein A
Protein C
Protein D
Protein B

similarity between two expression vectors must first be defined in order to group them. Calculations of distance are based on a "distance metric," typically classified into two types: metric and semi-metric (Quackenbush 2001). Next, decisions are made to determine which clusters should be joined. The distances, or similarities, between clusters is calculated using linkage methods. The "average linkage" algorithm works most favorably with most normalized microarray data (Shannon et al. 2003). The analysis starts with each variable in its own cluster. The algorithms then search for the pair of variables that have the smallest distance between them and merge these into a cluster. The distance metric is then repeatedly recalculated until a single cluster remains. The final result is displayed as a dendrogram, in which branch lengths reflect the degree of similarity between variables. The result can also be displayed as a "heat map" where a grid of color points provides a visual representation of protein expression. Hierarchical clustering has also been used for data mining from proteomic profiles: Using a two-way clustering algorithm, hepato-cellular carcinoma has been successfully differentiated from chronic liver disease (Poon et al. 2003).

There are two potential problems associated with hierarchical clustering. Firstly, unlike supervised statistical methods, such as *t*-test and ANOVA, the *p*-values associated with each cluster are not a measure of the statistical significance of the differences between clusters (Shannon et al. 2003), thereby leaving the results subjective and open to variable interpretation. External criteria are typically used to choose the number of clusters. For example, if splitting a hierarchical tree at a particular height results in mostly early stage tumor samples in one cluster and late stage samples in the other, the split would be considered relevant. Such a split suggests that some of the proteins in the tree may be involved with the biology of the tumor, these proteins would warrant further analysis. The problem with this approach is the subjective nature of deciding which external criteria to use (e.g., staging, prognosis, response to treatment). The second issue with hierarchical cluster

Fig. 15.3 (continued) Antibody microarray data analysis techniques. (**a**) Hierarchical clustering organizes all samples (or proteins) into groups with comparable patterns. Length of branch is inversely proportional to the degree of similarity. (**b**) *k*-Means split input data into predetermined *k* clusters by minimizing the distances of cluster members from the centroid. (**c**) Self-organizing maps provide a way of representing multidimensional data in much lower dimensional spaces. SOMs find variable-sized clusters of samples or proteins that are similar to each other, given the input number of clusters to find. (**d**) Principal component analysis is used as a visualization technique to observe the trend and scatter of clinical samples or proteins when viewed along two or three principal components. (**e**) Significance analysis of microarray carries out protein-specific *t*-tests using repeat permutations to avoid parametric assumptions. Average number of false positives can be adjusted to determine positive/negative expression. (**f**) Support vector machine transforms the input data set into a higher dimension in which separation becomes easier. By increasing dimensionality, a hyperplane can be found to classify data points or prediction of an outcome parameter. (**g**) A type of artificial neural networks uses multilayered perceptrons and the back-propagation algorithm to classify data points into multiple or continuous classes. (**h**) Bayesian network shows conditional probability and causality relationships between proteins. Nodes and arcs represent the proteins and their interactions

analysis is that the algorithm will produce clusters from any data (Quackenbush 2001). Since even unrelated data will produce clusters, caution is required in interpreting the number of reliable clusters. Although several methods have been developed to estimate the number of clusters, such as optimizing Gap statistics (Hastie et al. 2000), perturbation method (Bittner et al. 2000), and an approach based on Mantel statistics (Shannon et al. 2002), the problem of selecting the correct number of clusters remains open for further study. In spite of these problems, hierarchical cluster analysis remains a powerful and popular tool for microarray analysis. The advantage of the method is in the hierarchical nature of the output; the optimal or sensible cutoff point in the dendrogram can be chosen so that the clusters making biological sense can easily be identified. Unsupervised clustering is most suited for hypothesis generation, following which other methods should be used to assess relationships between variables.

15.5.2 K-Means Clustering and Self-Organizing Maps

Alternatives to hierarchical clustering are divisive clustering approaches, such as *k*-means or SOMs, to partition data (either by proteins or samples) into groups that have similar expression patterns (Barrios-Rodiles et al. 2005). If prior knowledge about the expected number of clusters is available, *k*-means clustering (Fig. 15.3b) is a good alternative to hierarchical methods (Gulmann et al. 2006; Story et al. 2008). In *k*-means clustering, objects are partitioned into a fixed number (*k*) of clusters such that the clusters are internally similar but externally dissimilar. While no dendrograms are produced hierarchical techniques could be used on each partition of the data. Not only the number of clusters can be specified but also the number of proteins, for each cluster. For example, *k*-means clustering can be used to classify patients with two morphologically similar but clinically distinct diseases, using microarray expression patterns. By setting $k=2$, the data will be partitioned into two groups. The challenge is to determine whether there are really only two distinct groups in the data. In this case, *k*-means clustering can be paired with other techniques, such as Principal component analysis (PCA, described below), to specify *k* and group proteins into related clusters. This technique alone should only be used for low-dimensional data sets of proteins, where a specified number of clusters can be appropriately selected.

A variation of the *k*-means method that allows samples to influence the location of neighboring clusters is known as SOM (Fig. 15.3c). This technique is particularly valuable for describing the relationships between clusters (Toronen et al. 1999). A SOM assigns proteins to a series of partitions on the basis of the similarity of their expression vectors to reference vectors that are defined for each partition (Kohonen 1995). It is the process of defining these reference vectors that distinguishes SOMs from *k*-means clustering. As with *k*-means clustering, the user has to rely on some other source of information to determine the number of clusters that best represent the available data. This artificial neural network-based clustering approach can also

be trained to recognize and classify complex patterns, for example discriminate extensive proteomic profiles of sera from cancer patients from sera of disease-free individuals (Poon et al. 2003).

15.5.3 Principal Component Analysis

PCA is a multivariate statistical tool for reducing the number of variables in high-dimensional data by identifying a subset of proteins that is responsible for the majority of observed differences (Joliffe 1986). The principle behind PCA involves presenting the data points in a three-dimensional cloud and rotating it so that it can be viewed from different perspectives (Fig. 15.3d). The technique finds the best views from which data can be separated into groups. Thus, PCA picks out patterns in the data while reducing the effective dimensionality of protein-expression space without significant loss of information.

PCA creates perpendicular axes (called principle components) in the three-dimensional cloud of data. The first principal component (P1) is the longest axis and contains the greatest fraction of the overall variance in the data. Each succeeding axis, or component, has increasingly smaller fractions of the remaining variability. There will be the same number of axes as there are variables (e.g., proteins). By plotting any two principal components, a two-dimensional view can be constructed that reflects the relative variation of the multidimensional data. Data points located near each other have similar characteristics (e.g., patterns of protein expression). Additionally, each variable is assessed to establish its contribution to the overall distribution of the data set. If a variable has a high correlation with a component, it has a strong influence on the distribution of the data. Thus, PCA will indicate which variables in a data set are important and which are of little consequence (Gulmann et al. 2006; Raychaudhuri et al. 2000).

PCA has been applied to antibody microarrays, to classify leukemia and lymphoma patients (Belov et al. 2006) and identify serum-protein profiles associated with lung cancer (Gao et al. 2005). In most implementations of PCA, it is difficult to accurately set the precise boundaries of distinct clusters in the data or to define proteins belonging to each cluster. However, PCA is a powerful tool when combined with other discriminative methods, such as *k*-means and SOMs, that require the user to specify the number of clusters. PCA can also be used to identify technical variables of importance, for example a batch effect where clusters in the PCA plot correspond to different batches of microarrays.

15.6 Differential Expression

The techniques discussed so far are unsupervised methods for identifying patterns of protein expression. Supervised methods provide an alternative or complementary approach if there is prior information for predicting clustering. Antibody

microarray analysis incorporates some of the simplest heuristic approaches for the identification of differentially expressed proteins. Most analyses are based on setting a threshold on the observed fold-change differences in expression between the state under study (e.g., disease or treatment) and the control. However, determination of fold changes simply addresses how the mean is behaving and does not consider variability in the measurements. A protein may exhibit a large fold-change, but may be unstable and thus unreliable. Statistical methods that take into account deviations in the mean as well as random variability in the system are preferred. A wide range of univariate feature ranking techniques are used for microarray data analysis. These techniques can be divided into two classes: parametric and model-free (nonparametric) methods.

15.6.1 Parametric Analysis

The student's *t*-test and analysis of variance (ANOVA) are among the most widely used techniques in microarray analysis. Parametric tests are more powerful and efficient than nonparametric tests. However, to use parametric approaches, fundamental statistical assumptions about sample distributions must be made. Microarray data may not conform to the assumption of a normal distribution, and have small numbers of replicates. The use of *t*-test or ANOVA analyses in their basic form often overlook the homogeneity of variance in the data (Jafari and Azuaje 2006). Modifications of the standard *t*-test to better deal with the small sample size and inherent noise of protein-expression data sets include a number of *t*- or *t*-test-like statistics (differing primarily in the way the variance is estimated). The *t*-test proposed by Welch (1947) is specifically designed to handle the possibility of having unequal variances (Pan 2002). Ruxton (2006) has shown the unequal variance *t*-test can effectively replace the student's *t*-test and Wilcoxon rank sum test (nonparametric version of *t*-test) for ranked data with non-normal distribution. Even when population variances are equal, the unequal variance *t*-test may perform just as well as the *t*-test and Wilcoxon rank sum test (Ruxton 2006).

ANOVA is a natural tool for studying data from experiments with multiple factors (Kerr et al. 2000). In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal. A fundamental assumption of ANOVA is that there exists a scale on which the various effects are additive. The crucial issue with ANOVA for microarray analysis is deciding whether these effects should be treated as fixed or random. Fixed effects are thought of as unknown constants, for example, the application of one or more treatments to the subjects. In contrast, random effects arise through random processes, where various factor levels are inherent to a large population. Some studies prefer the use of fixed effect models as a starting point (Kerr and Churchill 2007), but others argue that random effects are more appropriate in some cases (Wolfinger et al. 2001). Kerr and Churchill were hesitant of employing standard methods for random effects in their data, as the parameter

empirical distributions were decidedly non-normal. On the other hand, Wolfinger et al. (2001) performed ANOVA with random effects for the "blocking effects" in microarrays. Kerr and Churchill (2001) acknowledged the functionality of random effects and suggested they are more appropriate for exploratory microarray experiments. They also suggested that protein effects and protein interactions, including the effects of interest, should be treated as random. An advantage of model-based data analysis such as ANOVA is that the model helps the analysts explore the data. If one finds a model inadequate, discovering the source of the inadequacy can help to determine variations and bias in the data. Biological data are inherently variable, and statistical inference is required to draw conclusions.

15.6.2 Nonparametric Analysis

Due to uncertainty about the true underlying distribution of many protein-expression levels and difficulties in validating distributional assumptions because of small sample sizes, nonparametric methods have been widely proposed as an attractive alternative because they make less stringent model-free distributional assumptions (Troyanskaya et al. 2002).

The Wilcoxon rank sum test (or Mann-Whitney *U* test) is the classical nonparametric alternative for the *t*-test. It is more reliable for non-normal data and is thus well suited whenever the presence of outliers is suspected. However, the trade-off for robustness is the loss of power, as demonstrated by a significant reduction in differential gene expression (Thomas et al. 2001). The Wilcoxon test requires only that the samples being compared have distribution functions with the same shape (their location parameters may vary). This implies, strictly speaking, that it is not applicable if the expression levels of a protein have unequal variances under the two conditions.

Rank product (RP) is a nonparametric test developed specifically for detection of differentially expressed genes or proteins in microarray experiments (Breitling et al. 2004). It scores genes on the basis of their ranks in multiple comparisons. It is particularly well suited for noisy data sets with a small number of replicates, such as those encountered in the biological laboratory. Breitling and Herzyk (2005) used simulated microarray data to compare RP with Wilcoxon rank sum and *t*-statistics. RP outperformed both tests in data sets with a small number of replicates and non-normally distributed noise or lack of sample homogeneity. For a small sample size RP was on a par with *t*-statistic and Wilcoxon rank sum (Breitling and Herzyk 2005). The main weakness of RP is its sensitivity to variations in gene- or protein-specific variance, namely, the higher variance of weakly expressed genes or proteins. These limitations could be reduced by variance stabilizing normalization techniques or by using average ranks, that are less sensitive to variability in gene-specific variances.

15.6.3 Shrinkage Methods

A specific class of methods use a modified versions of the standard *t*-test to identify fidifferential genes or proteins. These methods, including the empirical Bayes (EB) method (Efron and Tibshirani 2002), linear models for microarray data (Limma) (Smyth 2004), significance analysis of microarray (SAM) (Tusher et al. 2001), and the mixture model method (MMM) (Pan et al. 2003), use different techniques to alleviate the problem of small sample sizes in microarray studies, enhancing the robustness against outliers. The idea is that with replicates of microarrays, one can estimate the distribution of random errors without strong parametric assumptions, making it possible to distinguish genuinely altered protein expression from noise with high confidence.

Such procedures often estimate prior distributions from the data, in direct contrast to standard Bayesian methods, where the prior distribution is fixed before data are collected. These techniques are primarily designed to "borrow information" across proteins and experimental conditions, in the hope that the borrowed information will lead to better estimates and, hence, more stable inferences. Some of these techniques, including Limma and SAM, enable adjusting for multiple testing based on a local FDR and associated *q*-statistic (SAM; Fig. 15.3e). The *q*-statistic gives an estimate of the probability of falsely identifying a protein as "significant" within all the groups of proteins that have *q* values lower than that for the protein in question. The local FDR gives a measure of the probability that a given feature is identified as significant by random chance. Performance comparison of these techniques have been discussed in detail in the literature (Pan 2002; Pan et al. 2003; Schwender et al. 2003). In brief, EB may produce slightly more conservative significance levels as only the lower bound of posterior probability is actually estimated, SAM may not be able to estimate small false positives well because false positives are estimated by a finite number of simulated null scores, while SAM is more robust to the use of null statistics than MMM. These procedures all take advantage of the existence of replicate samples to construct the null scores and a large number of proteins to estimate the null distribution (Pan et al. 2002), and therefore, they should only be applied to microarrays with very large number of proteins (over 500) to obtain meaningful results.

15.7 Multiple Testing

False positives tend to be particularly problematic in high-throughput proteomic studies where many candidates must be statistically tested. The strength of antibody microarrays, their ability to screen hundreds of proteins at a time, also creates many opportunities for spurious discoveries. For example, when comparing two conditions (such as normal vs. cancer) with ten samples per condition, the more proteins we test, the more likely we are to observe false "significantly" expressed proteins.

Therefore, a multiple testing correction procedure is required to adjust the statistical confidence measures based on the number of tests performed, regardless of which test is used.

The simplest and most widely used method is the Bonferroni adjustment, where a score is deemed significant only if the *p* value multiplied by the number of tests is below a set threshold (e.g., 0.01 or 0.05). The Bonferroni adjustment controls the familywise error (FWE): the probability that at least one protein is called significantly expressed while in reality it is not. In some microarray settings, with large numbers of tests and small sample numbers, this correction is often too strict (Noble 2009; Armstrong and van de Wiel 2004). An alternative is the control of false discovery rate (FDR); it is a less conservative procedure than Bonferroni correction, with greater power than FWE control but at a cost of increasing the likelihood of obtaining false positives. FDR can be computed from a list of *p* values using the Benjamini-Hochberg procedure by multiplying the univariate *p* values by the number of proteins and dividing by the rank of the *p* value (Benjamini and Hochberg 1995). In general, for a fixed significance threshold and fixed null hypothesis, performing multiple testing correction by means of an FDR estimation will always yield at least as many significant scores as using Bonferroni adjustment.

The choice of a multiple testing correction method depends upon the cost of false positives or false negatives. For example, FDR may be appropriate if a collection of follow-up validation experiments is planned and the user is willing to tolerate having a fixed percentage of those experiments as false positives. Alternatively, if follow-up focuses on a single sample, then the Bonferroni adjustment is more appropriate (Noble 2009). In summary, in any experimental setting in which multiple tests are performed, *p* values must be adjusted appropriately. Generally, antibody microarrays are less susceptible to multiplicity problems when compared to their DNA/cDNA counterparts. They contain fewer spots, antibodies are often selected based on biological knowledge of the test samples, and the smaller number of tests leads to less severe corrections.

15.8 Classification

Supervised approaches are well suited to categorizing samples into known phenotypes. The main goals for investigators are to identify the proteins that are most important for the classification and to develop a robust classifier with validation procedures that can successfully handle blinded test data. The results can provide a clinical tool for diagnosing a disease, predicting the outcome or treatment response, and/or insights into the underlying molecular mechanisms.

Supervised machine learning algorithms can be trained to recognize and characterize complex patterns, provided information is already available for the subclassification of test samples (e.g., cancer patients vs. normal controls). Machine learning approaches produce the best results when large numbers of examples (training set) are used to adapt the parameters of a model that can then be used for

performing predictions or classification on data. Support vector machines (SVMs) (Cristianini and Shawe-Taylor 2000), artificial neural networks (ANNs) (Bishop 1995), and Bayesian networks (BNs) (Friedman et al. 2000) are widely used examples for pattern recognition. This is achieved by adjusting the parameters of the data-fitting models by a process of error (e.g., misclassification) minimization through learning from experience with training samples. A report by Ringner et al. (2002) discusses the advantages and limitations of SVMs and ANNs. Generally, SVMs are better at handling the high-dimensional array data, whereas ANNs require some preprocessing to avoid overfitting. On the other hand, the results from ANNs allow for a straightforward probability interpretation, and ANNs are more easily generalized to multi-class classification problems. In addition, ANNs can be used to classify samples not only according to a dichotomous distinction (such as good prognosis vs. poor prognosis) but also according to more sample-specific phenotypes such as time of survival (a continuous variable).

15.8.1 Support Vector Machines

The basic idea of SVM is that the inputs are formulated as feature vectors, which are mapped into a feature space by using the kernel function. Finally, a division is computed in the feature space to optimally separate two classes of training vectors (Fig. 15.3f). SVM has been widely used in antibody microarray studies to identify complex patterns in the serum proteome of various patients, such as metastatic breast cancer (Carlsson et al. 2008), early pancreatic cancer (Ingvarsson et al. 2008), and ovarian cancer (Mor et al. 2005), systemic lupus erythematosus, and systemic sclerosis (Carlsson et al. 2011). Ingvarsson et al. (2008) chose not to perform any parameter tuning to avoid overfitting and filtered out nondiscriminatory proteins using the Wilcoxon rank score, followed by ranking proteins based on their predictive accuracy in a leave-one-out cross-validation scheme. Carlsson et al. (2011) also avoided parameter tuning but did not filter the data before training the SVM. All antibodies on the microarray were used, and proteins were ranked according to their contribution to defining the decision hyperplane, that is, according to their importance in classifying the sample. This unfiltered approach, in principle, allows each protein to be ranked for each sample. Since it is likely that distinct clinical behaviors are explained by different molecular mechanisms in different patients, this approach has the potential to use machine learning methods to profile an individual's proteome, thereby creating possibility of personalized medicine.

15.8.2 Artificial Neural Networks

The first generation of ANNs, called perceptrons, was simple linear logistic regression methods. More elaborate ANNs in the form of a multilayer perceptron were another machine learning approach that has proven to be powerful when classifying

tumor array-based mRNA expression data in DNA arrays (Khan et al. 2001). A multilayered perceptron consists of a set of layers of perceptrons, and for high-dimensional data, a large number of perceptrons are needed (Fig. 15.3g). The more perceptrons there are in the ANN, the more training samples are required for calibration. This becomes a problem for array data where the number of samples is much lower than the number of measured proteins which leads to a high risk of over-fitting. This problem may be overcome by reducing the dimensions of the data, using a dimensional reduction algorithm such as PCA, or by selecting a smaller set of proteins as input to the classifier in a supervised way by using a discriminatory score. This is followed by careful monitoring of the learning process using a cross-validation scheme to avoid over-training (Khan et al. 2001). So far, ANN has had limited use for antibody microarray data, but its versatility has been proven for gene expression microarrays. ANN was used to investigate the phenotype associated with estrogen receptor (ER) α status in human breast cancer; it was found that ANNs could accurately classify the tumors into ER-positive and ER-negative samples (Gruvberger et al. 2001). The advantage of ANN over SVM is that it can easily be adopted to predict continuous values instead of classes. For example, it can be used to predict protein levels of the ER receptor instead of classifying samples into binary ER classes. Such a prediction method can potentially be used to gain further insights into the relevant proteins and may be useful for patient outcome prediction, where survival times may be important.

15.8.3 Bayesian Networks

BN computational analysis can be used to identify patterns of protein expression over and across entire pathways (Fig. 15.3h). Implementation of BN has been for modeling protein signaling pathways (Friedman 2004) and for the discovery of novel molecular interactions (Sachs et al. 2005). These analyses combine all the features of BN, that is, the ability to learn from incomplete noisy data, to combine both expert knowledge and data to derive a suitable network structure, and to express causal relationships. BN analysis is ideal for microarray data, which can often contain nonspecific binding and/or high signal-to-noise ratio, as the parameters for BNs may be learned even when the training data set is incomplete (i.e., the values of some variables in some cases are unknown). Commonly, the expectation-maximization (EM) algorithm is used, that estimates the missing values by computing the expected values and updating parameters by re-using these expected values as if they were observed values. BN models qualitatively demonstrate the direct and indirect causative influences of pathway components on each other in the form of an influence graphic diagram containing nodes and arcs. Nodes represent the measured variables, and interconnecting arcs represent statistically meaningful relationships between them, both in a linear and nonlinear fashion. BN offers a solution to the common problem of overfitting in machine learning, by providing mechanisms for describing uncertainty and for adapting the number of parameters to the size of the data. This elevates the need for a large training data set when using complex learned

models. The ability of BN to identify probabilistic dependence relationships is particularly important in the discovery of new therapeutic targets. It also has potential in the field of phosphoproteomics, as inhibition of any one pathway branch may have significant consequential effects in other pathways. It has been applied to flow cytometry with success (Sachs et al. 2005) and could be a valuable technique for modeling complex interactions in antibody microarray data.

15.9 Open-Source Microarray Analysis Software

Most of the normalization and data analysis techniques discussed in this chapter can be accessed on open-source microarray analysis software. Table 15.2 presents some commonly used microarray analysis packages and their main features. TM4 (http://www.tigr.org/software), ARMADA (http://www.grissom.gr/armada), Bioconductor (http://www.bioconductor.org), BASE (http://base.thep.lu.se), Cluster (http://rana.lbl.gov/EisenSoftware.htm), and Java Treeview (jtreeview.sourceforge.net) represent different approaches to the same problem, and each has advantages and disadvantages. TM4 (Saeed et al. 2003) offers users a graphical interface that is easy to navigate, and the program architecture provides great flexibility for development of new algorithms. However, implementation of new statistical tools requires the creation of analysis libraries, and users have to install new software releases. ARMADA (Chatziioannou et al. 2009) provides a similar comprehensive data analysis program with a simple graphical user interface (GUI). The Bioconductor Web site (Gentleman et al. 2004) is a repository for packages written to analyze different types of biological data in the *R* statistical environment. It allows for the rapid development and dissemination of new methods. Perhaps the biggest problem with *R* language is that fist-time users can be discouraged by the complexity of the *R* command-line environment. Packages like tkWidgets may simplify the interface and provide GUI elements in *R*. However, the advantage of learning *R* is that one then has access to advanced statistical methods if standard programs are insufficient to analyze your experimental data. BASE (Saal et al. 2002) minimizes the software update problem by using a Web-based approach and, as such, could easily integrate the TM4 utilities, but it loses a good deal of the graphical functionality that local applications can provide. Cluster 3.0 (Eisen et al. 1998) and Java Treeview (Saldanha 2004) are specialized programs that provide powerful clustering analyses and visualization options. Although lacking in versatility, they provide more options in visualization of cluster data than any other program. The main limitation of these programs is that they were created for DNA/cDNA arrays and, as a result, image analysis can be challenging. In most cases, it is easier to import data already quantified using image analysis software (e.g., ImageJ or ImageQuant) in tabular form (Excel spreadsheet or tab-delimited text formats). Also, all of the integration databases are linked to genomic libraries and offer little use for proteomic analysis. These issues may only be temporary, as the advantage of open-source software is the availability of their source codes, that allows the community to modify the program and implement new functionality better suited for antibody microarray analysis.

Table 15.2 Some of the open-source microarray analysis software packages

Software	Function	Features/analytic techniques
TM4: Java-based system for microarray expression analysis		
TIGR Spotfinder	Semiautomated image analysis software	Dynamic thresholding algorithm for spot intensity determination, provides quality measures for each spot, and interfaces directly with AGED database (gene expression database)
MIDAS	Data normalization and filtering	Background- and quality-control trimming, lowess/loess normalization, replicate analysis and filtering, and SD regularization
MeV	Data visualization and analysis	Large number of techniques, including hierarchical clustering, *k*-means, SOM, PCA, ANOVA, SAM, SVM, *t*-test, and Bayesian networks
ARMADA: MATLAB implemented multi-analysis platform for microarray analysis		
ARMADA	Automated quality control, analysis, annotation, and visualization of microarray data	Image analysis: background subtraction, signal-to-noise ratio calculations, and spot quality filtering Normalization: global mean/median, lowess/loess, and quadratic/robust variants of lowess/loess, and rank invariant Data analysis: hierarchical clustering, *k*-means, fuzzy C-means, PCA, LDA, *k*-nearest neighbors, and SVM
Bioconductor: statistical analysis tools developed in R		
Biobase	Base functions for package	exprSet class provides a systematic representation of microarray expression data and designed to follow the MIAME standards
Marray package (classes, input, norm, plots)	Data handling and normalization	Diagnostic plots of microarray spots: boxplots, scatter plots, and spatial color images Normalization: 2D spatial, loess, internal control spots, and spiked-in controls
Genefilter, multtest, annotate	Data filtering and identify differentially expression	Filtration: ANOVA and Cox model Multiple testing procedure: familywise error rate and false discovery rate
Limma	Linear models for microarray data	Data analysis, linear models, and differential expression for microarray data

(continued)

Table 15.2 (continued)

Software	Function	Features/analytic techniques
BASE: Web-based approach to data management and annotation		
BASE	Database for ancillary annotation and microarray data. Plug-ins enable normalization and analysis	Plug-ins: Normalizer application—global mean/median normalization, lowess MDS module—multidimensional scaling 3D data viewer—visualize and explore data as 3D projection MeV integrated for data analysis
Cluster and Treeview: computational and graphical environment for microarray data analysis		
ScanAlyze	Microarray image analysis	Semiautomatic image processing, grid definition, and complex pixel/spot analysis
Cluster 3.0	Data filtering and analysis	Log transform, mean/median centering, hierarchical clustering, *k*-means, SOM, and PCA
Java Treeview	Cluster visualization	Numerous dendrogram visualization options, scatter plots, karyoscope plot, and dendrogram-like view of sequence data

SOM self-organizing map, *ANOVA* analysis of variance, *PCA* principal component analysis, *SAM* significance analysis of microarray, *SVM* support vector machine, *kNN* *k*-nearest neighbor algorithm, *LDA* linear discriminant analysis, *MIAME* minimum information about a microarray experiment

References

Angenendt P, Glokler J, Murphy D, Lehrach H, Cahill DJ. Toward optimized antibody microarrays: a comparison of current microarray support materials. Anal Biochem. 2002;309:253–60.

Armstrong NJ, van de Wiel MA. Microarray data analysis: from hypotheses to conclusions using gene expression data. Cell Oncol: Official J Int Soc Cell Oncol. 2004;26:279–90.

Barrios-Rodiles M, Brown KR, Ozdamar B, Bose R, Liu Z, Donovan RS, Shinjo F, Liu Y, Dembowy J, Taylor IW, et al. High-throughput mapping of a dynamic signaling network in mammalian cells. Science. 2005;307:1621–5.

Belov L, de la Vega O, dos Remedios CG, Mulligan SP, Christopherson RI. Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. Cancer Res. 2001;61:4483–9.

Belov L, Mulligan SP, Barber N, Woolfson A, Scott M, Stoner K, Chrisp JS, Sewell WA, Bradstock KF, Bendall L, et al. Analysis of human leukaemias and lymphomas using extensive immunophenotypes from an antibody microarray. Br J Haematol. 2006;135:184–97.

Benjamini Y, Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J R Stat Soc Ser B-Methodol. 1995;57:289–300.

Bishop CM. Neural networks for pattern recognition. Oxford: Clarendon; 1995.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 2000;406:536–40.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.

Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J Bioinform Comput Biol. 2005;3:1171–89.

Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004;573:83–92.

Carlsson A, Wingren C, Ingvarsson J, Ellmark P, Baldertorp B, Ferno M, Olsson H, Borrebaeck CA. Serum proteome profiling of metastatic breast cancer using recombinant antibody microarrays. Eur J Cancer. 2008;44:472–80.

Carlsson A, Wuttge DM, Ingvarsson J, Bengtsson AA, Sturfelt G, Borrebaeck CA, Wingren C. Serum protein profiling of systemic lupus erythematosus and systemic sclerosis using recombinant antibody microarrays. Mol Cell Proteomic: MCP. 2011;10:M110 005033.

Chatziioannou A, Moulos P, Kolisis FN. Gene ARMADA: an integrated multi-analysis platform for microarray data implemented in MATLAB. BMC Bioinformatics. 2009;10:354.

Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.

Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23:70–86.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95:14863–8.

Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004;303:799–805.

Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol J Comput Mol Cell Biol. 2000;7:601–20.

Gao WM, Kuick R, Orchekowski RP, Misek DE, Qiu J, Greenberg AK, Rom WN, Brenner DE, Omenn GS, Haab BB, Hanash SM. Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. BMC Cancer. 2005;5:110.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.

Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res. 2001;61:5979–84.

Gulmann C, Sheehan KM, Kay EW, Liotta LA, Petricoin 3rd EF. Array-based proteomics: mapping of protein circuitries for diagnostics, prognostics, and therapy guidance in cancer. J Pathol. 2006;208:595–606.

Hallborn J, Carlsson R. Automated screening procedure for high-throughput generation of antibody fragments. BioTechniques. 2002;(Suppl):30–7

Hamelinck D, Zhou H, Li L, Verweij C, Dillon D, Feng Z, Costa J, Haab BB. Optimized normalization for antibody microarrays and application to serum-protein profiling. Mol Cell Proteomic MCP. 2005;4:773–84.

Hanes J, Schaffitzel C, Knappik A, Pluckthun A. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. Nat Biotechnol. 2000;18:1287–92.

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 2000; 1:RESEARCH0003.

Ingvarsson J, Wingren C, Carlsson A, Ellmark P, Wahren B, Engstrom G, Harmenberg U, Krogh M, Peterson C, Borrebaeck CA. Detection of pancreatic cancer using antibody microarray-based serum protein profiling. Proteomics. 2008;8:2211–19.

Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak. 2006;6:27.

Joliffe T. Principle components analysis. Berlin: Springer; 1986.

Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 2001;98:8961–5.

Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet Res. 2007;89:509–14.

Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol J Comput Mol Cell Biol. 2000;7:819–37.
Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7:673–9.
Kohonen T. Self organizing maps. Berlin: Springer; 1995.
Kukar T, Eckenrode S, Gu Y, Lian W, Megginson M, She JX, Wu D. Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins. Anal Biochem. 2002;306:50–4.
Kusnezow W, Jacob A, Walijew A, Diehl F, Hoheisel JD. Antibody microarrays: an evaluation of production parameters. Proteomics. 2003;3:254–64.
Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, Vandesompele J. A novel and universal method for microRNA RT-qPCR data normalization. Genome Biol. 2009;10:R64.
Miller JC, Zhou H, Kwekel J, Cavallo R, Burke J, Butler EB, Teh BS, Haab BB. Antibody microarray profiling of human prostate cancer sera: antibody screening and identification of potential biomarkers. Proteomics. 2003;3:56–63.
Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, Yue L, Bray-Ward P, Ward DC. Serum protein markers for early detection of ovarian cancer. Proc Natl Acad Sci U S A. 2005;102:7677–82.
Noble WS. How does multiple testing correction work? Nat Biotechnol. 2009;27:1135–7.
Olle EW, Sreekumar A, Warner RL, McClintock SD, Chinnaiyan AM, Bleavins MR, Anderson TD, Johnson KJ. Development of an internally controlled antibody microarray. Mol Cell Proteomic MCP. 2005;4:1664–72.
Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics. 2002;18:546–54.
Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol. 2002; 3:research0022.
Pan W, Lin J, Le CT. A mixture model approach to detecting differentially expressed genes with microarray data. Funct Integr Genomics. 2003;3:117–24.
Pollard HB, Ji XD, Jozwik C, Jacobowitz DM. High abundance protein profiling of cystic fibrosis lung epithelial cells. Proteomics. 2005;5:2210–26.
Pollard HB, Srivastava M, Eidelman O, Jozwik C, Rothwell SW, Mueller GP, Jacobowitz DM, Darling T, Guggino WB, Wright J, et al. Protein microarray platforms for clinical proteomics. Proteomics Clin Appl. 2007;1:934–52.
Poon TC, Yip TT, Chan AT, Yip C, Yip V, Mok TS, Lee CC, Leung TW, Ho SK, Johnson PJ. Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. Clin Chem. 2003;49:752–60.
Quackenbush J. Computational analysis of microarray data. Nat Rev Genet. 2001;2:418–27.
Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pacific Symp Biocomput. 2000:455–66.
Ringner M, Peterson C, Khan J. Analyzing array data using supervised methods. Pharmacogenomics. 2002;3:403–15.
Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behav Ecol. 2006;17:688–90.
Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol. 2002; 3:SOFTWARE0003.
Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308:523–9.
Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003;34:374–8.
Saldanha AJ. Java Treeview–extensible visualization of microarray data. Bioinformatics. 2004;20:3246–8.

Schweitzer B, Wiltshire S, Lambert J, O'Malley S, Kukanskis K, Zhu Z, Kingsmore SF, Lizardi PM, Ward DC. Immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. Proc Natl Acad Sci U S A. 2000;97:10113–19.

Schwender H, Krause A, Ickstadt K. Comparison of the empirical Bayes and the significance analysis of microarrays. Technical Report // Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, 2003;44. http://hdl.handle.net/10419/49325.

Shannon WD, Watson MA, Perry A, Rich K. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. Genet Epidemiol. 2002;23:87–96.

Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. Pharmacogenomics. 2003;4:41–52.

Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004; 3:Article3.

Srivastava M, Eidelman O, Jozwik C, Paweletz C, Huang W, Zeitlin PL, Pollard HB. Serum proteomic signature for cystic fibrosis using an antibody microarray platform. Mol Genet Metab. 2006;87:303–10.

Story CM, Papa E, Hu CC, Ronan JL, Herlihy K, Ploegh HL, Love JC. Profiling antibody responses by multiparametric analysis of primary B cells. Proc Natl Acad Sci U S A. 2008;105:17902–7.

Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 2001;11:1227–36.

Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett. 1999;451:142–6.

Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002;18:1454–61.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98:5116–21.

Wang Y, Wu TR, Cai S, Welte T, Chin YE. Stat1 as a component of tumor necrosis factor alpha receptor 1-TRADD signaling complex to inhibit NF-kappaB activation. Mol Cell Biol. 2000;20:4505–12.

Welch BL. The generalisation of student's problems when several different population variances are involved. Biometrika. 1947;34:28–35.

White CN, Chan DW, Zhang Z. Bioinformatics strategies for proteomic profiling. Clin Biochem. 2004;37:636–41.

White SL, Belov L, Barber N, Hodgkin PD, Christopherson RI. Immunophenotypic changes induced on human HL60 leukaemia cells by 1alpha,25-dihydroxyvitamin D3 and 12-O-tetradecanoyl phorbol-13-acetate. Leuk Res. 2005;29:1141–51.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol J Comput Mol Cell Biol. 2001;8:625–37.

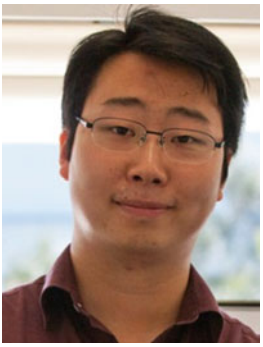
Wu X, Liu H, Liu J, Haley KN, Treadway JA, Larson JP, Ge N, Peale F, Bruchez MP. Immunofluorescent labeling of cancer marker Her2 and other cellular targets with semiconductor quantum dots. Nat Biotechnol. 2003;21:41–6.

Wylie D, Shelton J, Choudhary A, Adai AT. A novel mean-centering method for normalizing microRNA expression from high-throughput RT-qPCR data. BMC Res Notes. 2011;4:555.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30:e15.

Yang JY, Zong CS, Xia W, Wei Y, Ali-Seyed M, Li Z, Broglio K, Berry DA, Hung MC. MDM2 promotes cell motility and invasiveness by regulating E-cadherin degradation. Mol Cell Biol. 2006;26:7269–82.

Zhou J, Belov L, Solomon MJ, Chan C, Clarke SJ, Christopherson RI. Colorectal cancer cell surface protein profiling using an antibody microarray and fluorescence multiplexing. J. Vis. Exp. 2011;(55):e3322. DOI:10.3791/3322.



Jerry Zhou, Bachelor (Hon) Jerry Zhou is a Ph.D. candidate at the Cancer Proteomics Laboratory, University of Sydney, Australia. He completed his Bachelor of Molecular Biology and Genetics with first class Honors in 2008 and has since been investigating the potential of using antibody microarray (DotScan) to create a disease signature for colorectal cancer stratification and outcome prediction. His postgraduate research is funded by the University of Sydney Postgraduate Award and Cancer Institute New South Wales Research Scholar Award.

Larissa Belov, (Ph.D.) Dr Larissa Belov is a Senior Research Scientist with broad experience in medically oriented research and biotechnology, with specific interests in the fields of cancer, proteomics, immunology and inflammation. She has been involved in the development and utilisation of antibody microarrays (DotScan™) for cell surface immunophenotyping of viable cells populations, including leukaemia, lymphoma, colorectal cancer, melanoma and other cell types, with the aim of identifying prognostic markers and potential therapeutic targets.

Nicola Armstrong, (Ph.D.) Nicola Armstrong has focused on statistical applications in biology since she received her Ph.D. in Statistics from the University of California, Berkeley. She is currently a Senior Bioinformatics Officer at the Garvan Institute of Medical Research and holds conjoint Senior Lecturer positions in the School of Mathematics and the Prince of Wales Clinical School at the University of New South Wales. Her current research includes the analysis of genomic data in relation to clinical variables with the aim of identifying translational importance and integrative analysis of the epigenome.

Richard I. Christopherson, Ph.D., Professor. Richard Christopherson has focused on cancer research since his postdoctoral period in the USA where he was a fellow of the Damon Runyon-Walter Winchell Cancer Fund and then a special fellow of the Leukemia Society of America. He is a professor in the School of Molecular Bioscience at the University of Sydney and a scientific founder and chief scientific officer (part time) of the University spin-off company, Medsaic Pty Ltd (www.medsaic.com). His research has focused on elucidation of the mechanisms of action of anticancer drugs. His group determined the antipurine mechanism of the drug methotrexate used to treat leukemias and breast cancer. His current research includes investigation of the mechanisms of action of the anticancer drug fludarabine by identification of proteins that are differentially abundant in drug-treated cells. A CD antibody microarray (DotScan) has been developed that provides extensive surface expression profiles (immunophenotypes or disease signatures) of leukemias. The use of DotScan has been extended to profiling of colorectal cancers and melanoma, and it is being commercialized by Medsaic.

Chapter 16
Proteomics in Anaesthesia and Intensive Care Medicine

Ornella Piazza, Giuseppe De Benedictis, and Geremia F. Zito Marinosci

Abstract Proteomics is a systems-based methodology for describing the changes in protein expression present in biological samples that can occur with disease processes. Analysis of proteome allows individuating proteins that act as biomarkers, which have many potential roles in clinical practice: diagnosis, response to treatment, risk stratification and prognosis. This chapter summarises the most important proteomics application in anaesthesiology and critical care.

Keywords Biomarkers • ICU • Perioperative risk

16.1 Introduction

Proteomics has the potential of providing highly translatable results and applications for public health since proteins are the heart of human physiology and, consequently, of pathology. Proteomics is a systems-based methodology for describing the complex changes in protein expression and post-translational modifications present in the genome biological samples that can occur with disease processes. We already learnt that the genome is complex, and now the scientific community is facing the fact that proteome is even more complicated, redundant and well difficult to analyse and quantify: the exact number of human proteins is

O. Piazza, M.D. (✉)
Department of Medicine and Surgery,
Università di Salerno, Salerno, Italy
e-mail: opiazza@unisa.it

G. De Benedictis, M.D. • G.F.Z. Marinosci, M.D.
Department of Anesthesia and Intensive Care Medicine,
Università di Napoli Federico II, Naples, Italy

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_16,

unknown, but it is expected to be in the many million range; however, the real goal is understanding the fundamental principles of protein–protein interactions, enzyme catalysis and related signal transduction pathways. Aiming to maximum exemplification, we could say that a gene mutation is clinically irrelevant until a change in function of the protein end product is produced, but the next frontier is to move forward from descriptive lists of differentially expressed proteins to mechanistic insights to identify nodal points of protein network. Analysis of proteome allows to individuate proteins that act as biomarkers, which according to the WHO are "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" (WHO International Programme on Chemical Safety. Biomarkers in risk assessment: validity and validation 2001). A biomarker should be SMART, a mnemonic for Sensitive (and Specific), Measurable, Available, Responsive (and Reproducible) in a Timely fashion (Barnett and Ware 2011). New technologies are enhancing the study of proteomics (read also Chaps. 1 and 7 for detailed information) for patient-centred research strategies; this is very important since there are many potential roles for biomarkers in clinical practice: diagnosis, response to treatment, risk stratification and prognosis. This chapter summarises the most important proteomics application in anaesthesiology and critical care.

16.2 Clinical Application of Proteomics in Anaesthesiology

Despite the widespread clinical use of anaesthetics since the nineteenth century, a clear understanding of the mechanism of anaesthetic action is only just beginning to be understood in detail, in particular for inhalational anaesthetics.

Anaesthesia embraces some controlling yet reversible characteristics: immobility, muscle relaxation, amnesia, analgesia and unconsciousness. To obtain these effects, according to the "holistic vision", anaesthetic agents do not interfere with the function of limited groups of neurons but act widely in the central nervous system, from the spinal cord to the cortex, and even out of it: this enhances the difficulties to know exactly the targets and the mechanisms of action of anaesthetic drugs (Koch and Greenfield 2007).

The experiments carried out by Overton and Meyer at the turn of the nineteenth century suggested a common unitary mechanism for inhalation anaesthetics: the cell's lipid membrane was concluded to be the primary site of action of anaesthetics. This hypothesis included anaesthetic-induced volume expansion of the cell membrane, increased fluidity of cell membrane and increased lateral surface pressure (Kopp Lugli et al. 2009).

However, later experiments shifted the focus of anaesthetic action to proteins: now it is clear that anaesthetic drugs, similar to the most of other drugs, act by binding several membrane and cytosolic target proteins (Eckenhoff and Johansson 1997; Franks 2006). The fact that proteins also contain lipophilic domains supports the continuing importance of the Meyer–Overton correlation. Nevertheless, discovery

of a single fundamental mechanism for anaesthetic drug action has become less and less likely. According to multiple-target hypothesis, different anaesthetics produce various effects binding an increasing number of protein candidates, in particular membrane receptors and ion channels. Binding target proteins, anaesthetic drugs can modulate inhibitory and excitatory synapses by presynaptic or postsynaptic interactions and modify CNS cell metabolism (Kopp Lugli et al. 2009).

Thus, the proteocentric view of anaesthetic mechanism makes the proteomics a central science to know what are the target proteins and how anaesthetics interact with them in vivo and in vitro.

Using a combination of photolabelling, two-dimensional gel electrophoresis and mass spectrometry, it has been possible to identify anaesthetic-binding targets in the nervous tissue. In this way, it has been detected that halothane binds about 90 proteins in rat brain (Xi et al. 2004). These proteins can be classified into several functional groups, including carbohydrate metabolism, protein folding, oxidative phosphorylation, nucleoside triphosphatase, kinase activity, ion channels and membrane receptors (Pan et al. 2007).

Similar studies have reinforced the idea that mitochondria represent a target organelle for anaesthetics. Volatile anaesthetics are able to bind the components of the respiratory chain thereby reducing oxidative phosphorylation: these binding sites may contribute to anaesthetic action (Xi et al. 2004; Morgan et al. 2002).

These studies provide evidence for multiple anaesthetic-binding targets and suggest potential pathways involved in their actions.

Proteomic technologies can also highlight how anaesthetics interact with and alter protein function. Crystallographic studies have reinforced the notion that aesthetics anaesthetic act by binding to certain hydrophobic cavities on their protein targets, thereby modifying the thermodynamic stability and/or perturbing the tertiary or the quaternary structure of the target (Bhattacharya et al. 2000; Pidikiti et al. 2005). These interactions occur with low affinity and are unspecific since these hydrophobic cavities is an attractive generic for different hydrophobic ligands. In the case of halothane–nicotinic receptor interaction, halothane binds the hydrophobic cavity close to the M2–M3 loop. This binding modulates the dynamics of the M2–M3 loop, which is implicated in allosterically transmitting the effect to the channel gate. In potassium channels, anaesthetic molecules potentiate the open conformation (Vemparala et al. 2010). Proteomic experiments have also demonstrated that phosphorylation state of a target protein can modulate affinity for anaesthetic drugs (Koch and Greenfield 2007).

Another proteomics application to understand the pharmacodynamics of anaesthetic drugs concerns the study of protein expression profile (i.e. proteome) changes in tissues or fluids after exposure to an anaesthetic agent. Using proteomic technologies, such as two-dimensional electrophoresis, mass spectrometry (MS) and matrix-assisted laser desorption/ionisation–time-of-flight mass spectrometry (MALDI-TOF MS), most of the studies have demonstrated that anaesthetics are able to significantly modify proteome expression in vivo and in vitro.

As expected, the proteins whose expression level is up- or downregulated by anaesthetics are related to many biological neuronal activities, like cytoskeletal/

neuronal growth, cellular metabolism, antioxidation, signalling, synaptic plasticity and vesicle transport (Atkins and Johansson 2006).

In the Kalenka et al. study, 3 h of 1 MAC sevoflurane anaesthesia altered protein level expression of 11 cytosolic proteins, when analysed immediately after exposure, and 17 cytosolic proteins 72 h after the exposure. A first group of proteins altered after sevoflurane anaesthesia includes enzymes well known for their metabolic functions, like pyruvate kinase, transketolase and aconitase, according to the downregulation of energy-dependent cellular pathways induced by volatile anaesthetics. A second group includes cellular stress-responsive proteins. Ubiquitin carboxy-terminal hydrolase L1 (UCH L1) was strongly upregulated by sevoflurane. This protein serves as a multifunctional protein involved in proteolysis, apoptosis and synaptic function. Also 90-kDa heat shock protein (HSP90) level was modified by sevoflurane. This molecular chaperone plays a critical role in cell growth, signalling and neurotransmitter release. A third functional group of differentially regulated proteins is related to vesicle formation, transport and exocytosis, resulting in synaptic fatigue (Kalenka et al. 2007).

Propofol anti-inflammatory properties were clinically evident but not well understood. Proteomics analysis has demonstrated that propofol upregulates the protein annexin A1 thereby inhibiting the phosphorylation level of p38 and so the release of IL-1β, IL-6 and TNF-α in serum mononuclear cells (Tang et al. 2011).

It has also been found that propofol and sevoflurane have different, sometimes opposed, effects on proteome changes in brain rat. These data suggest different underlying mechanisms of proteomic regulation for different anaesthetics (Tsuboko and Sakamoto 2011).

Other studies have shown that inhaled anaesthetics elicit region-specific and age-related changes in protein expression in the mammalian brain (Pan et al. 2008; Duan et al. 2009).

Sevoflurane and isoflurane protect the heart against ischemia/reperfusion injury induced by mitochondria dysfunctions (i.e. delayed myocardial preconditioning), but the mechanism of action is not well understood. Proteomic studies have demonstrated that sevoflurane can modify mitochondria proteome in animal models of myocardial ischemia: the differentially expressed proteins are related to bioenergetic balance, suggesting an enhanced capacity to preserve ATP levels in ischemia (Xiao et al. 2011). Using phosphoproteomic approach, it has been shown that isoflurane influences the phosphorylation state of several mitochondria proteins in myocardial tissue. The same study was also able to detect a novel phosphorylation site in adenine nucleotide translocator-1 (ANT1), which may play an integral role in regulating ANT1 function and cellular respiration (Feng et al. 2008).

These are several examples that show how proteomic studies applied to anaesthesiology can help to translate from biochemical to clinical milieu and vice versa, that is, to understand the biochemical underlying mechanisms of action to explain drug effects previously highlighted only in the clinical setting.

Obviously, the differentially expressed proteins are not a specific target for anaesthetics. The expression changes may indicate changes in gene expression level that can be directly induced by the drug or indirectly, as a compensatory effort of the cell to adapt to the anaesthetic.

The same approach (study of the proteome change caused by anaesthetic drugs) has been useful to demonstrate that changes of protein expression persist in the tissues also after anaesthetic exposure has finished. Thus, in the rat hippocampus, 12 proteins did not restore to the basic level until the 7th day after propofol anaesthesia (Zhang et al. 2009). Similar results have been found using desflurane or isoflurane (Fütterer et al. 2004; Kalenka et al. 2010). As some of these proteins seem to be involved in Alzheimer disease, this persistent proteome change may be the pathogenetic basis for the development of postoperative cognitive dysfunction (POCD) (Kalenka et al. 2010).

An example of how new proteomic technologies can usefully integrate the "old" ones, like Western and Northern blotting, is provided by Hirota's study (Hirota et al. 2003). To examine the effect of halothane over the hypoxia-induced cellular responses, they analysed the activation of HIF-1 using molecular biological methods including Western blotting, Northern blotting, RT-PCR, gene-reporter assay and in vitro protein–protein interaction assay. Halothane significantly suppressed hypoxia-induced HRE-dependent gene expression, hypoxia-induced accumulation of HIF-1α protein and mRNA accumulation of its target genes. Moreover, it blocked hypoxia-induced HIF-1α transactivation activity by hyperactivating both the hypoxia sensors, which are two hydroxylases requiring molecular O_2 and Fe^{2+} for their enzymatic activity. The further application of proteomic strategy will be necessary to better understand the cellular hypoxia sensing machinery and the cellular adaptation to hypoxia.

In anaesthesiology too, proteomic science represents the forthcoming way to recognise new and more informative biomarkers able to predict perioperative risk stratification. For example, innovative technologies have facilitated the detection of several promising early biomarkers of acute kidney injury, such as neutrophil gelatinase-associated lipocalin (NGAL) or cystatin C (CyC) (Moore et al. 2010).

In patients subjected to major abdominal surgery for abdominal aortic aneurysm (AAA), proteomic approach has revealed changes in plasma proteome 6 h after surgery and has focused the attention on increased plasma levels of thrombin, which may be one of the most important mediators responsible for the systemic inflammatory reaction and for the hemodynamic instability that typically follow AAA surgery (Modesti et al. 2009).

Patients affected by postoperative cognitive dysfunction have been proved to show differences in serum proteome compared to controls. These regulated proteins might be used as a panel of biomarkers able to diagnose POCD with sensibility and specificity (Zhang et al. 2012).

Proteomic technique has also been used to study opioid tolerance in rat models. Shui et al. found that eight proteins were significantly upregulated or downregulated in the spinal cord after morphine tolerance developed, including proteins involved in targeting and trafficking of the glutamate receptor and opioid receptors and cytoskeletal proteins (Shui et al. 2007). Previous studies have demonstrated that protein kinase C (PKC) plays a key role in the development of morphine tolerance. In spinal cords taken from morphine-tolerant (MT) rats, Song et al. identified, by two-dimensional gel electrophoresis and MALDI–TOF MS, 13 differentially expressed proteins, comparing between MT rats with and without PKC knockdown (Song et al. 2012). It is

likely that some regulated proteins may mediate PKC signalling in course of opioid tolerance and serve as potential molecular targets for prevention of the development of morphine tolerance.

Proteomics will give a deeper understanding of intra- and interindividual variability in response to drugs: preoperative exams could soon include analysis of cytochrome p450 isoforms or serum esterase enzyme profiles, in order to guide the dosing of a range of medications (Atkins and Johansson 2006). Proteomic approach must be considered complementary to the pharmacogenetic one that seeks to link differences in gene structure with pharmacologic differences in drug action, such as proteomic must be complementary to the genomic studies.

16.3 Application of Proteomics in Pain Therapy

Proteomic technologies can usefully be applied to the study of the "pathologic" pain, in particular neuropathic pain, in order to identify pain-related proteins, which may serve as diagnostic markers or drug targets.

Neuropathic pain is a common syndrome that results from disease or dysfunction in the nervous system. It can arise from a wide variety of injuries to peripheral or central nerves, including metabolic disorders, traumatic injury, inflammation and neurotoxicity. Common causes of neuropathy are diabetes, herpes zoster infections, chronic or acute trauma and neurotoxins. Furthermore, neuropathic pain occurs frequently in cancer as a direct result of peripheral nerve damage (e.g. compression by a tumour) or as a side effect of many chemotherapeutic drugs. It is characterised by spontaneous ongoing or intermittent burning pain, an exaggerated response to painful stimuli (hyperalgesia) and pain in response to normally innocuous stimuli (allodynia) that can persist long after the initial injury is resolved (Attal 2012).

Neuropathic pain reflects both peripheral and central sensitisation mechanisms, which involve transcriptional and posttranscriptional modifications in sensory nerves (Xu et al. 2012). In particular it seems to be important the biological changes that develop in the grey matter of the spinal cord, where a neuronal network, traditionally called "gate", is able to modulate nociceptive information along the nervous way that codes for pain sensitivity. However, current treatment for neuropathic pain has limited success because the mechanisms that underlie the induction and maintenance of neuropathic pain are incompletely understood.

Many studies have highlighted proteome changes in peripheral or central nervous tissues taken from different animal models of neuropathic pain (Kühlein et al. 2011; Singh and Tao 2012; Lee et al. 2003; Oki et al. 2012; Zou et al. 2012; Alzate et al. 2004; Katano et al. 2006; Lu et al. 2012; Kang et al. 2006).

In a model of L5–L6 nerve ligation, five proteins with different expression levels in the spinal cord after nerve injury were identified. Creatine kinase B, which was decreased after nerve injury, might be particularly important for the development and maintenance of neuropathic pain: in fact it reduces glutamate levels and exhibits neuroprotective properties (Lee et al. 2003).

Using proteomics, in three patients, Oki et al. (2012) investigated whether injured peripheral nerves have altered protein profiles compared with fresh cadaver's control nerves. Metallothionein (a zinc-binding protein probably involved in regeneration after CNS damage) was absent in the injured nerves although it was readily detected in control nerves.

Proteomic technologies have also been used to elucidate the mechanisms by which protein kinase C (PKC) contributes to the cellular signalling responsible for central sensitisation in neuropathic pain. In a rat model of chronic constriction injury (CCI)-induced neuropathic pain, Zou et al. found 18 differentially expressed proteins in the spinal cord of rats with and without PKC knockout. Proteins were separated with two-dimensional gel electrophoresis, and gel images were analysed with PDQuest software; then the differentially expressed proteins were identified with MALDI–TOF MS (Zou et al. 2012).

In course of neuropathic pain, proteome changes can be highlighted also in samples that can be taken in a less invasive way, like the cerebrospinal fluid: this is of crucial importance to apply proteomics to human specimens (Lu et al. 2012).

The mentioned proteomic studies, together with many others, have delivered a huge number of proteins that may be involved in the pathogenesis of neuropathic pain. These proteins, based on their physiologic function, can be subdivided into fundamental categories, such as neuronal function proteins, heat shock proteins and chaperones, antioxidants, proteins related to cell cycle, signalling proteins, proteins related to the immune system and proteins related to protein synthesis and processing (Kang et al. 2006; Sharma et al. 2006; Niederberger and Geisslinger 2008).

Some regulated proteins are related to cellular metabolism: as these proteins are ubiquitous, it is not likely that they could be used as drug targets but rather as clinical biomarkers. In this sense, it is interesting to note that increased albumin levels in the spinal cord, consequence of an altered permeability of the blood–spinal cord barrier, have been found in neuropathy models (Gordh et al. 2006).

A number of regulated proteins outlined by proteomics studies are indeed involved in apoptosis. This is not surprising if we consider that apoptosis seems to induce neuronal sensitisation and loss of inhibitory systems, whereas nerve regeneration could improve neuropathic symptoms (Finnerup and Baastrup 2012).

It is remarkable that proteome changes are different depending on the tissue in which the analysis is realised and depending on neuropathic pain model. Also in humans neuropathic pain displays different pain syndromes and a number of different causes, thus indicating that a specific nerve injury might have a specific underlying mechanism. Moreover, proteome changes in neuropathic pain differ from those observed in inflammatory pain, indicating that inflammatory pain and neuropathic pain have distinct regulatory mechanisms (Kunz et al. 2005).

Proteomic studies can be used to better understand the mechanism of action of the therapies currently applied to improve painful symptoms. In spinal cord injury models, electro-acupuncture, which can to ameliorate neuropathic pain, has been proved to modify the protein expression profile in the spinal cord (Li et al. 2010). In another experiment, electro-acupuncture has shown to alter protein phosphorylation in the spinal cord dorsal horn and to alleviate inflammatory hyperalgesia previously

induced in animal models, thus sustaining the role of post-translational protein modifications also in inflammatory sensitisation (Lee et al. 2012). Electroconvulsive shock therapy is effective for treating refractory neuropathic pain, but its mechanism of action remains unknown. Kamagata et al. demonstrated that electroconvulsive stimulation can nearly normalise proteome changes induced by neuropathic pain injuries in the brainstem of rats (Kamagata et al. 2011).

Nevertheless, when applied to the study of the pain mechanisms, proteomic technologies show some limits.

At first, not all differentially expressed proteins might have an important role in the pathogenesis of the investigated diseases: proteomic analysis provides to screen potential target proteins, but the functional role of the regulated proteins has to be subsequently confirmed by other experiments that study specifically the biological role of the protein, such as studies based on knockout models.

Moreover, not all proteins in a specimen can be detected by 2D gel electrophoresis. Protein analysis is substrate limited because no amplification methods are available as yet; thus, several regulatory proteins, which are crucial in cell biology, are in such tiny quantities that they are almost impossible to detect. Similarly, high- and low-molecular-weight proteins as well as hydrophobic membrane proteins are difficult to separate by 2D gel electrophoresis. Another limitation is that 2D gel electrophoresis cannot differentiate the cellular origin of the proteins when they are extracted by tissue specimens composed by complex and heterogeneous mix of cells. Laser microdissection of single cells might yield a new method to analyse small groups of cells from neuronal tissues (Decarlo et al. 2011).

At the same time, evolution of proteome technology will soon provide new fields of research. The application of MS to imaging, or MS imaging (MSI), allows the direct investigation of tissue sections to identify biological compounds and determine their spatial distribution (Ketola and Mauriala 2012). A suggestive application of this technology has been provided by Monroe et al. (2008). They have revealed the spatial distribution of a number of phospholipids, proteins and neuropeptides in sections of rat spinal cord. these results help to understand the complex organisation of the grey matter in the spinal cord, composed by anatomically defined areas that include motor and sensory networks that are chemically different, including the networks that regulate nociceptive sensitivity and cause central sensitisation in pathologic pain.

16.4 Application of Proteomics in ICU

In every clinical field, protein biomarkers are growing in importance for the study of disease processes: troponin, C-reactive protein, tumour necrosis factor and brain natriuretic peptide are now routinely used, while protein biomarkers for diagnostic and prognostic use in cancer are helpful because of the differential protein expression in tumour (i.e. prostate, breast and colorectal cancer).

Proteomics methods are also applied in the study of hepatic and pulmonary disease.

In the intensive care unit, as in many other clinical fields, the applications of proteomics are in the diagnostic, prognostic and "therapeutic" areas, as surrogate outcome end points.

We elaborated a literature review with a definite search strategy. The articles were searched for in the US National Library of Medicine (PubMed) comprising the period from January 1, 2000, to October 4, 2012, with only studies performed in humans being selected. The terms used were:

- ICU/critical care + proteomics + diagnosis
- ICU/critical care + proteomics + treatment/therapy
- ICU/critical care + proteomics + prognosis

The search retrieved the citations distributed according to Table 16.1. Out of the total articles retrieved, among many subjective reviews, 11 papers were focusing on critical care-based proteomic studies.

The literature search results reflect the main scientific and clinical interests of intensive care physicians, sepsis, lung failure and acute kidney injury being the most studied topics.

This search is not exhaustive; other terms should bring more results, depending on the keywords highlighted by authors and experimental data may convey very useful information. Obviously, scientific data collected in different clinical settings can be translated to critical care, but for this chapter, we prefer now to focus on targeted proteomics studies in ICU patients.

When etiologic diagnosis of sepsis is concerned, we should mention that a proteomic approach, namely, matrix-assisted laser desorption/ionisation–time-of-flight mass spectrometry (MALDI–TOF MS), allows the early identification of microorganisms and may lead to earlier appropriate treatment of bacteraemia within the first 24 h (Vlek et al. 2012). This is of extraordinary importance to start appropriate antibiotics therapy timely, which may be life-saving.

Real-time proteomic analysis of patients with sepsis may allow rapid subclassification of the syndrome into variants that may better predict responsiveness to specific therapies.

In particular, proteomics has been used to evaluate the efficacy of continuous renal replacement therapy (CRRT) employed in severe sepsis treatment when acute renal failure has already taken place (Piazza et al. 2012): ten proteins were identified as differentially expressed during CRRT (Gong et al. 2009).

This could be important for clinical purposes since acute renal failure is heterogeneous and it is likely that diagnosis and classification will not be possible using one biomarker alone.

Nevertheless, neutrophil gelatinase-associated lipocalin (NGAL), cystatin C (CyC) and liver-type fatty acid-binding protein (L-FABP) have superior sensitivity and detect acute renal failure earlier than serum creatinine – as we already cited talking of perioperative risk stratification – enhancing the ability to demonstrate benefits and to justify the implementation of therapies as CRRT (Moore et al. 2010).

Table 16.1 Review of literature (2000-2012) focusing on proteomics in critical care setting

Search strategy	PubMed results: reviews	PubMed results: original articles
ICU + proteomics + diagnosis	3	Zhou H, et al. Kidney Int. 2006; 70(10):1847–57. Exosomal Fetuin-A identified by proteomics: a novel urinary biomarker for detecting acute kidney injury
Critical care + proteomics + prognosis (ICU and prognosis = 0)	6	1: Yeager ME, et al. Plasma proteomics of differential outcome to long-term therapy in children with idiopathic pulmonary arterial hypertension. Proteomics Clin Appl. 2012 Jun;6(5–6):257–67
		2: Metzger J, et al. Urinary excretion of twenty peptides forms an early and accurate diagnostic pattern of acute kidney injury. Kidney Int. 2010 Dec;78(12):1252–62
		3: Pinto-Plata V, et al. Use of proteomic patterns of serum biomarkers in patients with chronic obstructive pulmonary disease: correlation with clinical parameters. Proc Am Thorac Soc. 2006 Aug;3(6):465–6
ICU + proteomics + therapy	4	1: Gong Y et al. Serum proteome alteration of severe sepsis in the treatment of continuous renal replacement therapy. Nephrol Dial Transplant. 2009;24(10):3108–14
		2: Zhou H et al. Exosomal Fetuin-A identified by proteomics: a novel urinary biomarker for detecting acute kidney injury. Kidney Int. 2006;70(10):1847–57
Critical care + proteomics + therapy	32	1: Yeager ME et al. Plasma proteomics of differential outcome to long-term therapy in children with idiopathic pulmonary arterial hypertension. Proteomics Clin Appl. 2012 Jun;6(5–6):257–67
		2: Ko YC et al. Proteomic analysis of CD4+ T-lymphocytes in patients with asthma between typical therapy (controlled) and no typical therapy (uncontrolled) level. Hum Exp Toxicol. 2011 Jul;30(7):541–9
Critical care + proteomics + diagnosis	38	1: Metzger J, et al. Urinary excretion of twenty peptides forms an early and accurate diagnostic pattern of acute kidney injury. Kidney Int. 2010 Dec;78(12):1252–62
		2: Ko YC, et al. Proteomic analysis of CD4 + T-lymphocytes in patients with asthma between typical therapy (controlled) and no typical therapy (uncontrolled) level. Hum Exp Toxicol. 2011 Jul;30(7):541–9
		3: Lai X, et al A proteomic workflow for discovery of serum carrier protein-bound biomarker candidates of alcohol abuse using LC-MS/MS. Electrophoresis. 2009 Jun;30(12):2207–14
		4: Gao WM, et al. A gel-based proteomic comparison of human cerebrospinal fluid between inflicted and non-inflicted pediatric traumatic brain injury. J Neurotrauma. 2007 Jan;24(1):43–53
		5: Kalenka A, et al. Changes in the serum proteome of patients with sepsis and septic shock. Anesth Analg. 2006 Dec;103(6):1522–6
		6: Pinto-Plata V et al. Use of proteomic patterns of serum biomarkers in patients with chronic obstructive pulmonary disease: correlation with clinical parameters. Proc Am Thorac Soc. 2006 Aug;3(6):465–6
		7: Schnapp LM et al. Mining the acute respiratory distress syndrome proteome: identification of the insulin-like growth factor (IGF)/IGF-binding protein-3 pathway in acute lung injury. Am J Pathol. 2006 Jul;169(1):86–95

In 2006 Kalenka et al. demonstrated the usefulness of proteomics in outcome evaluation of sepsis patients, discovering a differential protein expression between survivors and non-survivors, by using two-dimensional gel electrophoresis which detected more than 200 spots per gel (Kalenka et al. 2006).

Similarly, protein expression profiling in acute respiratory distress syndrome may yield information to aid in the diagnosis and classification of the disorder and guide mechanism-based management.

Chang et al. (2008) by using 2-dimensional gel electrophoresis (DIGE) and PCA (principal component analysis) on bronchoalveolar lavage fluid of ARDS patients demonstrated that protein composition is different in ARDS and normal subjects, so confirming previous papers, but also that it changes during the disease evolution (days 1, 3 and 7 after the onset of illness). This study individuated S100 proteins (S100A8 and S100A9) as potential candidates for targeted intervention in ARDS because these proteins interact with several key modulators of ARDS, but are not highly connected to other noninflammatory proteins in the interactome. The S100A8 and S100A9 proteins belong to a family of calcium-binding proteins that are important mediators of inflammation in sepsis (Piazza et al. 2009), and the blockade of S100 proteins may result in an effective anti-inflammatory effect by simultaneously modulating several proinflammatory pathways without perturbing other important homeostatic pathways in the lungs (Fig. 16.1).

Similarly, changes in the expression of many protein groups such as inflammatory cytokines and cytoskeletal proteins in the CSF have been described after traumatic brain injury (TBI). Differences in the pattern of proteins in two-dimensional gels were detected between TBI and control patients by Cadosh et al. (2010): 14 proteins were only present in the serum of TBI patients, while other proteins were either up- or downregulated.

Annually there are two million traumatic brain injury cases in the USA alone, 100,000 deaths, 70,000–90,000 people with long-term disabilities and 2,000 that survive in permanent vegetative state.

The importance of neuro-proteomics is that it will help elucidate the currently poorly understood biochemical mechanisms or pathways underlying brain injury so that the diagnosis and treatment can be developed (Hergenroeder et al. 2008).

16.5 Conclusions

In its current state, proteomics is a conceptually appealing endeavour to understand human physiology and disease by a global characterisation of protein function and concentration. As proteomics technologies develop, there will be an increasing emphasis on the study of proteins in real time.

Even today, anaesthetic drugs are largely administered in the absence of exact pharmacodynamics details and selected on clinical experience and continuous assessment of patient. Proteomics allows a deeper understanding of patient-specific drug response, and this explains the vast potential for laboratory and clinical applications of proteomics

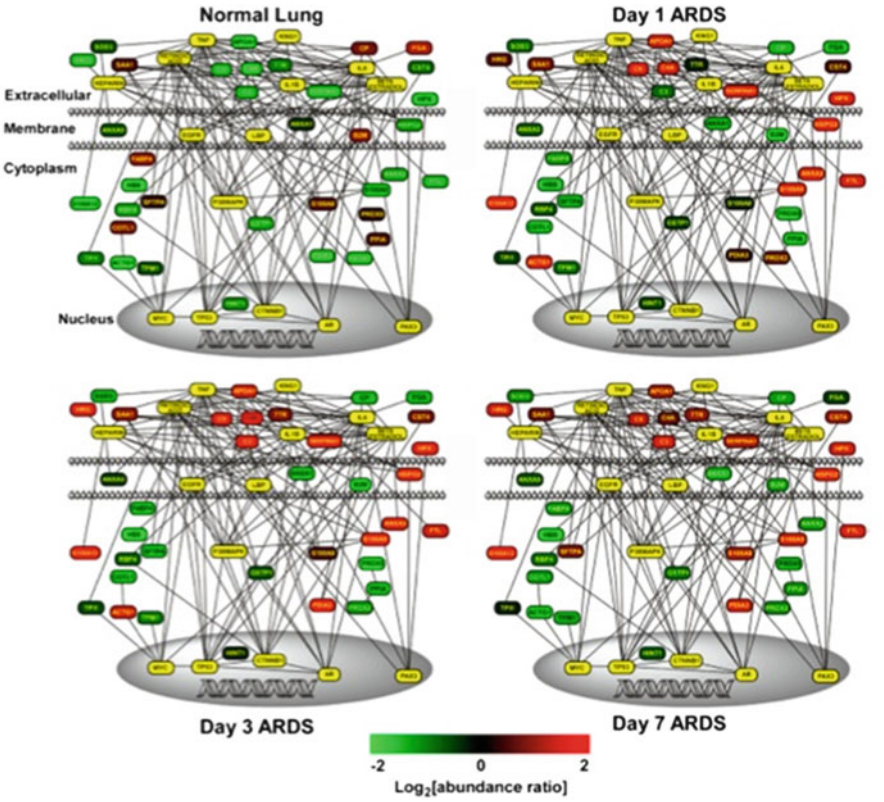


Fig. 16.1 The temporally dynamic characteristic of ARDS was revealed by the time course network analysis. Significant differences were seen in the expression of many members of the network between normal controls and day 1 of ARDS, while the differences between days 1 and 3 of ARDS were less dramatic. Nevertheless, some proteins, such as complement C3 and peroxiredoxin 2, changed in abundance between these times. These changes likely reflect changes in innate immune and oxidant pathways in the early stages of lung injury. The expression profile of the network at days 3 versus 7 of ARDS revealed several proteins that changed in expression: annexin A3 (decreased), surfactant protein-A (increased) and actin (decreased), among others. These changes likely reflect regeneration of the lung epithelium, decreased cellular injury and turnover, and the resolution of lung injury (From Chang et al., AJRCCM 2008, with permission)

technology in anaesthesia, from mechanisms of drug action to perioperative management for the reduction of cardiovascular, pulmonary, renal and neurologic injury. As critical care is concerned, numerous biological processes, including inflammation, apoptosis and thrombosis, are involved in the pathogenesis of trauma, sepsis and acute lung injury, to cite the commonest pathologies which the physicians treat in ICU.

Traditional research methods that explore single biological pathways cannot capture the complex interactions between these processes. In contrast, systems-based methodologies, such as proteomics, analyse global biological changes and provide an opportunity to examine the complexity that is inherent in critical care.

References

Alzate O, Hussain SR, Goettl VM, Tewari AK, Madiai F, Stephens Jr RL, et al. Proteomic identification of brainstem cytosolic proteins in a neuropathic pain model. Brain Res Mol Brain Res. 2004;128:193–200.

Atkins JH, Johansson JS. Technologies to shape the future: proteomics applications in anesthesiology and critical care medicine. Anesth Analg. 2006;102(4):1207–16.

Attal N. Neuropathic pain: mechanisms, therapeutic approach, and interpretation of clinical trials. Continuum (Minneap Minn). 2012;18(1):161–75.

Barnett N, Ware LB. Biomarkers in acute lung injury–marking forward progress. Crit Care Clin. 2011;27(3):661–83.

Bhattacharya AA, Curry S, Franks NP. Binding of the general anesthetics propofol and halothane to human serum albumin. High resolution crystal structures. J Biol Chem. 2000;275(49):38731–8.

Cadosch D, Thyer M, Gautschi OP, Lochnit G, Frey SP, Zellweger R, Filgueira L, Skirving AP. Functional and proteomic analysis of serum and cerebrospinal fluid derived from patients with traumatic brain injury: a pilot study. ANZ J Surg. 2010;80(7–8):542–7.

Chang DW, Hayashi S, Gharib SA, Vaisar T, King ST, Tsuchiya M, Ruzinski JT, Park DR, et al. Proteomic and computational analysis of bronchoalveolar proteins during the course of the acute respiratory distress syndrome. Am J Respir Crit Care Med. 2008;178(7):701–9.

Decarlo K, Emley A, Dadzie OE, Mahalingam M. Laser capture microdissection: methods and applications. Method Mol Biol. 2011;755:1–15.

Duan K, Ouyang W, Chen M, Xia Y, Wang S. Delayed effect of isoflurane on hippocampal proteome after anesthesia in adult and aged rats. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2009;34(7):589–94.

Eckenhoff RG, Johansson JS. Molecular interactions between inhaled anesthetics and proteins. Pharmacol Rev. 1997;49(4):343–67.

Feng J, Zhu M, Schaub MC, Gehrig P, Roschitzki B, Lucchinetti E, et al. Phosphoproteome analysis of isoflurane-protected heart mitochondria: phosphorylation of adenine nucleotide translocator-1 on Tyr194 regulates mitochondrial function. Cardiovasc Res. 2008;80(1):20–9.

Finnerup NB, Baastrup C. Spinal cord injury pain: mechanisms and management. Curr Pain Headache Rep. 2012;16(3):207–16.

Franks NP. Molecular targets underlying general anaesthesia. Br J Pharmacol. 2006;147 Suppl 1:S72–81.

Fütterer CD, Maurer MH, Schmitt A, Feldmann Jr RE, Kuschinsky W, Waschke KF. Alterations in rat brain proteins after desflurane anesthesia. Anesthesiology. 2004;100(2):302–8.

Gong Y, Chen N, Wang FQ, Wang ZH, Xu HX. Serum proteome alteration of severe sepsis in the treatment of continuous renal replacement therapy. Nephrol Dial Transplant. 2009;24(10):3108–14.

Gordh T, Chu H, Sharma HS. Spinal nerve lesion alters blood-spinal cord barrier function and activates astrocytes in the rat. Pain. 2006;124(1–2):211–21.

Hergenroeder G, Redell JB, Moore AN, Dubinsky WP, Funk RT, Crommett J, Clifton GL, Levine R, Valadka A, Dash PK. Identification of serum biomarkers in brain-injured adults: potential for predicting elevated intracranial pressure. J Neurotrauma. 2008;25(2):79–93.

Hirota K, Takabuchi S, Namba T, Fukuda K. Halothane reversibly blocks cellular hypoxic responses mediated by Hypoxia-Inducible Factor 1. Anesthesiology. 2003;99:A88.

Kalenka A, Feldmann Jr RE, Otero K, Maurer MH, Aschke KF, Fiedler F. Changes in the serum proteome of patients with sepsis and septic shock. Anesth Analg. 2006;103(6):1522–6.

Kalenka A, Hinkelbein J, Feldmann Jr RE, Kuschinsky W, Waschke KF, Maurer MH. The effects of sevoflurane anesthesia on rat brain proteins: a proteomic time-course analysis. Anesth Analg. 2007;104(5):1129–35.

Kalenka A, Gross B, Maurer MH, Thierse HJ, Feldmann Jr RE. Isoflurane anesthesia elicits protein pattern changes in rat hippocampus. J Neurosurg Anesthesiol. 2010;22(2):144–54.

Kamagata C, Tsuboko Y, Okabe T, Sato C, Sakamoto A. Proteomic analysis of rat brains in a model of neuropathic pain following exposure to electroconvulsive stimulation. Biomed Res. 2011;32(2):91–102.

Kang SK, So HH, Moon YS, Kim CH. Proteomic analysis of injured spinal cord tissue proteins using 2-DE and MALDI-TOF MS. Proteomics. 2006;6(9):2797–812.

Katano T, Mabuchi T, Okuda-Ashitaka E, Inagaki N, Kinumi T, Ito S. Proteomic identification of a novel isoform of collapsin response mediator protein-2 in spinal nerves peripheral to dorsal root ganglia. Proteomics. 2006;6(22):6085–94.

Ketola RA, Mauriala T. Mass spectrometric tools for cell and tissue studies. Eur J Pharm Sci. 2012;46(5):293–314.

Koch C, Greenfield S. How does consciousness happen? Sci Am. 2007;297(4):76–83.

Kopp Lugli A, Yost CS, Kindler CH. Anaesthetic mechanisms: update on the challenge of unravelling the mystery of anaesthesia. Eur J Anaesthesiol. 2009;26(10):807–20.

Kühlein HN, Tegeder I, Möser C, Lim HY, Häussler A, Spieth K, et al. Nerve injury evoked loss of latexin expression in spinal cord neurons contributes to the development of neuropathic pain. PLoS One. 2011;6(4):e19270.

Kunz S, Tegeder I, Coste O, Marian C, Pfenninger A, Corvey C, et al. Comparative proteomic analysis of the rat spinal cord in inflammatory and neuropathic pain models. Neurosci Lett. 2005;381(3):289–93.

Lee SC, Yoon TG, Yoo YI, Bang YJ, Kim HY, Jeoung DI, et al. Analysis of spinal cord proteome in the rats with mechanical allodynia after the spinal nerve injury. Biotechnol Lett. 2003;25(24):2071–8.

Lee SH, Kim SY, Kim JH, Jung HY, Moon JH, Bae KH, et al. Phosphoproteomic analysis of electroacupuncture analgesia in an inflammatory pain rat model. Mol Med Rep. 2012;6(1):157–62.

Li WJ, Pan SQ, Zeng YS, Su BG, Li SM, Ding Y, et al. Identification of acupuncture-specific proteins in the process of electro-acupuncture after spinal cord injury. Neurosci Res. 2010;67(4):307–16.

Lu J, Katano T, Nishimura W, Fujiwara S, Miyazaki S, Okasaki I, et al. Proteomic analysis of cerebrospinal fluid before and after intrathecal injection of steroid into patients with postherpetic pain. Proteomics. 2012. doi:10.1002/pmic.201200125.

Modesti PA, Gamberi T, Bazzini C, Borro M, Romano SM, Cambi GE, et al. Response of serum proteome in patients undergoing infrarenal aortic aneurysm repair. Anesthesiology. 2009;111(4):844–54.

Monroe EB, Annangudi SP, Hatcher NG, Gustein HB, Rubakhin SS, Sweedler JV. SIMS and MALDI MS imaging of the spinal cord. Proteomics. 2008;8(18):3746–54.

Moore E, Bellomo R, Nichol A. Biomarkers of acute kidney injury in anesthesia, intensive care and major surgery: from the bench to clinical research to clinical practice. Minerva Anestesiol. 2010;76(6):425–40.

Morgan PG, Hoppel CL, Sedensky MM. Mitochondrial defects and anesthetic sensitivity. Anesthesiology. 2002;96(5):1268–70.

Niederberger E, Geisslinger G. Proteomics in neuropathic pain research. Anesthesiology. 2008;108(2):314–23.

Oki G, Wada T, Iba K, Aiki H, Sasaki K, Imai S, Sohma H, Matsumoto K, Yamaguchi M, Fujimiya M, Yamashita T, Kokai Y. Metallothionein deficiency in the injured peripheral nerves of complex regional pain syndrome as revealed by proteomics. Pain. 2012;153(3):532–9.

Pan JZ, Xi J, Tobias JW, Eckenhoff MF, Eckenhoff RG. Halothane binding proteome in human brain cortex. J Proteome Res. 2007;6(2):582–92.

Pan JZ, Xi J, Eckenhoff MF, Eckenhoff RG. Inhaled anesthetics elicit region-specific changes in protein expression in mammalian brain. Proteomics. 2008;8(14):2983–92.

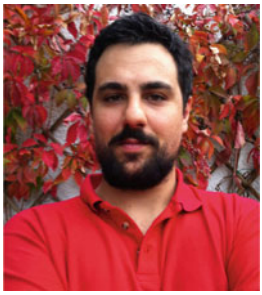
Piazza O, Cotena S, De Robertis E, Caranci F, Tufano R. Sepsis associated encephalopathy studied by MRI and cerebral spinal fluid S100B measurement. Neurochem Res. 2009;34(7):1289–92.

Piazza O, Pulcrano G, Fiori PL, Tufano R, Lonardo M, Rossano F, Catania MR. Toll-like receptor kinetics in septic shock patients: a preliminary study. Int J Immunopathol Pharmacol. 2012;25(2):425–33.

Pidikiti R, Zhang T, Mallela KM, Shamim M, Reddy KS, Johansson JS. Sevoflurane-induced structural changes in a four-alpha-helix bundle protein. Biochemistry. 2005;44(36):12128–35.
Sharma HS, Gordh T, Wiklund L, Mohanty S, Sjoquist PO. Spinal cord injury induced heat shock protein expression is reduced by an antioxidant compound H-290/51: an experimental study using light and electron microscopy in the rat. J Neural Transm. 2006;113(4):521–36.
Shui HA, Ho ST, Wang JJ, Wu CC, Lin CH, Tao YX, et al. Proteomic analysis of spinal protein expression in rats exposed to repeated intrathecal morphine injection. Proteomics. 2007;7(5):796–803.
Singh OV, Tao YX. Two-dimensional gel electrophoresis: discovering neuropathic pain-associated synaptic biomarkers in spinal cord dorsal horn. Method Mol Biol. 2012;851:47–63.
Song Z, Guo Q, Zhang J, Li M, Liu C, Zou W. Proteomic analysis of PKCγ-related proteins in the spinal cord of morphine-tolerant rats. PLoS One. 2012;7(7):e42068.
Tang J, Chen X, Tu W, Guo Y, Zhao Z, Xue Q, et al. Propofol inhibits the activation of p38 through up-regulating the expression of annexin A1 to exert its anti-inflammation effect. PLoS One. 2011;6(12):e27890.
Tsuboko Y, Sakamoto A. Propofol anaesthesia alters the cerebral proteome differently from sevoflurane anaesthesia. Biomed Res. 2011;32(1):55–65.
Vemparala S, Domene C, Klein ML. Computational studies on the interactions of inhalational anesthetics with proteins. Acc Chem Res. 2010;43(1):103–10.
Vlek AL, Bonten MJ, Boel CH. Direct matrix-assisted laser desorption ionization time-of-flight mass spectrometry improves appropriateness of antibiotic treatment of bacteremia. PLoS One. 2012;7(3):e32589.
WHO International Programme on Chemical Safety. Biomarkers in risk assessment: validity and validation. 2001. http://www.inchem.org/documents/ehc/ehc/ehc222.htm
Xi J, Liu R, Asbury GR, Eckenhoff MF, Eckenhoff RG. Inhalational anesthetic-binding proteins in rat neuronal membranes. J Biol Chem. 2004;279(19):19628–33.
Xiao YY, Chang YT, Ran K, Liu JP. Delayed preconditioning by sevoflurane elicits changes in the mitochondrial proteome in ischemia-reperfused rat hearts. Anesth Analg. 2011;113(2):224–32.
Xu B, Descalzi G, Ye HR, Zhuo M, Wang YW. Translational investigation and treatment of neuropathic pain. Mol Pain. 2012;8:15.
Zhang X, Liu Y, Feng C, Yang S, Wang Y, Wu AS, et al. Proteomic profiling of the insoluble fractions in the rat hippocampus post-propofol anesthesia. Neurosci Lett. 2009;465(2):165–70.
Zhang Q, Li SZ, Feng CS, Qu XD, Wang H, Zhang XN, et al. Serum proteomics of early postoperative cognitive dysfunction in elderly patients. Chin Med J (Engl). 2012;125(14):2455–61.
Zou W, Zhan X, Li M, Song Z, Liu C, Peng F, et al. Identification of differentially expressed proteins in the spinal cord of neuropathic pain models with PKCgamma silence by proteomic analysis. Brain Res. 2012;1440:34–46.

Ornella Piazza, MD, Italy Salerno professor at the Department of Medicine and Surgery at University. Prof. Piazza with her colleagues has started the Centre for Translational Science of the University of Salemo (CeTIS). She is editor of the open access journal *TranslationalMedicine@UniSa* (www.translationalmedicine.unisa.it). The clinical use of biomarkers of brain injury is the major research area of Prof. Piazza.

Geremia F. Zito Marinosci, MD, Italy Assistant Research Professor at the Department of Anesthesia and Intensive Care at Naples University School of Medicine, "Federico II". His field of interest involves the clinical application of biomarkers essay in the care of critically ill patients.



Giuseppe De Benedictis, MD, Italy Junior specialist physician in the Department of Anesthesia and Intensive Care Medicine, University of Naples "Federico II", Italy. He graduated with honors in the same University in March 2011.

ERRATUM

Erratum to: Chapter 5 Toward Development of Novel Peptide-Based Cancer Therapeutics: Computational Design and Experimental Evaluation

Elena Pirogova, Nahlah Makki A. Almansour, and Taghrid Istivan

DOI 10.1007/978-94-007-5811-7_17

Nahlah Makki A. Almansour (Applied Sciences, RMIT University) is the second author of Chapter 5, *Toward Development of Novel Peptide-Based Cancer Therapeutics: Computational Design and Experimental Evaluation*, but was not listed among the authors.

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_5,

Index

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7, © Springer Science+Business Media Dordrecht 2013

D

N