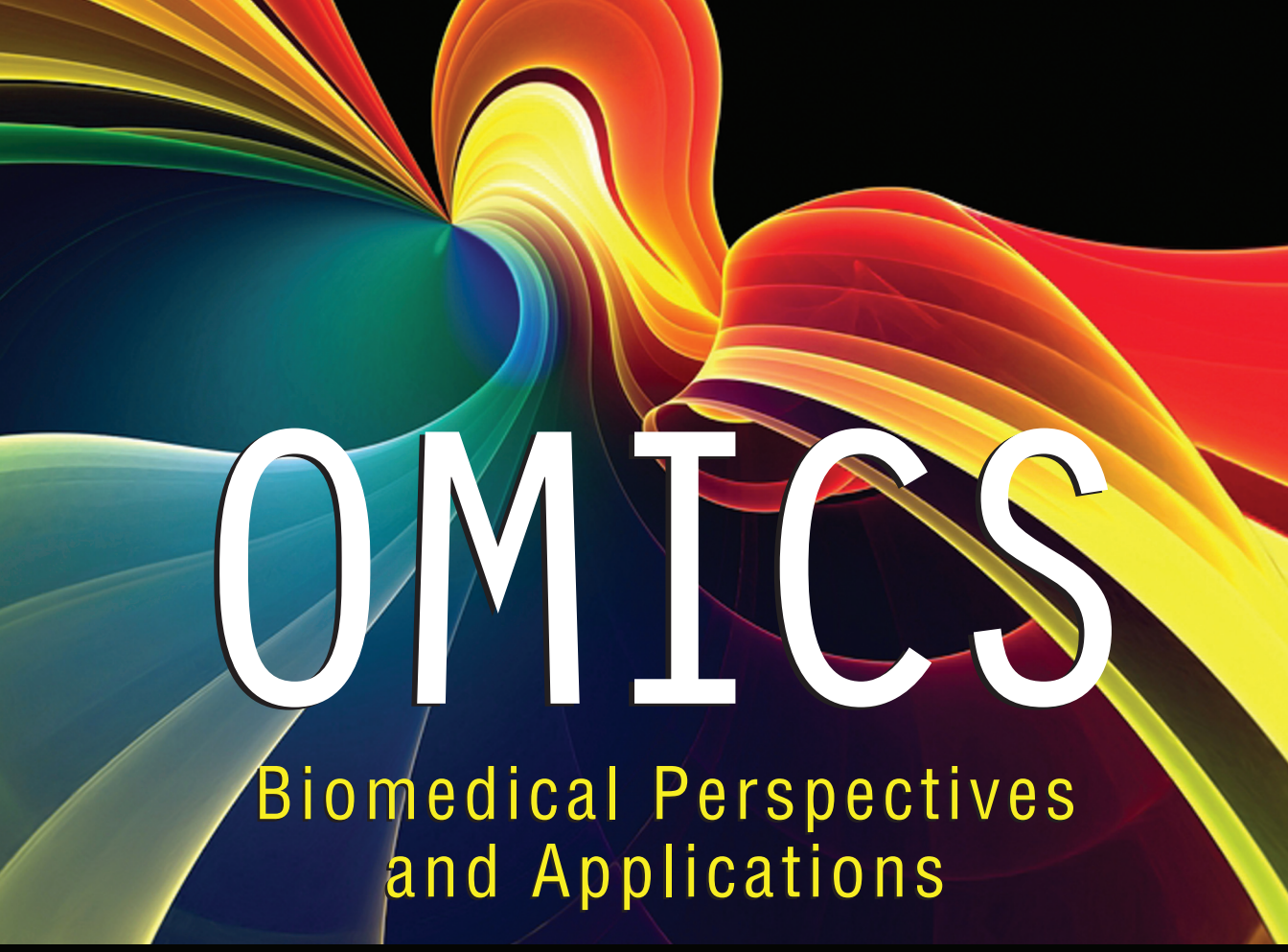

Edited by
Debmalya Barh
Kenneth Blum
Margaret Anne Madigan

CRC Press
Taylor & Francis Group

OMICS

Biomedical Perspectives and Applications

OMICS

Biomedical Perspectives and Applications

Edited by
Debmalya Barh
Kenneth Blum
Margaret Anne Madigan

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20111004

International Standard Book Number-13: 978-1-4398-5009-1 (eBook - PDF)

Purnendu Bhusan Barh (February 22, 1940–February 27, 2008)

This book is dedicated to Purnendu Bhusan Barh (S/O Ambika Bhusan Barh), an eminent academician, philosopher, career master, and transformator who is the soul and inspiration behind the establishment of the Institute of Integrative Omics and Applied Biotechnology (IIOAB) and all its activities.

Contents

Foreword

The term *omics* refers to a field of study in biological sciences ending in -omics, such as genomics, proteomics, or metabolomics. Basically, genomics is the discipline in genetics concerning the study of the genomes of organisms and fine-scale genetic mapping efforts. Proteomics is a well-accepted term for studying proteins on a large scale, and metabolomics is a term denoting investigation of the chemical fingerprints of small molecules' metabolite profiles.

Overall, one of the major challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic study of gene expression at the RNA level and metabolomic information to give a more complete picture of living organisms. Bioinformatics, another key term coined by Paulien Hogeweg and Ben Hesper in 1978, is an integrated term demonstrating the application of statistics and computer science to the field of molecular biology. The primary objective of bioinformatics is to increase and enhance the understanding of both biological and biochemical processes.

The list of authors and topics covered in this book is impressive. The editors are to be congratulated for bringing together such a unique group of experts from various fields of cutting-edge omics research. The book has twenty-seven chapters that deal with several cutting-edge features of novel technology.

The book starts with a chapter entitled "Overview of Omics" by Dr. Raghavachari that provides an overview of omics and omic technologies such as cellomics, glycomics, and lipidomics. The second chapter by Drs. Singh and Somvanshi focuses on bioinformatics and demonstrates how this can be an essential tool in omics. The third chapter provides a new twist and demonstrates the association of omics technology with nutrigenomics and nutraceuticals. Drs. López-Corrales, Stutzman, Miyoshi, Barh, and Azevedo discuss the various approaches of omics technology in toxicology research and applications in biomedical sciences in the fourth chapter. The fifth chapter covers the basic and versatile therapeutic applications of stem cells; it was written by Drs. Arya and Tripathi. Dr. Sandhiya provides an excellent chapter on the emerging trends of nanotechnology in omics-based drug discovery and development. This chapter provides a vivid description on how the integration of nanotechnology in the drug delivery system has the potential to improve specific drug targeting, drug release and interaction, and enhanced efficacy. The seventh chapter, by Drs. Zhang and Olin, discusses the biomedical applications of magnetic nanoparticles. Dr. Zaki demonstrates the usefulness of these state-of-the-art technologies on high-throughput screening in medical diagnosis and prognosis in Chapter 8. Drs. Kolukisaoglu and Thurow emphasize the applications of high-throughput omics technology in systems biology. Drs. Visaria, Prakash, and Shrivastava extensively discuss the safety aspects in diagnostic imaging techniques used in omics in Chapter 10. Dr. Gope and collaborators demonstrate the molecular genetics of human cancers in Chapter 11. Dr. Chatterjee highlights the intricate aspects on the functional identification of unknown genes in Chapter 12. Dr. Carranza-Cereceda and collaborators discuss their interesting research findings on the proteomics of phagosomal pathogens. In Chapter 14, Drs. Selvarajoo and Tsuchiya explore the governing principles of cellular networks from the perspective of systems biology. Dr. Fukunishi demonstrates the salient features of intermolecular interaction in biological systems. In Chapter 16, Dr. Zheng and collaborators demonstrate the application of neuromics and highlight how implanted brain machines interface in rats. Drs. Sharma and Munshi exhibit their concept on pharmacogenomics in the development of disease specific therapeutic strategy. Drs. Dhawan and Padh discuss the aspects of omics approaches in cancer drug discovery in Chapter 18. Drs. Ohdaira and Yoshida highlight the use of microRNA expression in the therapeutic strategy for tumors. Dr. Pereira and his collaborators extensively discuss marine metabolomics in

cancer chemotherapy. Drs. Hong, Xu, Mendrick, and Tong highlight the important findings and the present status of type 2 diabetes. Dr. Viero highlights the applications of genomics and proteomics in cardiac therapies. Drs. Davies and Flower demonstrate the applications of omics in the treatment of infectious diseases in Chapter 23. Dr. Verma and collaborators highlight their findings on AIDS and HIV with omics technologies. Dr. Archer highlights aspects of epigenetics in neuropsychiatry in Chapter 25, and in Chapter 26, Dr. Blum reviews the neurogenetics and nutrigenomics of reward deficiency syndrome. Finally, Dr. Barh et al. summarize these intricate aspects and issues together and project the future pathology.

Overall, this book will be intensively useful to scientists from both academia and industry, teachers and professors, health professionals, and mostly students, who should be encouraged to study and learn from its wisdom.

Debasis Bagchi, PhD, MACN, CNS, MAIChE
Department of Pharmacological and Pharmaceutical Sciences
University of Houston College of Pharmacy
Houston, Texas

REFERENCES

Bagchi, D., Bagchi, M., and Lau, F. C. (Eds.) (2010). *Genomics, proteomics and metabolomics in nutraceuticals and functional foods*. New York: Wiley Blackwell.

Baxevanis, A. D., Petsko, G. A., Stein, L. D., and Stormo, G. D. (Eds.) (2007). *Current protocols in bioinformatics*. New York: Wiley Blackwell.

Bagchi, D., Lau, F. C., and Ghosh, D. (Eds.) (2010). *Biotechnology in functional foods and nutraceuticals.* Boca Raton, Florida: Taylor & Francis/CRC Press.

Hartung, T. (2009). Toxicology for the 21st century. *Nature* 460, 208–212.

Preface

This book, *Omics: Biomedical Perspectives and Applications*, illustrates the direction that this rapidly emerging discipline is taking. Applications of omics technologies in the postgenomics era have swiftly expanded from rare monogenic disorders to multifactorial common complex diseases, pharmacogenomics, and personalized medicine.

Omics informally refers to a field of study in biology ending in -omics, such as genomics and proteomics. The related suffix -ome is used to address the objects of such explosive fields of study as the genome and protome, respectively. The field combines different omics techniques such as transcriptomics and proteonomics. The suffix -ome as used in molecular biology refers to a totality or systems biology. The -ome suffix originated as a variant of -oma and became productive in the last quarter of the 19th century. The *Oxford English Dictionary* suggests that the third definition originated as a backformation from *mitome*, which was later also reinforced by *chromosome*. Early attestations include *biome*, first used in 1916, and *genome*, first coined as the German *Genom* in 1920. Because *genome* refers to the *complete* genetic makeup of an organism, the new suffix *-ome* suggested itself as referring to *wholeness* or *completion*.

Interestingly, bioinformaticians and molecular biologists are considered the first scientists to start to apply the -ome suffix widely. Some early advocates were bioinformaticians in Cambridge, United Kingdom, where there were many early bioinformatics labs such as the Sanger Center and European Bioinformatics Institute. One such center run by the Medical Research Council is where the first genome and proteome projects were carried out. Many -omes beyond the original *genome* have become useful and have widely adopted by research scientists. *Proteomics* has become well established as a term for studying proteins at a large scale. *Omics* can provide an easy handle to encapsulate a field; for example, an interactomics study is clearly recognizable as relating to large-scale analysis of gene-gene, protein-protein, or protein-ligand interactions. Researchers have been rapidly taking up omes and omics, as shown by the explosion of the use of these terms in PubMed since the mid-1990s, making this exciting field relatively new.

Omics research now encompasses an assortment of technologies and academic disciplines aspiring to analyze the mysteries involved in cellular function at a molecular level within organisms. Genomics, transcriptomics, pharmacogenomics, toxicogenomics, epigenomics, lipidomics, glycomics, immunomics, and proteomics are all addressed in this book, whereas the technologies covered include bioinformatics, high-throughput sequencing involving DNA and protein microarrays and mass spectrometry, stem cell research, nanoparticle drug design, the uses of magnetic nanoparticles, and diagnostic imaging.

The study of omics has become increasingly important as a specialty area within medical genetics and systems biology. This domain, originally restricted to a few researchers, has now become a vast uncharted arena where scientists from very diverse fields, including biology, biochemistry, pharmacology, pathology, toxicology, botany, neurology, psychiatry, medical and population genetics, anthropology, molecular biology, and even to some degree medical ethics converge to explore biological systems.

The increased interest stems principally from advances in molecular genetic techniques, bioinformatics, the genome project, neurosciences, nutrition science, mathematics, particle physics, and other related disciplines. Many of the dedicated scientists in this emerging field have been encouraged by enhanced public awareness of the role of genes in somatic diseases like cancer, diabetes, and HIV and complex mental diseases like bipolar depression, schizophrenia,

Alzheimer's disease, reward deficiency syndrome, and addictive, impulsive, and compulsive behaviors. The announcement of genes associated with such devastating genetically based single-gene disorders such as Huntington's disease, cystic fibrosis, and muscular dystrophy, as well as complex polygenic diseases, such as lung cancer, breast cancer, diabetes, and most recently aging, have profoundly aroused the interest of professors, students, and people all over the globe.

This book serves as an important resource and review especially to students and researchers interested in the field of *integrative omics*. The volume is also addressed to basic scientists, clinicians, and other professionals who have a specialized or even a peripheral interest in not only molecular genetics and proteomics but the field of systems biology.

In a review volume of this size, it is not possible to convey every aspect of the subject; however, we as editors have attempted to compile an outline that is comprehensive and that could serve as a state-of-the-art framework for a rather new discipline. Every effort has been made to provide an informative, basic text that presents as wide a view as possible of the current status of integrative omics.

The omics overview provides an organizational framework upon which the "Methodology and Application" section is founded. This section includes works that introduce many of the omics fields and provide background technical information and expertise. These areas include: bioinformatics, nutrigenomics, toxicology, stem cell research, magnetic nanoparticles, high-throughput screening, and safety in diagnostic imaging.

The second section, "Empirical Research," includes omics research into such diverse areas as a "Forward Genetics Approach in Genomics: Functional Identification of Unknown Genes" and "Proteomics of Phagosomal Pathogens: Lessons from *Listeria monocytogens* and New Tools in Immunology."

The third section, "Computational and Systems Biology," provides very timely topics, such as "In a Quest to Uncover Governing Principles of Cellular Networks: A Systems Biology Perspective" and "Intermolecular Interaction in Biological Systems." This section also includes an interesting topic: "Implanted Brain Machine Interfaces in Rats: A Modern Application of Neuromics."

The fourth section focuses on the application of specific omics technologies to the discovery of omics-based diagnostic and therapeutic modalities for disease treatment. These processes' challenges and successes are described in chapters that look at, for example, metabolomic research into the development of chemotherapy and the application of larger-scale high-density genome-wide association studies in type 2 diabetes to shed light on the genetic etiology and explain the difficulties involved in replicating for biomarkers. This section also covers omics-based diagnosis and treatment approaches in cardiovascular disease and cancer.

Integrative applications in various omics fields have been responsible for moving the work forward. The recent exponential growth in omics is based on the explosion of bioinformatics and other biotechnologies and the integrative multi-omics approaches being applied to research. The fifth section, "Future Perspective," deals with these issues.

The original idea for this compendium came from Dr. Debmalya Barh, who convinced CRC Press to engage all of us to edit and publish the first text in this subject area. It is our wish that the contents of this compendium will be of use to researchers and students of biology, including technologists and scientists from all disciplines, by providing both a basic platform of methods and applications and a resource for enhanced cross-pollination in a multiomics approach to future endeavors in the fertile fields of omics research. We hope that from within these chapters, these estimable researchers will impart their great appreciation of the general principles of rigorous and arduous research that can lead to appropriate and productive approaches in the study of systems biology, leading to clinical strategies and potential disease cures.

The book is an initiative of the Institute of Integrative Omics and Applied Biotechnology towards fulfillment of the mission of promoting higher education. The book is dedicated to Purnendu Bhusan Barh, Dr. Barh's beloved father.

Debmalya Barh, Kenneth Blum, and Margaret A. Madigan, Editors

Editors

Debmalya Barh, MSc, MTech, MPhil, PhD, PGDM, is a scientist, consultant biotechnologist, and intellectual property rights specialist. He is the founder and president of Institute of Integrative Omics and Applied Biotechnology, India—a global platform for multideciplinary research and advocacy. He is a pioneer researcher in male breast cancer and cardiac myxoma disease signaling pathways and drug targets. His expertise includes bioinformatics, phamacogenomics, and integrative omics-based biomarker and targeted drug discovery. Since 2008, he has authored more than 30 international peer-reviewed publications with first authorship, contributed more than 10 book chapters, and written 2 books. During this period, he has also edited three books in the area of omics and biotechnology. Because of his significant contribution towards science and higher education, he was selected by *Who's Who in the World in 2010* (p. 168). He is one of the founding members and executive editors of *The IIOAB Journal* and *IIOAB Letters* and also serves as an editorial and review board member for several professional international research journals of global repute.

Kenneth Blum, PhD, is currently Chairman of the Board and Chief Scientific Officer of LifeGen, Inc., San Diego, California, and a managing partner of Reward Deficiency Solutions, LLC, San Diego, California. He serves as a consultant and senior scientific advisor for many companies and a foundation.

Dr. Blum was for 23 years a full professor of pharmacology at the University of Texas, San Antonio, Texas. Following his service as research professor at the Wake Forest College of Medicine, Winston-Salem, North Carolina, he is currently a full professor of the Department of Psychiatry and McKnight Brain Institute University of Florida College of Medicine, Gainesville, Florida. He has received numerous awards, including the NIDA Career Teacher Award, the American Chemical Society Speakers Award, the Gordon Conference Research Award, the Presidential Excellence Award (National Council of Alcoholism and Drug Abuse), and the 2011 Lifetime Achievement Award (National Association of Holistic Addiction Studies). Dr. Blum has authored and edited 12 books, published over 400 peer-reviewed papers, and coined the terms *brain reward cascade* and *reward deficiency syndrome*. He is credited with codiscovering the first gene to associate not only with alcoholism but reward dependence in general and was the lead author in the first association study of the dopamine D2 receptor gene with severe alcoholism (*Journal of the American Medical Association*, 1990). He is considered by many to be the father of psychiatric genetics. He serves on nine editorial boards and is the associate editor on two boards, including coeditor-in-chief of the *BMC IIOAB Journal*, the official journal of the Institute of Integrative Omics and Applied Biotechnology. He is also executive editor of *Journal of Genetic Syndromes and Gene Therapy* and is an ad hoc reviewer for 40 journals worldwide. He is also the inventor of neuroadaptagen amino-acid therapy for the recovery field.

Dr. Blum's research has been covered by major newspapers all over the world, and he has made numerous television and radio appearances. In 1984, his textbook *Handbook on Abusable Drugs* was a book of the month selection. His books *Alcohol and The Addictive Brain* (with James Payne) and *Overload* (with David Miller) have received high ratings from Amazon. His work has been cited by Allen King and the Australian Broadcasting Company and on *The NY Science Show* and *Law & Order*. He has appeared on the *TODAY*, *Good Morning America*, and *Sonja Live*, to name a few. His work has received both silver and bronze medals from the Natural Products Association in 2006 and 2007. Dr. Blum has chaired three Gordon Research Conferences on alcohol and psychiatric genetics. Dr. Blum has published in almost every major scientific journal worldwide: *Science*, *Lancet*, *Nature*, *Proceedings of the National Academy of Sciences*, and *Journal of the American Medical*

Association, among others. He is actively investigating the role of natural dopamine agonists as an anticraving DNA-directed therapeutic target for prevention of relapse.

Margaret A. Madigan, BSN, is a nursing practitioner by training and is currently a senior editor and assistant to the Chairman of the Board of the LifeGen, Inc. research center located in San Diego, California. She is a native of Sydney, Australia, and has been a long-time resident of Honolulu, Hawaii. She is a graduate of the University of Sydney New South Wales having fulfilled requirements earning her a Bachelor of Health Sciences (Nursing). She served as a registered nurse at Sutherland Hospital in Sydney working on an ICU step-down ward. Ms. Madigan is a member of The New South Wales College of Nursing. She has served as a registered nurse at the Kapi'olani Medical Center, Honolulu, Hawaii, and Palomar Pomerado Health Care System Hospitals in San Diego, California. She has been certified in numerous nursing specialties, including infection control, palliative care, advanced nursing interventions, basic critical care, oncology, cancer chemotherapy, pain management, advanced fetal monitoring, neonatal resuscitation, and basic EKG monitoring, among other disciplines. Ms. Madigan is licensed in Hawaii, California, Texas, and New South Wales, Australia. She has also completed courses in brain repair for addictive disorders and has experience in psychiatric nursing. Ms. Madigan has published in the fields of neuropharmacology, neurogenetics, nutrigenomics, clinical neurology, neuroimaging, and psychiatric genetics in peer-reviewed journals. She is a graphic artist and photographer and is credited with the cover art for an issue of *The IIOAB Journal*.

Contributors

Carmen Alvarez-Dominguez, MS, PhD
Servicio de Inmunologia, Fundación Marqués de Valdecilla-IFIMAV
Hospital Santa Cruz de Liencres
Liencres, Spain

Paula B. Andrade, PhD
EQUIMTE/Laboratório de Farmacognosia
Departamento de Química
Faculdade de Farmácia
University of Porto
Porto, Portugal

Trevor Archer, PhD
Department of Psychology
University of Gothenburg
Gothenburg, Sweden

Awadhesh Kumar Arya, PhD
Department of Medicine
Banaras Hindu University
Varanasi, India

Vasco Azevedo, DVM, MSc, PhD, FESC
Departamento de Biologia Geral
ICB/UFMG
Belo Horizonte, Brazil

Debmalya Barh, MSc, MTech, MPhil, PhD, PGDM
Institute of Integrative Omics and Applied Biotechnology
Nonakuri, India

Kenneth Blum, PhD
Department of Psychiatry
University of Florida College of Medicine and McKnight Brain Institute
Gainesville, Florida

Judith Boateng, PhD
Department of Food and Animal Sciences
Alabama Agricultural and Mechanical University
Normal, Alabama

Carlos Carranza-Cereceda, BS
Servicio de Inmunologia, Fundación Marqués de Valdecilla-IFIMAV
Hospital Santa Cruz de Liencres
Liencres, Spain

Bishwanath Chatterjee, PhD
Laboratory of Developmental Biology
National Heart, Lung, and Blood Institute
National Institutes of Health
Bethesda, Maryland

Weidong Chen, PhD
Qiushi Academy for Advanced Studies
College of Computer Science
Zhejiang University
Hangzhou, China

Georgina Correia-da-Silva, PhD
Laboratório de Bioquímica
Departamento de Ciências Biológicas
Faculdade de Farmácia
University of Porto
Porto, Portugal
and
IBMC–Instituto de Biologica Molecular e Celular
University of Porto
Porto, Portugal

Jianhua Dai, PhD
College of Computer Science
Zhejiang University
Hangzhou, China

Matthew N. Davies, PhD
Social, Genetic, and Developmental Psychiatry Centre
Institute of Psychiatry
London, United Kingdom

Dipali Dhawan, PhD
Department of Cellular and Molecular Biology
B. V. Patel Pharmaceutical Education and Research Development Centre
Ahmedabad, India

Premendra Dhar Dwivedi, MSc, PhD
Food Toxicology Division
Indian Institute of Toxicology Research
Lucknow, India

Lorena Fernandez-Prieto, PhD
Servicio de Inmunologia
Fundación Marqués de Valdecilla-IFIMAV
Hospital Santa Cruz de Liencres
Liencres, Spain

Darren R. Flower, PhD
Aston Pharmacy School
School of Life and Health Sciences
Aston University
Birmingham, United Kingdom

Yoshifumi Fukunishi, PhD
Biomedicinal Information Research Center
National Institute of Advanced Industrial Science and Technology
Tokyo, Japan

Mohan L. Gope, PhD
Department of Biotechnology
Presidency College
Bangalore, India

Rajalakshmi Gope, PhD
Department of Human Genetics
National Institute of Mental Health and Neurosciences
Bangalore, India

Huixiao Hong, PhD
Division of Systems Biology
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas

Xiaoling Hu, PhD
Department of Health Technology and Informatics
The Hong Kong Polytechnic University
Hung Hom, China

Üner Kolukisaoglu, PhD
Center for Life Science Automation
University of Rostock
Rostock, Germany

Nestor Luis López-Corrales, DVM, MSc, PhD
Departamento de Biologia Geral
ICB/UFMG
Belo Horizonte, Brazil

Margaret A. Madigan, BSN
LifeGen, Inc.
San Diego, California

Priyadarshini Mallick, MSc
Amity Institute of Biotechnology
Amity University
Noida, India

Donna L. Mendrick, PhD
Division of Systems Biology
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas

Rohan Mitra, MSc
Department of Human Genetics
National Institute of Mental Health and Neurosciences
Bangalore, India

Anderson Miyoshi, MSc, PhD
Laboratório de Genética Celular e Molecular
Instituto de Ciĕncias Biológicas
Universidade Federal de Minas Gerais (ICB/UFMG)
Belo Horizonte, Brazil

Anjana Munshi, MPhil, PhD
King Saud University
Riyadh, Saudi Arabia

Hiroaki Ohdaira, MS
Department of Life Sciences
Meiji University
Kawasaki, Japan

Håkan Olin, PhD
Department of Natural Sciences, Engineering, and Mathematics
Mid Sweden University
Sundsvall, Sweden

Harish Padh, PhD
B. V. Patel Pharmaceutical Education and Research Development Centre
Ahmedabad, India

David M. Pereira, PhD
REQUIMTE/Laboratório de Farmacognosia
Departamento de Química
Faculdade de Farmácia
University of Porto
Porto, Portugal

Surya Prakash, MD
Tej Bahadur Saproo Hospital
Allahabad, India

Nalini Raghavachari, PhD
Genetics and Development Biology Center
The National Institutes of Health
Bethesda, Maryland

José Ramós-Vivas, PhD
Servicio de Inmunologia, Fundación Marqués de Valdecilla-IFIMAV
Hospital Santa Cruz de Liencres
Liencres, Spain

Estela Rodriguez-Del Rio, BS
Servicio de Inmunologia, Fundación Marqués de Valdecilla-IFIMAV
Hospital Santa Cruz de Liencres
Liencres, Spain

Sandhiya Selvarajan, MD
Division of Clinical Pharmacology
Jawaharlal Institute of Post-Graduate Medical Education and Research
Pondicherry, India

Kumar Selvarajoo, PhD
Institute for Advanced Biosciences
Keio University
Tsuruoka, Japan

Sanjeev Sharma, MD
Apollo Hospital
Hyderabad, India

Devashish Shrivastava, PhD
Department of Radiology
University of Minnesota
Minneapolis, Minnesota

Abhishek Angelo Singh, MTech
Department of Bioinformatics-BiGCaT
Maastricht University
Maastricht, the Netherlands

Anchal Singh, MSc, PhD
Department of Microbiology and Immunology
Kirksville College of Osteopathic Medicine
A.T. Still University of Health Sciences
Kirksville, Missouri

Udai Pratap Singh, BTech, MTech
Amity Institute of Biotechnology
Amity University
Noida, India

Pallavi Somvanshi, PhD
Bioinformatics Centre
Lucknow, India

Tina Stutzman, BS
Department of Biological Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts

Natércia Teixeira, PhD
Laboratório de Bioquímica
Departamento de Ciências Biológicas
Faculdade de Farmácia
University of Porto
Porto, Portugal
and
IBMC–Instituto de Biologica Molecular e Celular
University of Porto
Porto, Portugal

Kerstin Thurow, PhD
Center for Life Science Automation
University of Rostock
Rostock, Germany

Weida Tong, PhD
Division of Systems Biology
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas

Kamlakar Tripathi, MD, DM
Department of Medicine
Institute of Medical Sciences
Banaras Hindu University
Varanasi, India

Masa Tsuchiya, PhD
Institute for Advanced Biosciences
Keio University
Tsuruoka, Japan

Patrícia Valentão, PhD
REQUIMTE/Laboratório de Farmacognosia
Departamento de Química
Faculdade de Farmácia
University of Porto
Porto, Portugal

J. Thomas Vaughan, PhD
Department of Radiology
University of Minnesota
Minneapolis, Minnesota

Ashish Swarup Verma, MSc, PhD
Amity Institute of Biotechnology
Amity University
Noida, India

Martha Verghese, PhD
Department of Food and Animal Sciences
Alabama Agricultural and Mechanical University
Normal, Alabama

Cedric Viero, PhD
Department of Cardiology
Wales Heart Research Institute
School of Medicine
Cardiff University
Cardiff, United Kingdom

Rachana Visaria, PhD
Division of Testing Services
MR Safe Devices LLC
Burnsville, Minnesota

Lei Xu, PhD
Division of Systems Biology
National Center for Toxicological Research
U.S. Food and Drug Administration
Jefferson, Arkansas

Kenichi Yoshida, DVM, PhD
Department of Life Sciences
Meiji University
Kawasaki, Japan

Maysaa El Sayed Zaki, PhD, MD
Faculty of Medicine
Mansoura University
Mansoura City, Egypt

Renyun Zhang, PhD
Department of Natural Sciences, Engineering and Mathematics
Mid Sweden University
Holmsgatan, Sweden

Shaomin Zhang, PhD
Qiushi Academy for Advanced Studies
College of Biomedical Engineering and Instrument Science
Zhejiang University
Hangzhou, China

Ting Zhao, PhD
Qiushi Academy for Advanced Studies
Zhejiang University
Hangzhou, China

Xiaoxiang Zheng, MD
Qiushi Academy for Advanced Studies
College of Biomedical Engineering and Instrument Science
Zhejiang University
Hangzhou, China

1 Overview of Omics

Nalini Raghavachari
National Heart, Lung, and Blood Institute/National Institutes of Health
Bethesda, Maryland

CONTENTS

1.1 INTRODUCTION TO THE FIELD OF OMICS

The central dogma of molecular biology (Figure 1.1a), enunciated by Crick (Crick, 1954), specified that the instruction manual is DNA (encoding genes) and that genes were transcribed into RNA to ultimately produce the basic operational elements of cellular biology, proteins whose interactions, through many levels of complexity, result in functioning living cells. This was the first description of the action of genes. After an enormous experimental effort spanning the last half-century, made possible by the development of many assays and technological advances in computing, sensing, and imaging, it has become apparent that the basic instruction manual and its processing are vastly more sophisticated than what was imagined in the 1950s. With the advent of these novel technologies, the primary focus of modern biology has shifted to link genotype to phenotype, interpreted broadly, from the level of the cellular environment to links with development and disease, and the central dogma has now been viewed as an integration of the -ome studies as depicted in Figure 1.1b.

In this context, biomedical research has been transformed recently by an exponential increase in the ability to measure biological variables of interest in grand scale (Abraham, Taylor et al., 2004). Diverse methods of large-scale measurements of biological processes have emerged in the past 15 years, and the list is growing rapidly. Remarkable technologies such as microarrays and their descendants, high-throughput sequencing, *in vivo* imaging techniques, and many others have enabled biologists to begin to analyze function at molecular and higher scales. The various aspects of these analyses have coalesced as *omics* (Wild, 2010). Omics is an emerging and exciting area in the field of science and medicine. Technologies that measure some characteristics of a large family of cellular molecules, such as genes, proteins, or small metabolites, have been named by appending

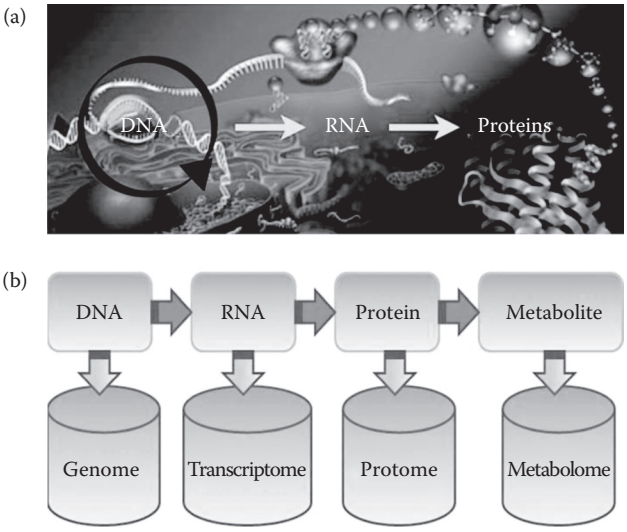

FIGURE 1.1 (**See color insert.**) (a) The central dogma (past and present) as explained by Crick. (b) Integration of the -ome studies.

the suffix *-omics*, as in *genomics*. Omics refers to the collective technologies used to explore the roles, relationships, and actions of the various types of molecules that make up the cells of an organism. These technologies encompass the following four major fields of study:

1. Genomics: the study of genome that stores the information in a cell to predict what can happen.
2. Transcriptomics: the study of mRNA or transcript that would depict what is really happening in a cell.
3. Proteomics: the study of protein molecules that would illustrate the functional roles of molecules in cellular function.
4. Metabolomics: the study of molecules involved in cellular metabolism that would eventually depict the phenotype of an organism.

Numerous promising developments have been elucidated using genomics, transcriptomics, epigenomics, proteomics, metabolomics, interactomics, cellomics, and bioinformatics (Wild, 2010). The omics technology that has driven these new areas of research consists of DNA and protein microarrays, mass spectrometry, and a number of other instruments that enable high-throughput analyses (Bier, von Nickisch-Rosenegk et al., 2008). Likewise, the field of bioinformatics has grown in parallel and with the help of the internet, rapid data analysis and information exchange are now possible. With these advancements, not only will omics have an impact on our understanding of biological processes, but the prospect of more accurately diagnosing and treating disease will soon become a reality. In an effort to understand the complex interplay of genes/proteins in disease processes, comparative genetic, transcriptomic, proteomic, and metabolomic analyses for individuals and populations are required. In particular, systems biology, more than the simple merge of omics technologies, aims to understand the biological behavior of cellular systems and enhance the capacity to test the probability of disease as early as possible through a noninvasive method of diagnosis as illustrated in Figure 1.2. The omics technologies are believed to open a new road to the field of personalized medicine in this postgenomic era. Understanding the existing and emerging technologies of genomics, transcriptomics, proteomics, and metabolomics is critical for widespread application of these technologies in the field of medicine.

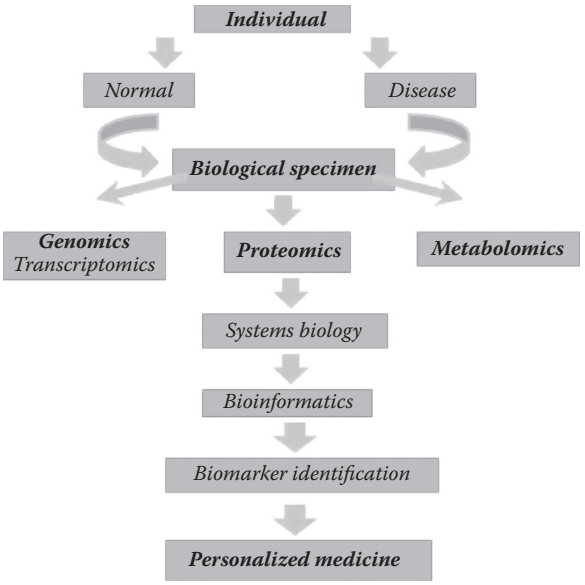

FIGURE 1.2 Flow chart for omics-based biomedical research.

Although the field of omics is ever expanding, currently genotyping, gene expression profiling, epigenomics, proteomics, and metabolomics are well established and widely used by scientists in clinical research. Technologies such as cellomics, glycomics, and lipidomics are now emerging as powerful tools for medical research. Almost invariably, these advances in omics have been associated with major expectations of transforming not only biological knowledge but also medicine and health. This chapter will provide valuable information about these powerful omics technologies.

1.2 GENOMICS

Genomics may be described as the comprehensive analysis of DNA structure and function and broadly refers to the analysis of all the genes and transcripts included within the genome (Bier, von Nickisch-Rosenegk et al., 2008). Understanding biological diversity at the whole-genome level will yield insight into the origins of individual traits and disease susceptibility. The aim of genomics is to analyze or compare the entire genetic complement of a species. Important areas of genomics are:

1. Structural genomics for the analysis of macromolecular structure using computational tools and theoretical frameworks
2. Comparative genomics (genomics for study of a species) by comparisons with model organisms
3. Functional genomics, a field of genomics attempting to make use of the vast wealth of data produced by genome sequencing projects to describe genome function
4. Pharmacogenomics, which aims to study how genes influence the response of humans to drugs, from the population to the molecular level, and uses genomic approaches and technologies for the identification of drug targets

These major fields of genomics are subclassified into genotyping, transcriptomics, pharmacogenomics, toxicogenomics, and epigenomics.

1.2.1 Genotyping

Although organisms such as humans are quite similar at the genetic level, differences exist at a frequency of about one in every 1000 nucleotide bases (Barron, 2008). This translates into approximately three million base differences between each individual. Such changes are referred to as single-nucleotide polymorphisms (SNPs), and a significant effort collectively referred to as genotyping is now underway in the research community to map the individual SNPs in humans and other organisms. SNPs may be found within gene coding regions or in noncoding regions. Their effects may be subtle, yielding slight changes in protein function, or profound, leading to the development of disease. A polymorphism is distinct from a mutation, in that mutation is considered rare, affecting less than 1% of the species, whereas polymorphism is relatively common, and its prevalence is no different from what is considered normal (Barron, 2008). Over the past decade, there has been an unprecedented surge of data directed at sequencing and categorizing all of the genes in the human genome, as well as those of other organisms. There has also been a concomitant acceleration in the technology dedicated to genomics research, including instrumentation, reagents, software, and databases. Since the introduction of array-based genotyping techniques, it has become possible to cover with varying resolution the entire genome in what are now commonly referred to as genome-wide association studies (GWAS). The GWAS have uncovered and will uncover in the future interesting and previously unknown polymorphic variants that are associated with a variety of chronic diseases (Seshadri, Fitzpatrick et al., 2010).

1.2.2 Transcriptomics

The abundance of specific mRNA transcripts in a biological specimen is a reflection of the magnitude of the expression levels of the corresponding genes. Gene expression profiling is the identification and characterization of the mixture of mRNA that is present in a biological sample. An important application of gene expression profiling is to associate differences in mRNA mixtures originating from different groups of individuals with phenotypic differences between the groups. In contrast to genotyping, gene expression profiling allows characterization of the level of gene expression. A gene expression profile provides a quantitative overview of the mRNA transcripts that were present in a sample at the time of collection (Ness, 2007). Therefore, gene expression profiling can be used to determine which genes are differentially expressed in disease conditions; these genes would then serve as disease biomarkers.

Recent advances in bioinformatics and high-throughput technologies such as microarray analysis are bringing about a revolution in our understanding of cell biology and the molecular mechanisms underlying normal and dysfunctional biological processes. This field of omics is also stimulating the discovery of new targets for the treatment of disease, which is aiding drug development, immunotherapeutics, and gene therapy. Gene expression profiling has enabled the measurement of thousands of genes in a single RNA sample. There are a variety of microarray platforms from companies such as Affymetrix, Agilent, NimbleGen, and Illumina that have been developed to accomplish this. The basic idea for each platform is simple: a glass slide or membrane is spotted or arrayed with DNA fragments or oligonucleotides that represent specific gene coding regions (Ness, 2007; Bier, von Nickisch-Rosenegk et al., 2008). Purified RNA is then fluorescently or radioactively labeled and hybridized to the slide/membrane. In some cases, hybridization is done simultaneously with reference RNA to facilitate comparison of data across multiple experiments. After thorough washing to remove nonspecific hybridization, the data can be analyzed by a variety of statistical algorithms by comparing the gene expression pattern of samples tested to identify differentially expressed genes (Holland, Smith et al., 2003) that could potentially serve as disease biomarkers.

The most popular platform is the short oligonucleotide chips produced by Affymetrix. The second major platform consists of printed cDNA fragments or a long oligonucleotide (45–80-mers) on glass slides or other types of solid support. Dissection of global changes in gene expression

during predisease states, during disease progression, and following clinical treatment can provide great insight into disease mechanism and treatment management. For example, early investigations using microarrays distinguished acute myeloid and acute lymphoblastic cell gene expression patterns (Golub, Slonim et al., 1999). Subsequent studies have used microarray technology to predict outcomes in breast and ovarian cancers (Berchuck, Iversen et al., 2005; Huang, Song et al., 2003). Additionally, it has been shown that classification of diffuse large B-cell lymphomas on the basis of gene expression profiles can identify clinically significant subtypes of cancer, and the new classification has significant prognostic implications (Alizadeh, Eisen et al., 2000; Alizadeh and Staudt, 2000). Examination of systemic lupus erythematosus using microarray technology identified a subgroup of patients who may benefit from new therapeutic options (Baechler, Batliwalla et al., 2003). Novel treatments for diseases, such as multiple sclerosis, have also been suggested by gene expression profiling (Chabas, Montfort et al., 2001; Chabas, Baranzini et al., 2001). Genomic biomarkers are currently being identified in cardiovascular diseases in a large-scale study using microarray technology (unpublished data).

1.2.3 Pharmacogenomics

Pharmacogenomics is the study of how an individual's genetic inheritance affects the body's response to drugs (Evans and Relling, 1999). The field of pharmacogenomics is an intersection of pharmaceuticals and genetics and specifically studies the variability in drug response caused by heredity. The way a person responds to a drug (in both a positive and negative manner) is a complex trait that is influenced by many different genes. Without knowing all of the genes involved in drug response, scientists have found it difficult to develop genetic tests that could predict a person's response to a particular drug. A person's response to a particular drug is the result of inherited variations in genes that dictate drug response and omics researchers are exploring the ways in which these variations can be used to predict whether a patient will have a good response or a bad response or no response at all to a particular drug. For example, in their study, Johnson and Evans et al. (2001) examined the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity. Pharmacogenomics is believed to be immensely helpful in reducing drug-caused morbidity and mortality (Algeciras-Schimnich, O'Kane et al., 2008). Pharmacogenomics has gained considerable momentum with the advent of new methods and technologies for genome analysis and is widely believed to play a major role in predictive and personalized medicine (Roden, Altman et al., 2006). It will have the most impact in areas such as oncology, where many therapies are available, but each one works only for a small percentage of cancer patients. Pharmacogenomics is also expected to help physicians and patients by enabling pharmaceutical companies to bring more drugs into the market that are targeted at those patients who are most likely to benefit from them. Pharmacogenomics holds great promise in personalized medicine by providing physicians an opportunity to individualize drug therapy for patients based on their genetic make-up.

1.2.4 Toxicogenomics

The field of toxicogenomics is used in the study of structure and output of the genome as it responds to adverse xenobiotic exposure and is very closely related to pharmacogenomics. Toxicology has traditionally been evaluated by the dosing of animals to define well-established cytologic, physiologic, metabolic, and morphologic endpoints (Ferrer-Dufol and Menao-Guillen, 2009; Ge and He, 2009; Luch, 2009) . The evaluation of the risk to humans cannot be performed in human individuals initially and thus must be derived from studies performed in other species. Typically, rodents are used to identify toxic substances such as carcinogens, reproductive toxins, and neurotoxins. Follow-up studies in nonrodent species (species extrapolation) can then be used to further define the effects of low doses and mechanism of action.

Although it is well recognized that intact animals are needed to reflect physiologic changes and mirror the effects of chronic dosing, such studies have disadvantages (Guguen-Guillouzo and Guillouzo, 2010; Mei, Fuscoe et al., 2010; Moreira, Yu et al., 2010; Pettit, des Etages et al., 2010; Thompson, 2010; Van Aggelen, Ankley et al., 2010). Experiments with animals may not be fully predictive of the response in humans because of species variation in physiology, anatomy, and metabolism. Also, toxicology studies require large numbers of animals to allow statistically significant conclusions to be drawn. Nevertheless, these numbers are still very small compared to the human population potentially at risk. In order to compensate for this relatively small sample size in these animal studies, the future risk to humans at therapeutic dosages is inferred by giving large doses of compound to these groups of animals. Finally, depending on the anticipated duration of exposure in the population, studies of up to 2 years are currently mandated to determine the carcinogenic potential. Thus, the traditional approach to toxicologic testing is costly, in terms of time, labor, and compound synthesis and, not least, the large numbers of animals.

Technological advances have now enabled scientists to simultaneously analyze thousands of genes of several species, including humans and rodents, quickly and in a reproducible manner. Current toxicogenomics applies genomics concepts and technologies to study adverse effects of chemicals. These studies use global gene expression analyses to detect expression changes that influence, predict, or help define drug toxicity. In essence, toxicogenomics combines the tools of traditional toxicology with those of genomics and bioinformatics (Zarbl, 2007). By evaluating and characterizing differential gene expression after exposure to drugs, it is possible to use complex expression patterns to predict toxicologic outcomes and to identify mechanisms involved with or related to the toxic event. Toxicogenomics thus combines conventional toxicology with the emerging technologies of genomics and bioinformatics. Gene and protein expression respond specifically to external stimuli such as pathological conditions or exposure to drugs. The corresponding genomic and proteomic technologies thus provide a new way of understanding biological systems and their response to toxic insult. This leads to a better understanding of the mechanisms of toxicity by the identification of toxicity-related gene expression signatures and the prediction of the toxic potential of unknown compounds by comparing their gene expression profiles to the fingerprints of known, similar compounds (Gant, 2003; Shostak, 2005). In addition, the identification of toxicity-related genes, together with the rapidly growing understanding of the human genome, is providing a basis for identifying and characterizing sequence variations in genes that might affect responses to chemicals. This is already having a great impact in pharmacology and toxicology, because it allows the prediction/differentiation of species-specific responses and also the identification of populations of responders and nonresponders (Mei, Fuscoe et al., 2010). The most optimistic estimates predict that the replacement of traditional methods of toxicology by toxicogenomics could eventually shorten the safety assessment of a new chemical entity from years to days and reduce costs by an estimated factor of four to six times. A more realistic picture with the data currently available suggests that toxicogenomics will reduce failure rates by helping select the right compounds for development early on and by accelerating toxicology testing and identifying suitable biomarkers amenable to screening using the generated data (Pettit, des Etages et al., 2010; Choudhuri, 2009). Toxicogenomics represents an exciting new approach to toxicology and has a great potential to influence the predictability and speed of preclinical safety assessments (Choudhuri, 2009). Published results so far show that genome-wide gene expression analysis is a powerful tool for compound classification and for the detection of new, specific, and sensitive markers for given mechanisms of toxicity (Gallagher, Tweats et al., 2009; Ge and He, 2009; Hirode, Omura et al., 2009; Smirnov, Morley et al., 2009). In addition, preliminary results support the theory that gene expression might be more sensitive than conventional toxicology endpoints. Therefore, compound classification could be performed during early, short-term (i.e., single-dose) animal studies. Hence, time, cost, and number of animals needed to identify the toxic potential of a compound would be greatly minimized. The potential identification and validation of possible marker genes are also gaining momentum. Such markers could be employed in automated, high-throughput assay systems that will provide indications regarding

toxicity potential that are fast and accurate, without incurring the high costs commonly associated with microarray analysis. Appropriately chosen markers are amenable to being tested in cell-based assays that will allow scientists to evaluate compounds much earlier in the developmental process, improving clinical candidate selection. The understanding of the molecular mechanisms underlying toxicity obtained through gene expression analysis after exposure of model systems (animals or cell cultures) to test compounds will also provide more insight into species-specific response to drugs regarding efficacy and toxicity. Hence, it is expected that extrapolation across species will become more accurate by enhancing the interpretation of preclinical observations and their meaning for the human situation. This should immensely increase the predictability of toxic liabilities and of potential risk accumulation for drug combinations or drug-disease interactions.

1.2.5 Epigenomics

Epigenomics, the merged science of epigenetics and genomics, has arisen as a new discipline with the aim of understanding genetic regulation and its contribution to cellular growth and differentiation, disease, and aging. Epigenetics is the study of heritable changes other than those in the DNA sequence and encompasses two major modifications of DNA or chromatin: DNA methylation; the covalent modification of cytosine; and post-translational modification of histones, including methylation, acetylation, and phosphorylation (Banerjee and Verma, 2009). Functionally, epigenetics acts to regulate gene expression, silence the activity of transposable elements, and stabilize adjustments of gene dosage, as seen in X inactivation and genomic imprinting (Herceg, 2007). The focus of epigenomics is to study epigenetic processes on a genome-wide scale. Epigenetic processes are mechanisms other than changes in DNA sequences that are involved in gene transcription and gene silencing (Schubeler, 2009). Epigenetic studies are currently based mainly on DNA methylation, histone modification interference by noncoding RNAs such as microRNA, and small interference RNA mechanisms (Schubeler, 2009). Generally, gene silencing is observed during genomic imprinting, x-chromosome inactivation, and tissue-specific gene expression. Alteration to these patterns of gene silencing by epigenetic modification is believed to play an important role in human disease (Herceg, 2007).

Historically, technology has limited large-scale approaches to epigenomics, but the emergence of highly reproducible quantitative high-throughput microarray technology has allowed virtually all epigenomics research to be read on microarray platforms, although the substrates, preprocessing, and data analysis differs substantially depending on the modification that is being addressed (Adorjan, Distler et al., 2002). Multiple complementary technologies are emerging now to analyze DNA methylation, protein binding patterns, and chromatin regulation on a genome-wide level. Early efforts are providing glimpses into the epigenetics of gene regulation and the mechanism of cancer and aging. It is hoped that the development of high-throughput technologies will continue to unravel the enigma of the epigenome. Early approaches to epigenomics used custom-made slide-based arrays of CpG-rich regions corresponding to methylated or unmethylated DNA (Adorjan, Distler et al., 2002). There has been a shift toward commercial high-density oligonucleotide arrays because of their greater precision and potential quantitative character. These include the photolithographic masked arrays of Affymetrix, photolithographic adaptive optics arrays of NimbleGen, inkjet arrays of Agilent, and, recently, the adaptation of bead arrays for epigenetic applications of Illumina. Each of these approaches offers potential advantages and disadvantages, but as yet, no direct comparison of epigenomic technology has been performed across platforms. An advantage of a flexible design for epigenomics is that one can tailor arrays to genomic targets of interest, such as imprinted genes, differentially methylated regions, and imprinting control regions.

An example of an early step in approaching the epigenome comes from recent studies by Fraga et al. (Fraga and Esteller, 2007; Fraga, Agrelo et al., 2007) that address the relationship between epigenetics and age. Another exciting work by Beth Israel Deaconess Medical Center and the Broad Institute created a map of histone modifications in fat cells, which led to the discovery of

two new factors that regulate fat formation, a key step on the road to better understanding obesity, diabetes, and other metabolic disorders (Mikkelsen, Thomsen et al., 2010). Epigenetics thus appears to be an exciting area of investigation with the potential for effective new therapies in areas of unmet medical need and the development of new diagnostic, screening, or pharmacogenomic tests.

1.3 PROTEOMICS

Proteomics is the study of proteins, including their location, structure, and function. Proteomics involves the systematic study of proteins in order to provide a comprehensive view of the structure, function, and regulation of biological systems (Patterson and Aebersold, 2003). Although all proteins are based on mRNA precursors, post-translational modifications and environmental interactions make it impossible to predict the abundance of specific proteins based on gene expression analysis alone (Patterson, 2003). In contrast to the genome, the proteome is highly variable over time between cell types and will change in response to its environment. A major challenge is the high variability in proteins and protein abundance in biological specimens (Patterson and Aebersold, 2003). Advances in instrumentation and methodologies have fueled an expansion of the scope of biological studies from simple biochemical analysis of single proteins to measurement of complex protein mixtures. Coupled with advances in bioinformatics, this approach to comprehensively describe biological systems will undoubtedly have a major impact on our understanding of the phenotypes of both normal and diseased cells. Initially, proteomics focused on the generation of protein maps using two-dimensional polyacrylamide gel electrophoresis (PAGE) (Patterson, 2003). The field has since expanded to include not only protein expression profiling, but also the analysis of post-translational modifications and protein-protein interactions. Protein expression, or the quantitative measurement of the global levels of proteins, may still be done with two-dimensional gels; however, mass spectrometry has been incorporated to increase sensitivity and specificity and to provide results in a high-throughput format (Domon and Aebersold, 2006). A variety of platforms such as mass spectrometry, tandem mass spectrometry (MS/MS), and protein microarrays are now available to conduct proteome analysis on a cellular, subcelluar, and organ level (Yates, Gilchrist et al., 2005; Cox and Mann, 2007). The study of protein-protein interactions has been revolutionized by the development of protein microarrays. Analogous to DNA microarrays, these biochips are printed with antibodies or proteins and probed with a complex protein mixture (Ressine, Marko-Varga et al., 2007). The intensity or identity of the resulting protein-protein interactions may be detected by fluorescence imaging or mass spectrometry. Other protein capture methods may be used in place of arrays, including the yeast two-hybrid system or the isolation of protein-protein complexes by affinity chromatography or other separation techniques (Ralser, Goehler et al., 2005).

Although DNA microarray technology provides a wealth of information about the expression and roles of RNA transcripts in states of disease, it is critically important to associate the events at the level of transcription with the actual proteins that are being encoded, translated, and modified. Using multidimensional gel electropheresis, high-throughput mass spectroscopy, various low density arrays for protein-protein interactions, or protein-specific antibody arrays, it is possible to study the proteomes of cells, tissues, and body fluids in search of disease-linked proteins. At the molecular and cellular level, biological functions are carried out by proteins rather than DNA or RNA (with the possible exception of ribozymes) (Kurian, Kirk et al., 1998). Thus, information obtained by proteomic analysis greatly complements data obtained from DNA microarrays.

A major technical challenge for proteomics is the significant increase in the complexity of the proteome, representing several hundred thousand or more proteins, as compared to the RNA transcriptome, which represents about 20,000–30,000 genes total. A major cause for this increased proteomic complexity is splice variants of genes that are manifested as different protein products. Another mechanism is that protein function and activity is regulated or restricted by

post-translational and covalent modifications of protein structure (i.e., phosphorylation, sulfation, methylation, and glycosylation), as well as other protein-protein interactions or protein-small molecule interactions. Thus, it is equally important to develop technologies to study the post-translational events of proteins that dictate the biological microenvironment of the cells and tissues and, thus, the entire organism (Sellers and Yates, 2003; Pan, Chen et al., 2008; Pan, Kumar et al., 2009; Pan, Aebersold et al., 2009).

Proteomic analysis is expected to have wide application in the field of medicine by providing unique information about cells and tissues and eventually creating noninvasive tests to monitor biomarkers in body fluids, such as urine or blood, that would correlate clinical analysis. A proteomics application to monitor transplantation acceptance was reported (Pan, Chen et al., 2008; Pan, Kumar et al., 2009; Pan, Aebersold et al., 2009) using 2D PAGE and matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) in a rat model of liver transplantation. The authors found that haptoglobin, which has been associated with inhibition of T-cell proliferation in studies of cancer patients and some *in vitro* culture assays, was up-regulated following liver transplantation. As additional proof, the level of RNA transcript expression and intracellular localization of haptoglobin correlated with the immune events in the liver, a good example of how proteomics can complement genomics. In the field of kidney transplantation, one of the earliest searches to identify potential biomarker candidates from the urine was performed with surface enhanced laser desorption ionization (SELDI)-TOF mass spectroscopy (Clarke, Silverman et al., 2003). A study in human kidney transplantation using the same technology, SELDI-TOF mass spectroscopy, profiled urinary protein spectra from five groups of subjects: acute rejection, acute tubular necrosis, recurrent or de novo glomerulopathy, stable transplant patients with excellent function, and normal urine donor controls (Schaub, Rush et al., 2004; Schaub, Wilkins et al., 2004a; Schaub, Wilkins et al., 2004b). Two distinct urine protein patterns were observed when comparing the normal controls and stable transplant groups to the acute rejection group. A more recent study looked at the differentiation of BK virus-associated nephropathy from acute allograft rejection in kidney-transplant recipients (Jahnukainen, Malehorn et al., 2006). A plethora of biomarkers exist for diagnosis of nutritional status, metabolic diseases (carbohydrate, amino acid, and fatty acid metabolism), inflammation (C-reactive protein, haptoglobin, orosomucoid, and anti-trypsin) (Agarwal, Binz et al., 2005; Sadrzadeh and Bozorgmehr, 2004; Kanikowska, Grzymislawski et al., 2005; Kanikowska, Hyun et al., 2005), hormonal imbalance (insulin, thyroxine, adrenaline, and pituitary hormones), tissue damage (aspartate transaminase and alanine transaminase for liver and heart, collagen for joints) (Collier, Lecomte et al., 2002; Conigrave, Davies et al., 2003; Collier and Bassendine, 2002; Poole, 2003), cancer (CA15.3, CA27.29, CEA, PSA, S100-β, and hCG) (Rosai, 2003; Shitrit, Zingerman et al., 2005), neurodegeneration (amyloid plaques, and β-amyloid peptide) (Aslan and Ozben, 2004; Bossy-Wetzel, Schwarzenbacher et al., 2004; Teunissen, de Vente et al., 2002), and autoimmune diseases (autoantibodies) (Pender, Csurhes et al., 2000; Masaki and Sugai, 2004; Weetman, 2004a; Weetman, 2004b).

Candidate biomarkers have been identified for a number of diseases, including cancers of different origins (e.g., ovary, breast, and prostate) (Rapkiewicz, Espina et al., 2004), neurological disorders (Austen, Frears et al., 2000), and pathogenic organisms (Lancashire, Schmid et al., 2005), and Alzheimer's disease and diabetes.

1.4 METABOLOMICS

The metabolome consists of small molecules that are involved in the energy transmission in the cells by interacting with other biological molecules following metabolic pathways. In cells, the rate of enzymatic reactions is also regulated by metabolites. The metabolome is highly variable and time dependent and consists of a wide range of chemical structures (Fridman and Pichersky, 2005). It is also important to point out here that metabolomics and metabonomics are generally

interchangeable terms. Metabolic phenotypes are the by-products that result from the interaction between genetic, environment, lifestyle, and other factors (Fridman and Pichersky, 2005). Metabolomics, as a method to define the small molecule diversity in the cell and to display differences in small molecule abundance, shows many advantages in terms of metabolic analyses because metabolites are the functional entities within the cells, and their concentration levels vary as a consequence of genetic or physiological changes. An important challenge of metabolomics is to acquire qualitative and quantitative information concerning the metabolites that are perturbed because of changes in environmental factors. Metabolomics analysis is typically performed by employing gas chromatography time-of-flight mass spectrometry, high performance liquid chromatography-mass spectrometry, or capillary electrophoresis mass spectrometry instruments, nuclear magnetic resonance spectroscopy, and more recently vibrational spectroscopy (Robertson, Reily et al., 2005). Metabolome analysis can also be performed through combined application of several technologies together in order to achieve wide coverage and better identification. Compared with transcriptomics and proteomics, improvements in instrumentation and data analysis software are still needed for metabolomic studies.

In animals and humans, metabolic profiling of body fluids to characterize metabolic disorders has been ongoing since the introduction of gas chromatography and mass spectrometry. Nuclear magnetic resonance (NMR) techniques have also been applied for a wide range of components of blood and urine. Current metabolomic studies are making use of technologies such as mass spectrometry (MS), gas chromatography/MS, and NMR to produce metabolic profiles or signatures of toxicity, disease, and drug efficacy. A major aspect of organismal biology is the metabolism and elimination of proteins, hormones, and exogenous molecules, including drugs. In fact, if a given drug therapy resulted in a set of molecular events that created a unique metabolome detected in blood plasma, for example, these metabolic biomarkers could be highly specific as metrics for therapeutic efficacy but actually not be comprised of any of the metabolites of the drug. In other settings, it is hoped that metabolomic profiles of drugs will also correlate with unwanted and dangerous side effects and could therefore be used to enhance the safety of drug therapy.

Metabolic signatures provide prognostic, diagnostic, and surrogate markers for a disease state. For example, NMR spectroscopy of urine and plasma samples was used to examine early graft dysfunction in a pig ischemia/reperfusion model (Holland, Smith et al., 2003) in order to assess and predict early graft dysfunction (Kurian, Flechner et al., 2005). In another study, NMR spectroscopy in combination with pattern recognition tools was used to investigate the composition of organic compounds in urine from patients with multiple sclerosis, patients with other neurological diseases, and healthy controls (Holland, Pfleger et al., 2005). Using the marmoset monkey model of experimental autoimmune encephalomyelitis, the relation of disease progression and alteration of the urine composition was investigated and compared with the measurements obtained with the human patient samples. A recent study has led to the development of a new statistical paradigm to coanalyze NMR and ultra performance liquid chromatography combined with orthogonal acceleration TOF-MS data (Heverhagen, Hartlieb et al., 2002; Hutcheson, Canning et al., 2002) across different samples of urine. Application of these tools has been shown to improve the efficiency of biomarker identification. Finally, another source for metabonomic biomarkers is the low-molecular-weight range serum proteome, the peptidome, which may also contain disease-specific information (Hu, Ye et al., 2009). This seems to be an untapped resource of candidates for new and specific biomarkers, because it is comprised of a multitude of small protein fragments that present a recording or snapshot of events taking place at the level of disease-associated microenvironments. Because intact tissue proteins are too large to passively diffuse through the cell and across the endothelial basement membranes into the circulation, the peptidome could provide an accessible portal to identify and quantify a wide range of protein changes that are taking place in all of the cells and tissues (Hu, Ye et al., 2009). Therefore, metabolomics appears to be a valuable platform for studies of complex diseases and for the development of new therapies both in nonclinical disease model characterization and in clinical settings.

1.5 CELLOMICS

The field of cellomics was driven by the need to define the functions of genes and the proteins that they encoded. It was apparent by the mid-1990s that knowing the human genome was the start not the end of the biological challenge for basic research and drug discovery. Light microscopy, especially digital imaging fluorescence microscopy on living cells, was chosen as the best approach to defining the functions of genes and proteins (Yasuda, 2010). Human interactive imaging methods were pretty well developed by the 1980s, and fundamental information about the temporal and spatial dynamics of cells and their constituents was being published by a growing academic community. However, the human interactive imaging tools in the absence of automated imaging methods and informatics tools to archive, mine, and display complex imaging data made the process of studying cells time consuming and complicated (Yasuda, 2010). Similar to the field of genomics, there was a need for the development of an automated system to acquire, process, analyze, display, and mine massive amounts of cellular data derived from arrays of cells treated in various ways. This need for high-content screening of cells has paved the way for developing novel technologies such as automated digital microscopy and flow cytometry and Arrayscan, to offer a complete solution for single cell analysis. These technologies are currently being put to use in biomedicine.

1.6 LIPIDOMICS

Lipidomics, the systems-level analysis of lipids and their interacting partners, can be viewed as a subdiscipline of metabolomics. An enormous number of chemically distinct molecular species arise from the various combinations of fatty acids with backbone structures (Blanksby and Mitchell, 2010; Shevchenko and Simons, 2010). Lipidomics is the emerging field of systems-level analysis of lipids and factors that interact with lipids (Wenk, 2005). Although important, the study of lipids has been hampered by analytical limitations. Lipids are molecules that are highly soluble in organic solvents. It is clear, however, that without special precautions many classes of lipid molecules (such as the very polar phosphoinositides) will escape into the aqueous milieu during phase partitioning (Brown and Murphy, 2009).

Lipids, the fundamental components of biological membranes, play multiple important roles in biological systems. The most important functions are creating in the cell a subsystem in the context of the whole and relatively independent of the exterior environment through lipid bilayer structures, providing an appropriate hydrophobic medium for the functional implementations of membrane proteins and their interactions and producing second messengers by enzyme reactions (Brown and Murphy, 2009). Abnormal lipid metabolism has been observed in numerous human diseases such as diabetes, obesity, atherosclerosis, and Alzheimer's disease, leading to tremendous interest in lipid research in biomedical research (Aukrust, Muller et al., 1999; Hjelmesaeth, Hartmann et al., 2001). Current research on lipids tends to shift from determining the individual molecular structures of single lipids in biological samples to characterizing global changes of lipid metabolites in a systems-integrated context in order to understand the crucial role of lipids in physiopathology (Wenk, 2005). Traditional strategies for lipid analysis usually prefractionate lipids into classes using thin-layer chromatography normal-phase liquid chromatography, or solid-phase extraction and then separate particular classes of lipids into individual molecular species by high-performance liquid chromatography coupled with either ultraviolet or evaporative light-scattering detector. However, such classical techniques often either lack sensitivity or require large sample volumes and multi-step procedures for sample preparation, and the resolution is limited, i.e. only a limited set of individual molecular species are analyzed. Recent advancements in mass spectrometry and innovations in chromatographic technologies have largely driven the development of high-throughput analysis of lipids. With the advent of soft ionization, technologies such as matrix-assisted laser desorption/ionization, electrospray ionization, and atmospheric pressure chemical ionization for MS, possibly coupled to liquid chromatography (LC) rapid and sensitive analysis of the majority or a substantial

fraction of lipids possible in one analysis, is currently possible. Most common strategies currently used in lipidomics include direct-infusion electrospray ionization (ESI)-MS and ESI-MS/MS, LC coupled with ESI-MS or MS/MS, and MALDI combined with Fourier transform ion cyclotron resonance MS or MALDI-TOF-MS) (van Meer, 2005; Wenk, 2005).

Despite all advances recently made, the diversity of structures and properties and the wide range of concentrations of lipids provide a huge and almost impossible challenge for analytical methodology when aiming at a single technological platform capable of measuring and identifying all lipids in a single sample simultaneously. As a consequence, multiple, often complementary, analytical approaches are currently used in the field of lipidomics (van Meer, 2005).

1.7 GLYCOMICS

The term glycomics is derived from the chemical prefix for sweetness or a sugar, *glyco*, and was formed to follow the naming convention established by genomics and proteomics (Liang, Wu et al., 2008). Glycomics is an integrated approach to study structure-function relationships of complex carbohydrates or glycans such as glycolipids, glycoproteins, lipopolysaccharides, peptidoglycans, and proteoglycans. Comparative studies of specific carbohydrate chains of glycoproteins can provide useful information for the diagnosis, prognosis, and immunotherapy of tumors. Glycan-based drugs have generated much excitement and provided important insight into the power of glycan-based therapeutics. However, the ultimate promise of glycans as drugs is only beginning to be exploited. The emerging omics domain of glycomics has lagged behind that of genomics and proteomics, mainly because of the inherent difficulties in analysis of glycan structure and functions. A wide variety of technologies are now being brought to bear on the technically difficult problems of glycan structural analysis and investigation of functional roles. Enabling technologies such as high-throughput mass spectroscopy, glycan microarrays, aminoglycoside antibiotic microarrays and glycan sequencing, quantum dots, and gold nano particles are currently helping to unravel the complexity resulting from diverse glycans in biological systems (Liang, Wu et al., 2008). In an effort to harness the promise of glycans as therapeutics, advances have been made in analyzing glycan structures in a rapid manner using a minimum of material, in synthesizing glycan structures *in vitro*, and in harnessing endogenous glycosylation pathways *in vivo* to create new reproducible glycan structures. Recently, there has been a marked increase in the reporting of techniques that have been successfully applied to the analysis of complex glycans and glycoconjugates, including MS (Kaji, Saito et al., 2003; Zhang, Cocklin et al., 2003) and capillary electrophoretic (Que, Mechref et al., 2003) techniques. Many of these technologies have distinct advantages compared with traditional analytical methodologies, including the ability to analyze minute amounts of biologically based material.

Comparative studies of glycans can provide useful information for the diagnosis, prognosis, and therapy for several diseases. For example, glycoproteomics analysis was used to discover serum markers in liver cancer; GP73, a glycoprotein, was found to be elevated in hepatocellular carcinoma, and this marker have been successfully used as a positive predictor of diagnosis and treatment (Marrero and Lok, 2004).

1.8 FUTURE PROSPECTS OF OMICS: INTEGRATION OF OMICS TECHNOLOGIES AND APPLICATIONS IN CLINICAL MEDICINE

The omics field is now transforming biomedical research where one gene or protein was studied at a time to a world in which whole organelles and pathways are studied simultaneously using less biological material. An integrative approach using the data collected by various omics platforms developed by companies shown in Table 1.1 is expected to fulfill the dream of specific disease biomarkers, individualized care, and treatment of human diseases. Whereas high-throughput omics approaches to analyze molecules at different cellular levels are rapidly becoming available, it is

TABLE 1.1
List of Biotechnology Companies and Their Resources

Biotechnology Companies	Technology	Tools
Affymetrix	Genomics, microarrays, sequencing	Gene chips
Agilent Technologies	Genomics, protomics, microarrays, sequencing	Glass arrays
Applied Biosystems/Life Technologies	Genomics, protomics, microarrays, sequencing	Arrays
Axcell Biosciences	Informatics	Proteomic database
Bio-Rad	Genomics	Arrays
Celera	Genomics	Informatics
Cellomics Inc.	Cellomics	High content screening
Ciphergen Biosystems	Proteomics, toxicogenomics	Proteinchip systems
Decode Genetics	Genomics, proteomics	Bioinformatics
Epigenomics	Epigenomics, pharmacogenomics	Arrays
Global Lipidomics	Lipidomics	Service
Human Metabolome Technologies	Metabolomics	CE-MS tool
Illumina	Genomics	Bead arrays
Incyte Genomics	Proteomics, toxicogenomics	Informatics
Large Scale Biology Corp	Proteomics, toxicogenomics	2-DE
Nimblegen Systems	Genomics	Arrays
PerkinElmer	Genomics, proteomics	Reagents
Proteome Inc.	Proteomics, toxicogenomics	Service
Proteome Sciences plc.	Proteomics, toxicogenomics	2D gel electrophoresis
Sequenom Inc.	Genomics	DNA analysis tools
Sigma-Aldrich	Genomics	Reagents, arrays

also becoming clear that any single omics approach may not be sufficient to characterize the complexity of biological systems (Gygi, Han et al., 1999; Gygi, Rochon et al., 1999). For example, the expression level of a given gene does not indicate the amount of protein produced nor its location, biological activity, or functional relationship with metabolomes. Moreover, in cells, many levels of regulation occur after genes have been transcribed, such as post-transcriptional, translational, and post-translational regulation and all forms of biochemical control such as allosteric or feedback regulation. For example, in a study by ter Kuile and Westerhoff (2001), control of glycolysis was shown to be shared between metabolic, proteomic, and genomic levels, thereby suggesting that the functional genomics cannot stop at the mRNA level or any single level of information. Integrated multiomics approaches have been applied recently, and the studies have enabled researchers to unravel global regulatory mechanisms and complex metabolic networks in various eukaryotic organisms (Hegde, White et al., 2003; Mootha, Bunkenborg et al., 2003; Ray, Mootha et al., 2003; Alter and Golub, 2004). These early studies have clearly demonstrated that integrated omics analysis may be a key to decipher complex biological systems. It is widely believed that the application of the currently available omics technologies as depicted in Figure 1.3 will not only have an impact on our understanding of biological processes, but will also improve the prospect of more accurately diagnosing and treating diseases. The reality of applying omics technologies to unravel disease processes, identify disease biomarkers, and finally make a recommendation for personalized medicine is expected to revolutionize medical practice. Several examples of different omics technologies that have been integrated and methods used in studying diseases in human and animal models have been summarized in Table 1.2.

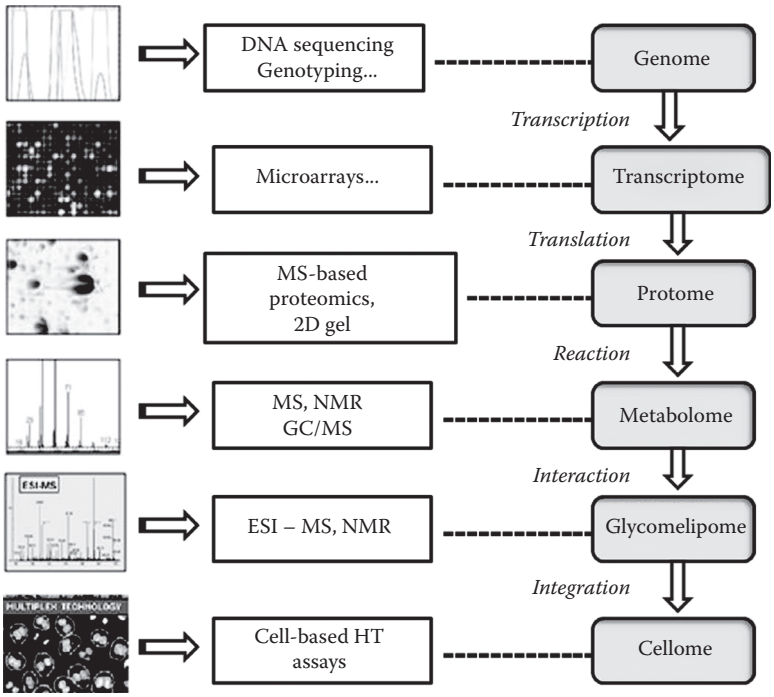

FIGURE 1.3 Technologies for omics analysis.

TABLE 1.2
Integration of Omics Technologies and Applications in Clinical Medicine

Disease Pathology	Omics Technology	References
Alzheimer's disease, Parkinson's disease, and multiple sclerosis	Metabolomics, mass spectrometry	Alimonti, Ristori et al. (2007)
Coronary disease	Lipidomics, liquid chromatography-mass spectrometry	Bergheanu, Reijmers et al. (2008)
Parkinson disease	Metabolomics, high-performance liquid chromatography, electrochemical colorimetric array detection	Bogdanov, Matson et al. (2008)
Biomarkers/kideney transplanation	SELDI-TOF, mass spectrometry	Clarke, Silverman et al. (2003); Schaub, Wilkins et al. (2004)
Diabetes, obesity, coronary heart disease	Functional genomics, metabonomics, NMR spectroscopy, mass spectrometry	Brindle (2002); Griffin and Vidal-Puig (2008)
Muscular dystrophy in mice	Metabolomics, NMR	Griffin (2006)
Crohn's disease and ulcerative colitis	Proteomics, genomics, protein microarrays	Kader, Tchernev et al. (2005)
Ovarian cancer	Glycomics, mass spectrometry (MALDI-FTMS)	Leiserowitz, Lebrilla et al. (2008)
Leigh syndrome, mitochondrial complex I deficiency	Proteomics, PAGE, LC-MS/MS, genomics, homozygosity mapping, Affymetrix GeneChip mapping	Pagliarini, Calvo et al. (2008)
Various cancers	Genomics, transcriptomics, RNA interference	Pai, Lin et al. (2006)
Organ transplantation/rejection	Proteomics/2D PAGE/MALDI-TOF	Pan, Jain et al. (2004)

TABLE 1.2 (CONTINUED)
Integration of Omics Technologies and Applications in Clinical Medicine

Disease Pathology	Omics Technology	References
Kidney cancer	Proteomics, metabolic profiling, PAGE, MS, immunoblotting	Perroud, Lee et al. (2006)
Type II diabetes and dyslipidemia	Metabolomics, biofluid NMR spectroscopy	Ringeissen, Connor et al. (2003)
Amyotrophic lateral sclerosis in a mouse model	Genomics, proteomics, immunochemistry, genetic engineering, gene silencing	Saito, Yokota et al. (2005)
Gene expression in human liver	Genomics, expression profiling, genotyping	Schadt, Molony et al. (2008)
Phenylketonuria	Genomics, population genetics, metabolomics	Scriver (2007)
Alzheimer's disease	Genotyping, genomics	Seshadri, Fitzpatrick et al. (2010)
Normal glucose metabolism, homeostasis, insulin sensitivity	Metabolic profiling, metabolomics, LC-MS/MS, radioimmunoassay, hexokinase assay	Shaham, Wei et al. (2008)
Rheumatoid arthritis, hypertension, Crohn's disease, coronary artery disease, bipolar disorder, diabetes	Genomics, genome-wide association, genotyping, GeneChip arrays	The Wellcome Trust Case Control Consortium (2007)
Crohn's disease and ulcerative colitis	Genomics, expression microarrays, quantitative reverse transcription-polymerase chain reaction	Wu, Dassopoulos et al. (2007)
Plant storage proteins, allergens	Proteomics, affinity columns, PAGE	Yano and Kuroda (2008)

REFERENCES

Abraham, V. C., Taylor, D. L., et al. (2004). High content screening applied to large-scale cell biology. *Trends Biotechnol.* 22: 15–22.

Adorjan, P., Distler, J., et al. (2002). Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* 30: e21.

Agarwal, R., Binz, T., et al. (2005). Analysis of active site residues of botulinum neurotoxin E by mutational, functional, and structural studies: Glu335Gln is an apoenzyme. *Biochemistry* 44: 8291–8302.

Algeciras-Schimnich, A., O'Kane, D. J., et al. (2008). Pharmacogenomics of tamoxifen and irinotecan therapies. *Clin. Lab. Med.* 28: 553–567.

Alimonti, A., Ristori, G., et al. (2007). Serum chemical elements and oxidative status in Alzheimer's disease, Parkinson disease and multiple sclerosis. *Neurotoxicol*. 28: 450–456.

Alizadeh, A. A., Eisen, M. B., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503–511.

Alizadeh, A. A., and Staudt, L. M. (2000). Genomic-scale gene expression profiling of normal and malignant immune cells. *Curr. Opin. Immunol.* 12: 219–225.

Alter, O., and Golub, G. H. (2004). Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc. Natl. Acad. Sci. U.S.A.* 101: 16577–16582.

Aslan, M., and Ozben, T. (2004). Reactive oxygen and nitrogen species in Alzheimer's disease. *Curr. Alzheimer Res.* 1: 111–119.

Aukrust, P., Muller, F., et al. (1999). Enhanced levels of soluble and membrane-bound CD40 ligand in patients with unstable angina: Possible reflection of T lymphocyte and platelet involvement in the pathogenesis of acute coronary syndromes. *Circulation* 100: 614–620.

Austen, B. M., Frears, E. R., et al. (2000). The use of seldi proteinchip arrays to monitor production of Alzheimer's betaamyloid in transfected cells. *J. Pept. Sci.* 6: 459–469.

Baechler, E. C., Batliwalla, F. M., et al. (2003). Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. U.S.A.* 100: 2610–2615.

Banerjee, H. N., and Verma, M. (2009). Epigenetic mechanisms in cancer. *Biomark Med.* 3: 397–410.

Barron, A. (2008). DNA sequencing and genotyping. *Electrophoresis* 29: 4617.

Berchuck, A., Iversen, E. S., et al. (2005). Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin. Cancer Res.* 11: 3686–3696.

Bergheanu, S. C., Reijmers, T., et al. (2008). Lipidomic approach to evaluate rosuvastatin and atorvastatin at various dosages: Investigating differential effects among statins. *Curr. Med. Res. Opin*. 24: 2477–2487.
Bier, F. F., von Nickisch-Rosenegk, M., et al. (2008). DNA microarrays. *Adv. Biochem. Eng. Biotechnol.* 109: 433–453.
Blanksby, S. J., and Mitchell, T. W. (2010). Advances in mass spectrometry for lipidomics. *Annu. Rev. Anal. Chem.* 3: 433–465.
Bogdanov, M., Matson W. R., et al. (2008). Metabolomic profiling to develop blood biomarkers for Parkinson's disease. *Brain* 131: 389–396.
Bossy-Wetzel, E., Schwarzenbacher, R., et al. (2004). Molecular pathways to neurodegeneration. *Nat. Med.* 10(Suppl.): S2–S9.
Brindle, K. M. (2002). Detection of apoptosis in tumors using magnetic resonance imaging and spectroscopy. *Adv. Enzyme Regul*. 42: 101–112.
Brown, H. A., and Murphy, R. C. (2009). Working towards an exegesis for lipids in biology. *Nat. Chem. Biol.* 5: 602–606.
Chabas, A., Montfort, M., et al. (2001). Mutation and haplotype analyses in 26 Spanish Sanfilippo syndrome type A patients: Possible single origin for 1091delC mutation. *Am. J. Med. Genet.* 100: 223–228.
Chabas, D., Baranzini, S. E., et al. (2001). The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* 294: 1731–1735.
Choudhuri, S. (2009). Looking back to the future: From the development of the gene concept to toxicogenomics. *Toxicol. Mech. Methods* 19: 263–277.
Clarke, W., Silverman, B. C., et al. (2003). Characterization of renal allograft rejection by urinary proteomic analysis. *Ann. Surg.* 237: 660–665.
Collier, J., and Bassendine, M. (2002). How to respond to abnormal liver function tests. *Clin. Med.* 2: 406–409.
Collier, T. L., Lecomte, R., et al. (2002). Assessment of cancer-associated biomarkers by positron emission tomography: Advances and challenges. *Dis. Markers* 18: 211–247.
Conigrave, K. M., Davies, P., et al. (2003). Traditional markers of excessive alcohol use. *Addiction* 98(Suppl. 2): 31–43.
Cox, J., and Mann, M. (2007). Is proteomics the new genomics? *Cell* 130: 395–398.
Crick, F. H. (1954). The Complementary Structure of DNA. *Proc. Natl. Acad. Sci. U.S.A.* 40: 756–758.
Domon, B., and Aebersold, R. (2006). Mass spectrometry and protein analysis. *Science* 312: 212–217.
Evans, W. E., and Relling, M. V. (1999). Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* 286: 487–491.
Ferrer-Dufol, A., and Menao-Guillen, S. (2009). Toxicogenomics and clinical toxicology: An example of the connection between basic and applied sciences. *Toxicol. Lett.* 186: 2–8.
Fraga, M. F., Agrelo, R., et al. (2007). Cross-talk between aging and cancer: The epigenetic language. *Ann. N.Y. Acad. Sci.* 1100: 60–74.
Fraga, M. F., and Esteller, M. (2007). Epigenetics and aging: The targets and the marks. *Trends Genet.* 23: 413–418.
Fridman, E., and Pichersky, E. (2005). Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Curr. Opin. Plant Biol.* 8: 242–248.
Gallagher, W. M., Tweats, D., et al. (2009). Omic profiling for drug safety assessment: Current trends and public-private partnerships. *Drug Discov. Today* 14: 337–342.
Gant, T. W. (2003). Application of toxicogenomics in drug development. *Drug News Perspect.* 16: 217–221.
Ge, F., and He, Q. Y. (2009). Genomic and proteomic approaches for predicting toxicity and adverse drug reactions. *Expert Opin. Drug Metab. Toxicol.* 5: 29–37.
Golub, T. R., Slonim, D. K., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537.
Griffin, J. L. (2006). Understanding mouse models of disease through metabolomics. *Curr. Opin. Chem. Biol*. 10: 309–315.
Griffin, J. L., and Vidal-Puig, A. (2008). Current challenges in metabolomics for diabetes research: A vital functional genomic tool or just a ploy for gaining funding? *Physiol. Genom*. 34: 1–5.
Guguen-Guillouzo, C., and Guillouzo, A. (2010). General review on in vitro hepatocyte models and their applications. *Methods Mol. Biol.* 640: 1–40.
Gygi, S. P., Han, D. K., et al. (1999). Protein analysis by mass spectrometry and sequence database searching: Tools for cancer research in the post-genomic era. *Electrophoresis* 20: 310–319.
Gygi, S. P., Rochon, Y., et al. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19: 1720–1730.

Hegde, P. S., White, I. R., et al. (2003). Interplay of transcriptomics and proteomics. *Curr. Opin. Biotechnol.* 14: 647–651.

Herceg, Z. (2007). Epigenetics and cancer: Towards an evaluation of the impact of environmental and dietary factors. *Mutagenesis* 22: 91–103.

Heverhagen, J. T., Hartlieb, T., et al. (2002). Magnetic resonance cystometry: Accurate assessment of bladder volume with magnetic resonance imaging. *Urology* 60: 309–314.

Hirode, M., Omura, K., et al. (2009). Gene expression profiling in rat liver treated with various hepatotoxic-compounds inducing coagulopathy. *J. Toxicol. Sci.* 34: 281–293.

Hjelmesaeth, J., Hartmann, A., et al. (2001). Metabolic cardiovascular syndrome after renal transplantation. *Nephrol. Dial. Transplant.* 16: 1047–1052.

Holland, N. T., Pfleger, L., et al. (2005). Molecular epidemiology biomarkers: Sample collection and processing considerations. *Toxicol. Appl. Pharmacol.* 206: 261–268.

Holland, N. T., Smith, M. T., et al. (2003). Biological sample collection and processing for molecular epidemiological studies. *Mutat. Res.* 543: 217–234.

Hu, L., Ye, M., et al. (2009). Recent advances in mass spectrometry-based peptidome analysis. *Expert Rev. Proteomics* 6: 433–447.

Huang, G., Song, Y., et al. (2003). [The bystander effect of HSV-tk/GCV system on human cervical carcinoma cell line ME180]. *Sichuan Da Xue Xue Bao Yi Xue Ban* 34: 51–54.

Hutcheson, J. C., Canning, D. A., et al. (2002). Magnetic resonance imaging of fetal urinoma. *Urology* 60: 697.

Jahnukainen, T., Malehorn, D., et al. (2006). Proteomic analysis of urine in kidney transplant patients with BK virus nephropathy. *J. Am. Soc. Nephrol.* 17: 3248–3256.

Johnson, W. E., Evans, H., et al. (2001). Immunohistochemical detection of Schwann cells in innervated and vascularized human intervertebral discs. *Spine* 26: 2550–2557.

Kader, H. A., Tchernev, V. T., et al. (2005). Protein microarray analysis of disease activity in pediatric inflammatory bowel disease demonstrates elevated serum PLGF, IL-7, TGF-beta1, and IL-12p40 levels in Crohn's disease and ulcerative colitis patients in remission versus active disease. *Am. J. Gastroenterol.* 100: 414–423.

Kaji, H., Saito, H., et al. (2003). Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat. Biotechnol.* 21: 667–672.

Kanikowska, D., Grzymislawski, M., et al. (2005). Seasonal rhythms of acute phase proteins in humans. *Chronobiol. Int.* 22: 591–596.

Kanikowska, D., Hyun, K. J., et al. (2005). Circadian rhythm of acute phase proteins under the influence of bright/dim light during the daytime. *Chronobiol. Int.* 22: 137–143.

Kurian, E., Kirk, W. R., et al. (1998). Affinity of fatty acid for rRat intestinal fatty acid binding protein: Further examination. *Biochemistry* 37: 6614.

Kurian, S. M., Flechner, S. M., et al. (2005). Laparoscopic donor nephrectomy gene expression profiling reveals upregulation of stress and ischemia associated genes compared to control kidneys. *Transplantation* 80: 1067–1071.

Lancashire, L., Schmid, O., et al. (2005). Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics* 21: 2191–2199.

Leiserowitz, G. S., Lebrilla, C., et al. (2008). Glycomics analysis of serum: A potential new biomarker for ovarian cancer? *Int. J. Gynecol. Cancer* 18: 470–475.

Liang, P. H., Wu, C. Y., et al. (2008). Glycan arrays: Biological and medical applications. *Curr. Opin. Chem. Biol.* 12: 86–92.

Luch, A. (2009). Preface in *Molecular, clinical and environmental toxicology*. Berlin: Department for Product Safety and ZEBET, Federal Institute for Risk Assessment.

Marrero, J. A., and Lok, A. S. (2004). Newer markers for hepatocellular carcinoma. *Gastroenterology* 127(Suppl. 1): S113–S119.

Masaki, Y., and Sugai, S. (2004). Lymphoproliferative disorders in Sjogren's syndrome. *Autoimmun. Rev.* 3: 175–182.

Mei, N., Fuscoe, J. C., et al. (2010). Application of microarray-based analysis of gene expression in the field of toxicogenomics. *Methods Mol. Biol.* 597: 227–241.

Mikkelsen, J. D., Thomsen, M. S., et al. (2010). Use of biomarkers in the discovery of novel anti-schizophrenia drugs. *Drug Discov. Today* 15: 137–141.

Mootha, V. K., Bunkenborg, J., et al. (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 115: 629–640.

Moreira, E. G., Yu, X., et al. (2010). Toxicogenomic profiling in maternal and fetal rodent brains following gestational exposure to chlorpyrifos. *Toxicol. Appl. Pharmacol.* 245: 310–325.

Ness, S. A. (2007). Microarray analysis: basic strategies for successful experiments. *Mol. Biotechnol.* 36: 205–219.

Pagliarini, D. J., Calvo, S. E., et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134: 112–123.

Pai, S. I., Lin, Y. Y., et al. (2006). Prospects of RNA interference therapy for cancer. *Gene Ther*. 13: 464–477.

Pan, C., Jain, A., et al. (2004). Analysis of causes of late mortality in liver transplant recipients. *Chinese Critical Care Med*. 16: 547–551.

Pan, C., Kumar, C., et al. (2009). Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. *Mol. Cell Proteomics* 8: 443–450.

Pan, J., Chen, H. Q., et al. (2008). Comparative proteomic analysis of non-small-cell lung cancer and normal controls using serum label-free quantitative shotgun technology. *Lung* 186: 255–261.

Pan, S., Aebersold, R., et al. (2009). Mass spectrometry based targeted protein quantification: methods and applications. *J. Proteome Res.* 8: 787–797.

Patterson, S. D. (2003). Proteomics: Evolution of the technology. *Biotechniques* 35: 440–444.

Patterson, S. D., and Aebersold, R. H. (2003). Proteomics: The first decade and beyond. *Nat. Genet.* 33(Suppl.): 311–323.

Pender, M. P., Csurhes, P. A., et al. (2000). Surges of increased T cell reactivity to an encephalitogenic region of myelin proteolipid protein occur more often in patients with multiple sclerosis than in healthy subjects. *J. Immunol.* 165: 5322–5331.

Perroud, B., Lee, J., et al. (2006). Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol. Cancer* 5: 64.

Pettit, S., des Etages, S. A., et al. (2010). Current and future applications of toxicogenomics: Results summary of a survey from the HESI Genomics State of Science Subcommittee. *Environ. Health Perspect.* 118: 992–997.

Poole, A. R. (2003). Biochemical/immunochemical biomarkers of osteoarthritis: Utility for prediction of incident or progressive osteoarthritis. *Rheum. Dis. Clin. North Am.* 29: 803–818.

Que, A. H., Mechref, Y., et al. (2003). Coupling capillary electrochromatography with electrospray Fourier transform mass spectrometry for characterizing complex oligosaccharide pools. *Anal. Chem.* 75: 1684–1690.

Ralser, M., Goehler, H., et al. (2005). Generation of a yeast two-hybrid strain suitable for competitive protein binding analysis. *Biotechniques* 39: 165–166, 168.

Rapkiewicz, A. V., Espina, V., et al. (2004). Biomarkers of ovarian tumours. *Eur. J. Cancer* 40: 2604–2612.

Ray, H. N., Mootha, V. K., et al. (2003). Building an application framework for integrative genomics. *AMIA Annu. Symp. Proc.* 981.

Ressine, A., Marko-Varga, G., et al. (2007). Porous silicon protein microarray technology and ultra-/superhydrophobic states for improved bioanalytical readout. *Biotechnol. Annu. Rev.* 13: 149–200.

Ringeissen, S., Connor, S. C., et al. (2003). Potential urinary and plasma biomarkers of peroxisome proliferation in the rat: Identification of N-methylnicotinamide and N-methyl-4-pyridone-3-carboxamide by 1H nuclear magnetic resonance and high performance liquid chromatography. *Biomarkers* 8: 240–271.

Robertson, D. G., Reily, M. D., et al. (2005). Metabonomics in preclinical drug development. *Expert Opin. Drug Metab. Toxicol.* 1: 363–376.

Roden, D. M., Altman, R. B., et al. (2006). Pharmacogenomics: Challenges and opportunities. *Ann. Intern. Med.* 145: 749–757.

Rosai, J. (2003). Immunohistochemical markers of thyroid tumors: Significance and diagnostic applications. *Tumorigenesis* 89: 517–519.

Sadrzadeh, S. M., and Bozorgmehr, J. (2004). Haptoglobin phenotypes in health and disorders. *Am. J. Clin. Pathol.* 121(Suppl.): S97–S104.

Saito, Y., Yokota, T., et al. (2005). Transgenic small interfering RNA halts amyotrophic lateral sclerosis in a mouse model. *J. Biol. Chem*. 280: 42826–42830.

Schadt, E. E., Molony, C., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*. 6: e107.

Schaub, S., Rush, D., et al. (2004a). Proteomic-based detection of urine proteins associated with acute renal allograft rejection. *J. Am. Soc. Nephrol.* 15: 219–227.

Schaub, S., Wilkins, J., et al. (2004b). Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int.* 65: 323–332.

Schaub, S., Wilkins, J. A., et al. (2004). Proteomics in renal transplantation: Opportunities and challenges. *Clin. Transpl*. 253–260.
Schubeler, D. (2009). Epigenomics: Methylation matters. *Nature* 462: 296–297.
Scriver, C. R. (2007). The PAH gene, phenylketonuria, and a paradigm shift. *Hum. Mutat*. 28: 831–845.
Sellers, T. A., and Yates, J. R. (2003). Review of proteomics with applications to genetic epidemiology. *Genet. Epidemiol.* 24: 83–98.
Seshadri, S., Fitzpatrick, A. L., et al. (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA* 303: 1832–1840.
Shaham, O., Wei, R., et al. (2008). Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Mol. Syst. Biol*. 4: 214.
Shevchenko, A., and Simons, K. (2010). Lipidomics: Coming to grips with lipid diversity. *Nat. Rev. Mol. Cell Biol.* 11: 593–598.
Shitrit, D., Zingerman, B., et al. (2005). Diagnostic value of CYFRA 21-1, CEA, CA 19-9, CA 15-3, and CA 125 assays in pleural effusions: Analysis of 116 cases and review of the literature. *Oncologist* 10: 501–507.
Shostak, S. (2005). The emergence of toxicogenomics: A case study of molecularization. *Soc. Stud. Sci.* 35: 367–403.
Smirnov, D. A., Morley, M., et al. (2009). Genetic analysis of radiation-induced changes in human gene expression. *Nature* 459: 587–591.
ter Kuile, B. H., and Westerhoff, H. V. (2001). Transcriptome meets metabolome: Hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* 500: 169–171.
Teunissen, C. E., de Vente, J., et al. (2002). Biochemical markers related to Alzheimer's dementia in serum and cerebrospinal fluid. *Neurobiol. Aging* 23: 485–508.
Thompson, K. (2010). Toxicogenomics and studies of genomic effects of dietary components. *World Rev. Nutr. Diet* 101: 115–122.
Van Aggelen, G., Ankley, G. T., et al. (2010). Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environ. Health Perspect.* 118: 1–5.
van Meer, G. (2005). Cellular lipidomics. *EMBO J.* 24: 3159–3165.
Weetman, A. P. (2004a). Autoimmune thyroid disease. *Autoimmunity* 37: 337–340.
Weetman, A. P. (2004b). Cellular immune responses in autoimmune thyroid disease. *Clin. Endocrinol.* 61: 405–413.
The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
Wenk, M. R. (2005). The emerging field of lipidomics. *Nat. Rev. Drug Discov.* 4: 594–610.
Wild, C. P. (2010). OMICS technologies: An opportunity for two-way translation from basic science to both clinical and population-based research. *Occup. Environ. Med.* 67: 75–76.
Wu, F., Dassopoulos, T., et al. (2007). Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflamm. Bowel Dis*. 13: 807–821.
Yano, H., and Kuroda, S. (2008). Introduction of the disulfide proteome: Application of a technique for the analysis of plant storage proteins as well as allergens. *J. Proteome Res*. 7: 3071–3079.
Yasuda, K. (2010). [On-chip Cellomics technology for drug screening system using cardiomyocyte cells from human stem cell]. *Yakugaku Zasshi* 130: 545–557.
Yates, J. R., 3rd, Gilchrist, A., et al. (2005). Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* 6: 702–714.
Zarbl, H. (2007). Toxicogenomic analyses of genetic susceptibility to mammary gland carcinogenesis in rodents: Implications for human breast cancer. *Breast Dis.* 28: 87–105.
Zhang, Y., Cocklin, R. R., et al. (2003). Rapid determination of advanced glycation end products of proteins using MALDI-TOF-MS and PERL script peptide searching algorithm. *J. Biomol. Tech.* 14: 224–230.

Section I

Methodology and Application

2 Bioinformatics
A Brief Introduction to Changing Trends in Modern Biology

Abhishek Angelo Singh
Maastricht University
Maastricht, the Netherlands

Pallavi Somvanshi
Bioinformatics Centre
Lucknow, India

CONTENTS

2.1 INTRODUCTION

Computers have become ubiquitous in the field of biological science. They serve a range of purposes, including collecting and processing signals detected by DNA sequencers, storing data in public repositories, and annotating data and performing simulations on the stored data. Biological molecules like DNA, RNA, and proteins are information carriers, instructing the system when and where a particular biological process needs to take place. In each human cell, approximately 5,000 different proteins are expressed (Celis et al., 1991); efficient handling of this massive amount of data requires strong computational resources. In the world of computer science, data consist of a numerical value or a value that can be processed by a computer, whereas in the biological world, data consist of raw knowledge and observations and can be seen as pieces of a puzzle that needs to be put in correct perspective for a clear vision. Contributions from the fields of computer science and mathematics have been very significant in expanding our understanding of cell mechanisms at the molecular level. The continuous advancements in computing technologies, supercomputers, and computer clusters are bringing significant changes to modern biology. The past few decades have witnessed steady progress in the field of bioinformatics. The late 1960s witnessed the development of algorithms responsible for the construction of phylogenetic trees and protein sequence alignment (Fitch and Margoliash, 1967; Cantor, 1968); the 1970s witnessed the development of algorithms for secondary structure prediction of RNA and proteins (Tinoco et al., 1971; Chou and Fasman, 1974); the 1980s witnessed development in sequence analysis, protein structure prediction (tertiary), molecular evolution, and database development; while the past decade witnessed development in methods of *in silico* drug designing.

The role of bioinformatics is pivotal in the postgenomic era. Today it is hard to imagine designing a biological experiment without taking into account biological databases. Bioinformatics aids in terms of the application of knowledge to a biological problem and cuts down on cost and time for experimental design. Put simply, bioinformatics provides necessary computational tools and

databases for efficient and successful running of a biological project. Good examples are developments in the field of *omics* involving handling, processing, and analyzing large-scale genomic data arising from various genome-sequencing projects. Bioinformatics precision extends beyond the study of genes and pathways to include the study of drug targets and therapeutic drugs.

The strength of bioinformatics resides in its ability to link diverse research and academic fields such as molecular biology, genetics, biochemistry, clinical genetics, molecular diagnostics, pharmacogenomics, biomedical informatics, mathematics, statistics, informatics, artificial intelligence, physics, chemistry, medicine, and biology, making it an interdisciplinary field (Figure 2.1). Hence, the field of bioinformatics can be seen as a fine amalgamation of various disciplines. Bioinformatics is a dynamic and rapidly developing branch of modern science that has the capability to change the rule of thumb of biology: predictions are not based on general principles (Lake and Moore, 1998; Howard, 2000; Rashidi and Buehler, 2000).

Bioinformatics tools are widely used by the scientific community for a variety of tasks including comparison of biological sequences, establishing ancestral relationship, structure prediction of a biomolecule, primer designing, genome-map construction, restriction-map construction, high-throughput data analysis, pathway analysis, and *in silico* drug designing.

2.2 UNIQUENESS OF BIOINFORMATICS

Successful running of an *in silico* analysis (bioinformatics-based analysis) to decode the language of biomolecules requires modest hardware, and most bioinformatics tools and biological repositories are freely accessible. The results obtained through the *in silico* analysis unarguable cut down in laboratory time and cost. The efficiency of the bioinformatics analysis is increased several-fold when linked to different repositories. Bioinformatics acts as a knowledge bridge in comprehending the consequences of mutations in DNA, revealing gene structure, establishing ancestral relationships, and determining the consequences of a structural disorder. Bioinformatics provides its selfless services (Foster, 2005) to the scientific community, for example, the myGrid e-science project (http://www.mygrid.org.uk) (Hey and Trefethen, 2005).

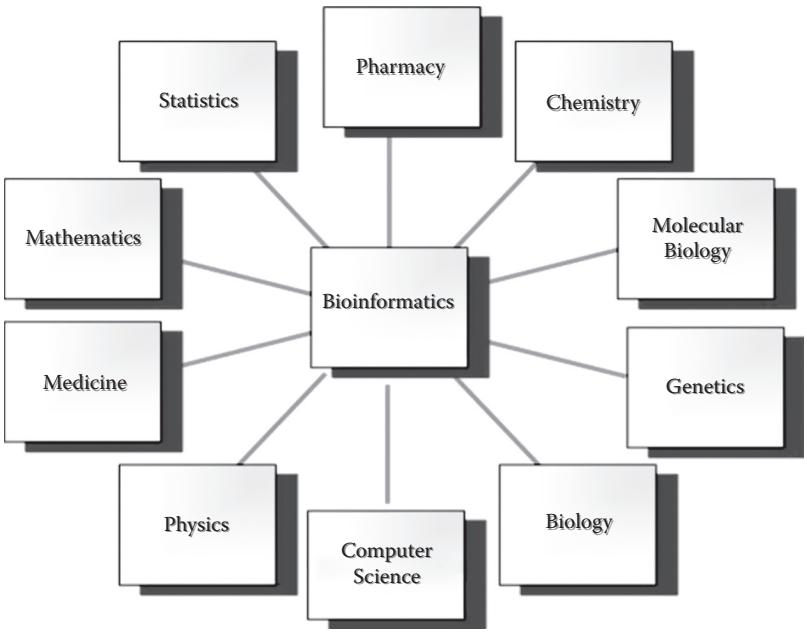

FIGURE 2.1 The strength of bioinformatics resides in its ability to link diverse research and academic fields.

2.3 AIMS OF BIOINFORMATICS

The fundamental aims of bioinformatics are as follows:

1. Construction of biological databases: for example, the Protein Data Bank (PDB) for 3D macromolecular structures (Bernstein et al., 1977; Berman et al., 2000).
2. Development of algorithms for identification of relationships among the members of data set: for example, in order to identify homologous sequence, programs like Basic Local Alignment Search Tool are required (Altschul et al., 1997).
3. Comprehensive analysis and interpretation of biological data.

It is evident that computers are the core part of bioinformatics, but it is necessary that computers are fed with accurate data input.

2.3.1 Biological Databases

Biological databases store and organize large biological data sets. The biological data sets are composed of information from scientific experiments, scientific literature, and analysis on these datasets. These may hold genomics, proteomics, phylogenetic, and gene expression information (Altman, 2004). Biological databases can be mined for information associated with gene function, structure, localization, clinical effects of mutations, and similarity between biological sequences and structures; this information helps us in building and understanding a large number of biological phenomena ranging from the structure and interactions of biomolecules to the evolution of species.

Based on the type of data, biological databases can be grouped into primary, secondary, and specialized databases. Primary databases contain unprocessed biological sequence or structural data submitted by research groups across the world. Examples of primary databases are the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank), DNA Data Bank of Japan (DDBJ), GenBank of the National Center for Biotechnology Information (NCBI), Swiss-Prot, and Protein Data Bank.

Secondary databases are well-curated databases. Although the information content is derived from primary databases, it is of higher quality compared with its source. Examples of secondary databases are PROSITE and PRINTS.

Specialized databases are the tailored databases and serve a particular interest. Examples of specialized databases are the HIV Sequence Database, TRANSFAC for transcription factors, and dbSNP for single nucleotide polymorphism.

Primary databases act as central storing and distribution hubs for raw biological information and are a support system for many biological databases. They can be seen as the first level of information holders and may provide templates for building other databases. Users should be cautious while using databases, because primary databases (data sources) may contain errors, duplication of records, and ambiguity in the presentation of data. Some of the commonly used databases are discussed in this chapter.

2.3.1.1 Protein Data Bank

The Research Collaboratory for Structural Bioinformatics (RCSB) PDB houses information about experimentally validated structures of biomolecules, including protein, nucleic acids, and complex assemblies (Bernstein et al., 1977; Berman et al., 2000). RCSB PDB is also member of worldwide PDB (wwPDB) and follows norms laid down by the consortium in curating and annotating RCSB PDB data (Berman et al., 2003). It also provides tools and resource to support users in performing advanced searches based on sequence, structure, and function. The homepage of RCSB PDB (http://www.pdb.org/pdb/home/home.do) can be mined for more information.

2.3.1.2 Molecular Modeling Database

The Molecular Modeling Database (MMDB) stores experimentally determined structures of proteins and nucleic acids derived from the PDB with some additional features to give more information about the structures (Wang et al., 2007). The added features include chemical graphs, computationally identified 3D domains, links to literature, and links to similar sequences, etc. More details can be seen on the home page of MMDB (http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml).

2.3.1.3 PRINTS

The PRINTS database accumulates conserved protein motifs that act as fingerprint marker to describe a particular protein family (Attwood et al., 2004). Fingerprints can determine protein folds and functionalities comprehensively as compared to single motif. More details can be seen on the home page (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php).

2.3.1.4 CATH

CATH (Orengo et al., 1997; Pearl et al., 2001) uses protein domain information to classify hierarchically protein structures in PDB. The four levels of hierarchy are: class, architecture, topology, and homologous superfamily. Crystal structures having resolution better than 4.0 angstron are considered together with nuclear magnetic resonance (NMR) structures (Orengo et al., 1997, 1998, 1999). These resources can be found on the home page (http://www.cathdb.info).

2.3.1.5 SCOP

The Structural Classification of Protein (SCOP) database comprises protein structures classified according to their evolutionary and structural relationships, i.e., protein domains from all species are classified into families, superfamilies, folds, and classes (Murzin et al., 1995). More information can be accessed from the home page (http://scop.mrc-lmb.cam.ac.uk/scop/).

2.3.1.6 GenBank

The GenBank database has DNA sequences (Benson et al., 2008). It also contains bibliographic and biological annotations. GenBank, along with DDBJ and EMBL, is a member of the International Nucleotide Sequence Database Collaboration. These organizations try to maintain uniformity in their content by exchanging data on a regular basis. More information can be acquired at http://www.ncbi.nlm.nih.gov/genbank.

2.3.1.7 dbSNP

dbSNP houses data on single-nucleotide polymorphisms (SNPs) and helps in identifying genotype-phenotype associations (Sherry et al., 2001). It contains SNP data (location in gene, nucleotides, and affected amino acids) and also information about the organisms from which SNPs are derived, including population details. The data from dbSNP can aid researchers in identifying SNPs in a gene, in determining SNPs that influence phenotype, and in obtaining functional information of gene product. More information can be found at http://www.ncbi.nlm.nih.gov/projects/SNP.

2.3.1.8 KEGG

Kyoto Encyclopedia of Genes and Genomes, commonly known as KEGG, is an integrated database resource of 16 databases roughly grouped into systems information, genomic information, and chemical information (Kanehisa et al., 2006). It contains data on genome sequences, metabolic pathways, orthologs, and compound structures. KEGG aids in providing biological understanding of large-scale datasets generated by high-throughput experimental technologies. More information can be obtained at http://www.genome.jp/kegg.

2.3.1.9 TRANSFAC

TRANSFAC is a database of transcription factors along with their experimentally verified binding sites and regulated genes (Matys et al., 2006). Academic and nonprofit organizations have free access to these data. More information can be obtained at the TRANSFAC home page (http://www.biobase-international.com/pages/index.php?id=transfac).

2.3.1.10 OMIM

The Online Mendelian Inheritance in Man (OMIM) is a database of genes and genetic disorders in humans (Hamosh et al., 2002). It is focused on inherited or heritable genetic diseases. It acts as a phenotypic resource to the human genome project. A lot of information is available at http://www.ncbi.nlm.nih.gov/omim.

2.3.2 Algorithms and Biology

Algorithms are a set of instructions, implemented using computers to perform complex tasks in bioinformatics analysis. The range of tasks that algorithms perform in bioinformatics includes comparative analysis of sequences, building phylogenetic trees, predicting gene and protein structure, visualization of structures, probing databases, simulation of biological processes, and image processing. The efficiency of any bioinformatics analysis depends upon the grade of algorithm being used.

The field of bioinformatics has not yet reached its adolescence, as most of the algorithms being used to solve confounding problems in modern biology are not comprehensive enough to reflect a clear picture of the biological system in question. Furthermore, the algorithms that intend on producing a good-enough picture of *in vivo* processes have very high computational demands.

There is a huge demand for developing algorithms that may give a comprehensive picture of *in vivo* biological processes. However, the biggest challenge for algorithm developers is to find a middle road, i.e., an algorithm that is computationally less demanding and at the same time comprehensive enough. Researchers are striving to overcome these hurdles posed by the complexities of biological systems.

2.3.3 Analysis and Interpretation of Experimental Data

The combination of biological data with a statistical or mathematical model illustrates a more comprehensive picture of cellular activities. A strong statistical model is important for characterization of various analyses measured in an experiment. Integration of various biological data resources significantly contributes to the accuracy of analysis and can provide insights into complex diseases and traits, for example:

1. Integrating nonsynonymous single-nucleotide polymorphism data with the native structure of protein can help in identifying deleterious nonsynonymous single-nucleotide polymorphisms (Singh et al., 2011a).
2. In molecular biomarker studies, the proficient integration of molecular, clinical, and imaging data may aid in reaching a conclusive result.

The projects in biological sciences vary in their demand for analysis and integration of resources; hence statistical models and analysis are customized according to the need of the project. The correct analysis and interpretation of results is also important for successfully running a wet lab protocol. In this area, researchers from the bioinformatics discipline can provide aid by integrating multiple sources of biological knowledge (databases and tools), which in turn helps in correct interpretation and elucidation of biological processes (Gerstein, 2000).

2.4 BIOINFORMATICS AND RESEARCH AREAS IN MODERN BIOLOGY

Bioinformatics methods are indispensable in biological science. Below we describe a few of the major research areas in bioinformatics; a review of all research areas is out of the scope of this chapter.

2.4.1 Genomics

The successful completion of large numbers of genome-sequencing projects has resulted in an ocean of genomic data. The precise analysis of these data for the correct identification of the gene, its location, genetic control elements, tagging structural and functional information, and identification of genetic variation(s) is the greatest challenge to the scientific community. Bioinformatics helps in the organization and storage of high-throughput data and provides tools for analyzing these data. It was through the availability of computational tools and databases that the field of comparative genomics has made progress in leaps and bounds. The analysis of a single sequence reveals lots of information, but there are still certain regions in genomes that remain unexplored, and the functionality of these regions can only be explored by comparing them with other genomes that are less divergent or homologous in nature (Kellis et al., 2003). The availability of freely accessible biological databases and computational tools for the mining and analysis of genomes and their genes has made it possible to answer a few of the many confounding questions of modern biology. Biological databases such as dbSNP (Sherry et al., 2001) can be mined for genetic variations of a given gene, whereas TRANSFAC (Matys et al., 2006) is for transcription factors, and a computational tool like Sorting Intolerant from Tolerant (SIFT) (Ng and Henikoff, 2001, 2002, 2003; Kumar et al., 2009) can be used for predicting deleterious SNPs. Some web servers such as SNPnexus (Chelala et al., 2009) are completely dedicated to the analysis of mutations in the regulatory regions of genes, and Regulatory Sequence Analysis Tools (RSAT), available at http://rsat.ulb.ac.be, provides several computational tools for detecting regulatory signals in genomes (Turatsinze et al., 2008). These are some of the prominent bioinformatics tools used in the field of genomics. Ongoing projects for genome annotation include ENCODE, the ENCyclopedia of DNA Elements (http://www.genome.gov/10005107); Entrez Gene (http://www.ncbi.nlm.nih.gov/gene); Ensembl (http://www.ensembl.org/index.html); the Gene Ontology Consortium (http://www.geneontology.org); RefSeq (http://www.ncbi.nlm.nih.gov/projects/RefSeq); and the Vertebrate and Genome Annotation Project (http://vega.sanger.ac.uk/index.html).

2.4.2 Proteomics

Proteomics refers to a comprehensive study of proteins, particularly their structures, functions, and interactions (Anderson and Anderson, 1998). Structural and functional information about the proteins is vital in elucidating their role in the regulation of cellular processes and in biological pathways (Morel et al., 2004). Bioinformatics provides a pool of databases and computational tools for studies related to proteomics, such as for protein expression analysis. Bioinformatics techniques can be employed to match a large amount of data against predicted data from protein sequence databases followed by statistical analysis. Examples of bioinformatics tools and databases used in protein expression analysis are as follows:

1. ExPASy 2D PAGE databases and services (http://expasy.org/ch2d/2d-index.html) (Hoogland et al., 1999), which contain references to known 2D PAGE database servers and to 2D PAGE-related services
2. GelScape (http://www.gelscape.ualberta.ca) (Young et al., 2004), a platform-independent program for analyzing standard 1D and 2D protein gels
3. NCI Flicker (http://www-lmmb.ncifcrf.gov/flicker) (Lemkin et al., 1979), which compares 2D sample gels against 2D gel database maps, possibly suggesting putative protein spot identification

Bioinformatics methods help in providing a snapshot of proteins present in biological samples. Also in protein structure prediction, correct determination of secondary structure is important because it governs the shape acquired by the polypeptide chain. Some of the commonly used methods for secondary structure prediction are the Chou and Fasman method (Chou and Fasman 1974), the GOR method (Garnier et al., 1978), and the PHD program (Rost and Sander, 1993a, 1993b). Tertiary or native structure of protein is responsible for its biological activity. Hence, a lot of focus has been on developing tools or protocols for correct or approximate prediction of tertiary structure. Computationally predicted structures serve as templates to build hypotheses for further research in structural and functional biology. The gold standards in structure prediction are set by NMR and X-ray crystallography methods.

2.4.3 Transcriptomics

Only 5% of human DNA is transcribed; the rest of the genome controls and regulates expression of this 5% of the genome. The expression information of individual genes at the mRNA level can help in establishing the relationship between gene expression patterns and its biological significance (Carulli et al., 1998; Scheel et al., 2002; Morel et al., 2004). Bioinformatics can be used for mining of the transcriptomics data followed by its analysis, this could help in creating an understanding of how certain genes are activated and deactivated and how the levels of gene products produced. Understanding transcriptomics data and its analysis is particularly important to researchers studying the process of cellular differentiation and carcinogenesis.

2.4.4 Metabolomics

Metabolomics refers in particular to the study of metabolite profiles, the study of chemical fingerprints left after a completion of cellular processes (Daviss, 2005). Metabolic profiling is important because it can give us a snapshot of the cellular physiology that can augment the information about the cellular physiology that mRNA gene expression data and proteomic analysis does not reveal. The field of bioinformatics with its data management, data processing, statistical analysis, data mining, data integration, and the mathematical modeling of metabolic networks' abilities can assist in the development of metabolomics (Shulaev, 2006).

2.4.5 Cytomics

Cytomics refers to single cell study of cell-system heterogeneity using image or flow cytometry (Gomase and Tagore, 2008a). By studying cell phenotypes, we can establish a correlation between the disease process as the sum of the respective genotype and exposure influences. The cell phenotype contains the information about the cell health (disease status) and helps in predicting therapy-dependent future developments. The integration of information from cytomics with proteomics may assist in identifying cells with a specific set of phenotype characteristics; this may aid in the identification of tumor markers (Bernas et al., 2006). Bioinformatics tools can be exhaustively used in extracting molecular cell phenotype(s).

2.4.6 Physiomics

Physiomics refers to an integrative study of genome, proteome, and metabolome (Gomase and Tagore, 2008b). It comprehensively uses experimental databases and computer algorithms for correct identification of physiological phenotypes of genes and proteins. Overall, bioinformatics tools and databases may aid the field of physiomics in constructing physiological features associated with genes, proteins, and interactions among and between them. The field of physiomics is extremely useful in the development of drugs and biochips (Gomase and Tagore, 2008b).

2.4.7 Glycomics

Glycomics refers to a comprehensive study of the entire sugar compliment of an organism, encompassing its genetic, pathologic, physiologic, and other aspects (Aoki-Kinoshita, 2008). The integration of glycomics data with proteomic and genomics data can help in elucidating the relationship between the glycome and genome (Pilobello and Mahal, 2007). This in turn will help in determining the role of carbohydrates or sugar moloecules in various pathways.

2.4.8 Lipidomics

In biological systems, lipidomics refers to comprehensive study of the pathways and networks of cellular lipids (Wenk, 2005; Watson, 2006). Lipids are important because they perform structural, energy storage, and signaling roles (Wenk, 2005). Lipidomics, along with the data integration and data analysis capabilities of bioinformatics, can help elucidate important biological phenomena and their influence(s) on biological pathways.

2.4.9 Interactomics

Interactomics refers to the study of interactions between and among protein(s) and other biomolecules inside cells (Kiemer and Cesareni, 2007). A network of such interactions is called an *interactome*. Comparison of interactomes among and between species may help reveal the traits of such networks.

2.5 FREQUENTLY USED BIOINFORMATICS TOOLS

A list of commonly used bioinformatics tools, which can be accessed freely over the World Wide Web, is presented. A comprehensive list of bioinformatics tools used in modern biology is beyond the scope of this chapter.

2.5.1 Electronic Polymerase Chain Reaction

Electronic polymerase chain reaction (e-PCR) aids in the identification of sequence-tagged sites within DNA sequences (Schuler, 1997). The e-PCR, aids in searching subsequences that closely match PCR primers and have the right order, orientation, and spacing. The program employs a fuzzy matching strategy that improves search sensitivity of the program and also allows incorporating gaps in primer alignment. The latest release of e-PCR provides a search mode using a query sequence against a sequence database (Rotmistrovsky et al., 2004). More information about this program can be accessed on http://www.ncbi.nlm.nih.gov/Tools, and the program can be accessed through http://www.ncbi.nlm.nih.gov/projects/e-pcr. Researchers have used the *in silico* designing of primers and validated using PCR. For example, *in silico* primers have been designed for the amplification of all the structural regions of the quasi species of dengue viruses that could be useful in molecular diagnostic (Somvanshi and Seth, 2008). Using this technique, numbers of *in silico* primers have been designed, synthesized, and validated in wet laboratories for the detection of other bacterial pathogens, especially *Aeromonas hydrophila* (Singh et al., 2009; Singh et al., 2010b; Singh et al., 2011a; Singh and Somvanshi, 2009c).

2.5.2 Map Viewer

Map Viewer displays integrated chromosomal maps of many organisms including vertebrates, invertebrates, fungi, protozoa, and plants (Wheeler et al., 2008). The tool aids in locating genes along with other biological features. More information about the program can be found at http://

www.ncbi.nlm.nih.gov/Tools, and the program can be accessed through http://www.ncbi.nlm.nih.gov/mapview.

2.5.3 Model Maker

Through Model Maker, the user can view the evidence that was used to construct a gene model on an assembled genomic sequence, and it can be accessed from sequence maps (Wheeler et al., 2008). Moreover, Model Maker enables the user to construct his version of the model by selecting desired exons.

2.5.4 Open Reading Frame Finder

Open Reading Frame (ORF) Finder detects all open reading frames of selected minimum size in a query sequence (Wheeler et al., 2008). The program detects open reading frames using standard or alternative genetic codes. ORF Finder may assist in preparing complete and precise sequence submissions and is packaged in Sequin (http://www.ncbi.nlm.nih.gov/projects/Sequin), a sequence submission software. The program can be accessed at http://www.ncbi.nlm.nih.gov/projects/gorf. It has already been used to find the ORF that could be used for the expression of genes (Singh et al., 2009, 2011a; Singh and Somvanshi, 2009c).

2.5.5 TaxPlot

TaxPlot is a three-way comparison of genomes based on the protein encoded by them (Wheeler et al., 2008). The tool selects a reference genome, compares two other genomes to it, and uses pre-computed BLAST results to plot a point for each predicted protein in reference genome. The program can be accessed at http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi.

2.5.6 VAST Search

VAST Search is a structure similarity search program; compares 3D coordinates of a recently ascertained protein structure with the existing ones in the MMDB/PDB databases (Wheeler et al., 2008). The program computes and generates a list of structure neighbors that could be browsed interactively. The program can be accessed at http://structure.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html.

2.5.7 Basic Local Alignment Search Tool

Basic Local Alignment Search Tool (BLAST) is a set of similarity searching programs intended to search all of the accessible sequence databases regardless of whether the query is protein or DNA; it seeks local alignments and identifies relationships between sequences sharing isolated regions of similarity (Altschul et al., 1990, 1997). Variants of BLAST are as follows:

1. BLASTn: Input nucleotide sequence is compared with nucleotide sequence database to find sequences containing regions homologous to input sequence.
2. BLASTp: Input protein sequence is compared with protein sequence database to find sequences containing regions homologous to input sequence.
3. BLASTx: Input nucleotide sequence is translated and compared with protein sequence database to find sequences containing regions homologous to input sequence.
4. tBLASTn: Input protein sequence is compared with translated nucleotide sequence database to find sequences containing regions homologous to input sequence.
5. tBLASTx: Input nucleotide sequence is translated and compared with translated nucleotide sequence database to find sequences containing regions homologous to input sequence.

Several applications of this tool are used to find the homology of genes and proteins present in available organisms such as influenza virus (Somvanshi et al., 2008a), in genes (Somvanshi et al., 2008b), hemolysin (Singh et al., 2009), and aerolysin (Singh and Somvanshi, 2009c; Singh et al., 2010b). BLAST can be accessed through NCBI (Johnson et al., 2000) at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

2.5.8 SIFT

SIFT is used for identifying potential deleterious nonsynonymous single-nucleotide polymorphism(s) (nsSNPs) (Ng and Henikoff 2001, 2002, 2003; Kumar et al., 2009). SIFT uses sequence homology to predict deleterious nsSNPs; amino acids at specific positions that are important for protein function must remain conserved in the alignment of homologous sequences. The tool (and more information about it) can be accessed at http://sift.jcvi.org.

2.5.9 Clustal

Clustal is a multiple sequence alignment program and is available in command line (ClustalW) as well as in graphical interface (ClustalX) (Chenna et al., 2003; Larkin et al., 2007). The program is written in C++ programming language. The program can be executed either on default parameters or on customized parameters; the main parameters of the program that could be adjusted are gap-opening penalty and gap-extension penalty. Precompiled executables for most of the operating systems can be downloaded from http://www.clustal.org.

2.5.10 PHYLIP

PHYLogeny Inference Package, commonly known as PHYLIP (Felsenstein, 1989), is a freely available computational package of programs for inferring phylogenies. The package consists of 35 portable programs (http://bioweb2.pasteur.fr/docs/phylip/doc/main.html). The source code is written in C programming language, and precompiled executables are available for Windows, Mac OS, and Linux systems and can be downloaded from http://evolution.genetics.washington.edu/phylip.html.

2.5.11 FoldX

FoldX program provides a quick and quantitative estimate of the effect of mutation on the stability of proteins. The program exploits atomic description in the structure of proteins (Schymkowitz et al., 2005). The FoldX program (and more information about it) can be obtained from http://foldx.crg.es.

2.5.12 MODELLER

MODELLER is commonly used for homology modeling of the tertiary structure of proteins (Eswar et al., 2006; Marti-Renom, 2000). Target and template sequence alignment is fed as an input into the program, and then it automatically computes a model containing all of the nonhydrogen bond atoms. Other tasks performed by MODELLER include: *de novo* modeling of loop structure, optimization of protein model, comparison of protein structures, etc. The strength of MODELLER lies in its ability to model structures by satisfying spatial constraints (Sali and Blundell, 1993; Fiser et al., 2000). MODELLER has been frequently used to generate the 3D structure of uncrystallized protein from different origins such as bacterial, viral, human, and so on (Somvanshi and Singh, 2008, 2010; Singh et al., 2009; Singh and Somvanshi, 2009a, 2009b,

2009c, 2009d, 2009e, 2010). The tool (and more information about it) can be obtained from http://www.salilab.org/modeller.

2.5.13 GROMACS

Groningen Machine for Chemical Simulations (GROMACS) is a package written in ANSI C programming language and performs molecular dynamics simulation. The program does not have its own force field, but it is compatible with the following force fields: GROMOS, OPLS, AMBER, and ENCAD (Van Der Spoel et al., 2005). The program offers flexibility to customize force routines and tabulated functions. The program can be downloaded from http://www.gromacs.org.

2.5.14 AutoDock

AutoDock (Goodsell et al., 1996; Morris et al., 1998, 2009) is a cluster of automated docking tools. It assists in understanding the binding of a ligand molecule to a receptor of known three-dimensional structure. The docking tool consists of two main programs:

1. AutoDock, which performs docking of ligand molecule to grid(s) characterizing target protein
2. AutoGrid, which computes the grid(s) significant for docking

AutoDock can assist organic synthetic chemists in designing better binders. Some of the application areas of AutoDock are structure-based drug design, lead optimization, virtual screening, protein-protein docking, and chemical mechanism study. This tool became very popular, and researchers are continuously using it for screening drug molecules on the active region of 3D proteins (Somvanshi and Singh, 2008; Singh and Somvanshi, 2009a, 2009b, 2009c, 2009d, 2009e). The tool (and more information about it) can be obtained from http://autodock.scripps.edu/downloads.

2.5.15 Gene Designer for Gene Designing and Codon Optimization

Gene Designer (https://www.dna20.com/index.php?pageID=220) is commonly used for designing genes in a given expression host. Codon optimization technique is used to improve the protein expression in living organisms by increasing the translational efficiency of a gene of particular interest (Mani et al., 2010). The nucleotides of a DNA sequence are separated into triplets (codons), and then the codon is replaced with a new (degenerate codon), generated with a given frequency distribution. This amino acid will be the same, but codon with the low frequency of an amino acid will be replaced with a codon of high frequency, according to the desired species frequency distribution. Optimizer software is used for optimization and calculation of CAI, G+C, and A+T (Puigbo et al., 2007). CAIcal and MrGene are used for optimization of DNA sequences at maximum suitable threshold level (Sharp and Li, 1987). It has become popular to improve the level of expression in the host (Singh et al., 2010a; Mani et al., 2010, 2011). It is also useful for overexpression in the host and provides new insights into emerging research in synthetic biology.

2.6 PRACTICAL APPLICATIONS OF BIOINFORMATICS

The field of bioinformatics is very vast and consequentially has a vast area of application. Below we have listed some of the practical applications of bioinformatics.

2.6.1 Identification of Drug Targets

A comprehensive understanding of disease mechanisms and an efficient use of computational tools can lead to the identification and validation of potential novel drug targets. This will lead to development of more specific medicine(s) that act on cause and have fewer side effects.

2.6.2 Personalized Medicine

Developments in the field of pharmacogenomics will lead to a smooth transition of clinical medicines into personalized medicines. Knowledge about an individual genetic profile can help in prescribing the best possible drug therapy and dosage.

2.6.3 Preventive Medicine

Advances in the field of genomics have unraveled many specific details about the genetic basis of a disease. Preventive measures such as change in lifestyle or medication at an early stage could shield an individual from many diseases.

2.6.4 Gene Therapy

Comprehensive knowledge and understanding of gene expression can aid us in targeting and manipulating expression of genes that lead to disease. This will lead to treatment, cure, or even prevention of disease. Simulation programs can be run in order to find the most probable gene to be targeted for manipulation. Knowledge from simulated gene expression could be obtained before implementation of any manipulation in a real biological system.

2.6.5 Evolutionary Study

Over time biologists have been aided by bioinformatics in studying various aspects of evolution, including:

1. Measuring changes in DNA during the evolution of a certain organism
2. Studying complex evolutionary events such as speciation and gene duplication
3. Building computational models to study a population and making predictions for the population

2.6.6 Agriculture

The agricultural community has benefited vastly from genome sequencing programs for plants. With the aid of bioinformatics tools, specific genes can be targeted in genomes to produce insect- and disease-resistant plants with healthier more productive offspring.

2.6.7 Veterinary Science

The information revealed through sequencing projects of farm animals such as, cattle, pigs, and sheep has helped veterinary scientists immensely in improving the production and health of livestock. More details about the practical applications of bioinformatics can be found at http://www.ebi.ac.uk/2can/bioinformatics/bioinf_realworld_1.html.

2.7 LIMITATIONS

Bioinformatics analysis greatly depends upon the quality and quantity of data generated by the biologist(s) and the comprehensiveness of the algorithm being used to analyze data. The field of bioinformatics is still in its infancy, and many algorithms are in the developmental stage and thus are not comprehensive enough to reflect the complete *in vivo* picture of a biological system.

2.8 WEB RESOURCE FOR BIOINFORMATICS

Through the internet, it is possible to access any database and any computational tool (most of the databases and tools are freely accessible) needed to perform *in silico* analysis. Table 2.1 provides

TABLE 2.1
Web Servers Used in Bioinformatics and Computational Biology Research

Name	Description	Web Link
	Important bioinformatics web resource locators	
ArrayExpress	Public curated repository of array data such as gene expression, comparative genome hybridization data, and chip-on-chip data	http://www.ebi.ac.uk/arrayexpress
ExPASy	Dedicated to *in silico* protein study	http://expasy.org
miRBase	Resource of microRNA	http://www.mirbase.org
RNA World	Lists web links on related RNA topics	http://www.imb-jena.de/RNA.html
Pasteur	Miscellaneous links	http://bioweb.pasteur.fr/intro-en.html
	Important biological databases	
GenBank/DDBJ/EMBL	Nucleotide sequences	http://www.ncbi.nlm.nih.gov/nuccore
PDB	Protein structures	http://www.pdb.org/pdb/home/home.do
Swiss-Prot	Protein sequences	http://expasy.org/people/swissprot.html
OMIM	Genetic diseases and genetic disorders	http://www.ncbi.nlm.nih.gov/omim
KEGG	Biological pathways	http://www.genome.jp/kegg
dbSNP	Single nucleotide polymorphism	http://www.ncbi.nlm.nih.gov/projects/SNP
PubMed	Literature references	http://www.ncbi.nlm.nih.gov/pubmed
	Important bioinformatics tools	
BLAST	Homology search	http://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA	Homology search	http://www.ebi.ac.uk/Tools/sss
Entrez	Database search	http://www.ncbi.nlm.nih.gov/gene
Clustalw2	Multiple sequence alignment	http://www.ebi.ac.uk/Tools/msa/clustalw2
GenScan	Gene structure prediction	http://genes.mit.edu/GENSCAN.html
PredictProtein	Protein structure prediction	http://www.predictprotein.org
Mfold	RNA structure prediction	http://mfold.rna.albany.edu/?q=mfold
PHYLIP	Phylogeny inference package	http://www.phylip.com
RasMol	Structure visualization	http://www.umass.edu/microbio/rasmol

a list of major resource locators for bioinformatics, important biological databases, and important software used in the field. Although this is not a comprehensive list, and the use of databases and tools varies from analysis to analysis, the databases and tools listed may provide a starting point to those entering the field of bioinformatics.

REFERENCES

Altman, R. B. (2004). Building successful biological databases. *Briefings in Bioinformatics*, 4–5.

Altschul, S. F., Gish, W., Miller, W., et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 3389–3402.

Anderson, N. L., Anderson, N. G. (1998). Proteome and proteomics: New technologies, new concepts, and new words. *Eloectrophoresis*, 1853–1861.

Aoki-Kinoshita, K. F. (2008). An introduction to bioinformatics for glycomics research. *PLOS Computational Biology*, e1000075.

Attwood, T. K., Bradley, P., Gaulton, A., et al. (2004). *The PRINTS protein fingerprint database: Functional and evolutionary applications.* New York: John Wiley & Sons.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., et al. (2008). GenBank. *Nucleic Acids Research,* D26–D31.

Berman, H. M., Henrick, K., Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, 98.

Berman, H. M., Westbrook, J., Febg, Z., et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, 235–242.

Bernas, T., Gregori, G., Asem, E. K., et al. (2006). Integrating cytomics and proteomics. *Molecular Cell Proteomics*, 2–13.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., et al. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 535–542.

Cantor, C. R. (1968). The occurrence of gaps in protein sequences. *Biochemical and Biophysical Research Communications*, 410–416.

Carulli, J. P., Artinger, M., Swain, P. M., et al. (1998) High throughput analysis of differential gene expression. *Journal of Cell and Biochemistry Supplement*, 30–31.

Celis, J. E., Leffers, H., Rasmussen, H. H., et al. (1991). The master two-dimensional gel database of human AMA cell proteins: Towards linking protein and genome sequence and mapping information. *Electrophoresis*, 765–770.

Chelala, C., Khan, A., Lemoine, N. R. (2009). SNPnexus: A web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 655–661.

Chenna, R., Sugawara, H., Koike, T., et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, 3497–3500.

Chou, P. Y., Fasman, G. D. (1974). Prediction of Protein Conformation. *Biochemistry*, 222–245.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, New York: D. Appleton and Company.

Daviss, B. (2005). Growing pains for metabolomics. *The Scientist*, 25.

Eswar, N., Marti-Renom, M. A., Webb, B., et al. (2006). Comparative protein structure modeling with MODELLER. *Current Protocols in Bioinformatics*, 5.6.1–5.6.30.

Felsenstein, J. (1989). PHYLIP: PHYLogeny Inference Package. *Cladistics*, 164–166.

Fiser, A., Do, R. K., Sali, A. (2000). Modeling of loops in protein structures. *Protein Science*, 1753–1773.

Fitch, W. M., Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 270–284.

Foster, I. (2005). Service-oriented science. *Science*, 814–817.

Garnier, J., Osguthorpe, D. J., Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 97–120.

Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nature Structural Biology*, 960–963.

Gomase, V. S., Tagore, S. (2008a). Cytomics. *Current Drug Metabolism*, 263–266.

Gomase, V. S., Tagore, S. (2008b). Physiomics. *Current Drug Metabolism*, 259–262.

Goodsell, D. S., Morris, G. M., Olson, A. J. (1996). Automated docking of flexible ligands: Applications of AutoDock. *Journal of Molecular Recognition*, 1–5.

Hamosh, A., Scott, A. F., Amberger, J., et al. (2002). Online Mendelian Inheritance in Man (OMIM): A knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 52–55.

Hey, T., Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, 817–821.

Hoogland, C., Sanchez, J. C., Tonella, L., et al. (1999). The SWISS-2DPAGE database: What has changed during the last year. *Nucleic Acids Research*, 289–291.

Howard, K. (2000). The bioinformatics gold rush. *Scientific American*, 58–63.

Johnson, M. Zaretskaya, I. Raytselis, Y., et al. (2000). NCBI BLAST: A better web interface. *Nucleic Acids Research*, W5–W9.

Kanehisa, M., Goto, S., Hattori, M., et al. (2006). From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Research*, D354–D357.

Kellis, M., Patterson, N., Endrizzi, M., et al. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 241–254.

Kiemer, L., Cesareni, G. (2007). Comparative interactomics: Comparing apples and pears? *Trends in Biotechnology*, 448–454.

Kumar, P., Henikoff, S., Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocol*, 1073–1082.

Lake, J. A., Moore, J. E. (1998). Phylogenetic analysis and comparative genomics. *Trends in Biotechnology*, 22–23.

Larkin, M. A., Blackshields, G., Brown, N. P., et al. (2007). ClustalW and ClustalX version 2.0. *Bioinformatics*, 2947–2948.

Lemkin, P., Merril, C., Lipkin, L., et al. (1979). Software aids for the analysis of 2D gel electrophoresis images. *Computer and Biomedical Research*, 517–544.