MICROARRAY IMAGE AND DATA ANALYSIS

THEORY AND PRACTICE
Digital Imaging and Computer Vision Series

Series Editor

Rastislav Lukac
Foveon, Inc./Sigma Corporation
San Jose, California, U.S.A.

Microarray Image and Data Analysis: Theory and Practice, by Luis Rueda
Perceptual Digital Imaging: Methods and Applications, by Rastislav Lukac
Image Restoration: Fundamentals and Advances, by Bahadir Kursat Gunturk and Xin Li
Image Processing and Analysis with Graphs: Theory and Practice, by Olivier Lézoray and Leo Grady
Visual Cryptography and Secret Image Sharing, by Stelvio Cimato and Ching-Nung Yang
Digital Imaging for Cultural Heritage Preservation: Analysis, Restoration, and Reconstruction of Ancient Artworks, by Filippo Stanco, Sebastiano Battiato, and Giovanni Gallo
Computational Photography: Methods and Applications, by Rastislav Lukac
Super-Resolution Imaging, by Peyman Milanfar
MICROARRAY IMAGE
AND
DATA ANALYSIS
THEORY AND PRACTICE
EDITED BY
Luis Rueda
Dedication

I would like to dedicate this book to my wife, Ivanna, and my children, Maria, Gabriel and Rocio, for their constant love and support.
Contents

Preface.................................................................................................................................................. ix
Editor................................................................................................................................................... xiii
Contributors......................................................................................................................................... xv

Chapter 1  Introduction to Microarrays............................................................................................. 1
    Luis Rueda and Adnan Ali

Chapter 2  Biological Aspects: Types and Applications of Microarrays ..................................... 41
    Adnan Ali

Chapter 3  Gridding Methods for DNA Microarray Images............................................................. 77
    Iman Rezaeian and Luis Rueda

Chapter 4  Machine Learning-Based DNA Microarray Image Gridding..................................... 109
    Dimitris Bariamis, Michalis Savelonas, and Dimitris Maroulis

Chapter 5  Non-Statistical Segmentation Methods for DNA Microarray Images........................... 129
    Shahram Shirani

Chapter 6  Statistical Segmentation Methods for DNA Microarray Images................................... 149
    Meng-Yuan Tsai, Tai-Been Chen, and Henry Horng-Shing Lu

Chapter 7  Microarray Image Restoration and Noise Filtering...................................................... 171
    Rastislav Lukac

Chapter 8  Compression of DNA Microarray Images................................................................. 195
    Miguel Hernández-Cabrero, Michael W. Marcellin, and Joan Serra-Sagristà
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Image Processing of Affymetrix Microarrays</td>
<td>227</td>
</tr>
<tr>
<td></td>
<td><em>Jose Manuel Arteaga-Salas</em></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Treatment of Noise and Artifacts in Affymetrix Arrays</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td><em>Caroline C. Friedel</em></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Quality Control and Analysis Algorithms for Tissue Microarrays as Biomarker Validation Tools</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td><em>Todd H. Stokes, Sonal Kothari, Chih-wen Cheng, and May D. Wang</em></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CNV-Interactome-Transcriptome Integration to Detect Driver Genes in Cancerology</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td><em>Maxime Garcia, Raphaèle Millat-Carus, François Bertucci, Pascal Finetti, Arnaud Guille, José Adélaide, Ismahane Bekhouche, Renaud Sabatier, Max Chaffanet, Daniel Birnbaum, and Ghislain Bidaut</em></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Mining Gene-Sample-Time Microarray Data</td>
<td>339</td>
</tr>
<tr>
<td></td>
<td><em>Yifeng Li and Alioune Ngom</em></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Systematic and Stochastic Biclustering Algorithms for Microarray Data</td>
<td>369</td>
</tr>
<tr>
<td></td>
<td><em>Wassim Ayadi, Mourad Elloumi, and Jin-Kao Hao</em></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Reconstruction of Regulatory Networks from Microarray Data</td>
<td>401</td>
</tr>
<tr>
<td></td>
<td><em>Yiqian Zhou, Rehman Qureshi, Francis Bell, and Ahmet Sacan</em></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Multidimensional Visualization of Microarray Data</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td><em>Urška Cvek and Marjan Trutschl</em></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Bioconductor Tools for Microarray Data Analysis</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td><em>Simon Cockell, Matthew Bashton, and Colin S. Gillespie</em></td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>481</td>
</tr>
</tbody>
</table>
Preface

Since their discovery, microarrays have emerged as one of the most important biotechnological tools for studying the behavior of thousands of genes and genomic features on a genome-wide scale. The development of accurate and affordable technologies for effective and quick detection becomes paramount when considering the cost and time allowed to conduct a biological experiment. This has influenced the way in which transcriptional, proteomic and cell developmental studies are being carried out, providing a systematic way to study gene expression, elucidation of the proteome, discovering binding sites and methylation studies, all of these in a synchronized way. Transcriptional studies are one of the most important applications of microarrays, and for this, the development of gene expression arrays has helped enormously. Measuring the transcriptome or the entire repertoire of transcripts in an organism or cell lines provides a wealth of data about the relationships among DNA, transcription, translation, and phenotype. Advances in the microarray field have also allowed accurate detection that can uncover specific genetic mutations that may result in a number of human disorders. In this regard, microarrays, whether on full-length genes or short oligonucleotide assays, constitute an efficient tool for parallel and systematic analysis.

Although different shapes, resolutions, and features of microarrays may vary substantially across platforms, techniques and applications, they can be defined in general as a small glass slide with a collection of spots or elements in which biochemical reactions take place in order to reveal quantitative information about gene or protein expression, or other biological interaction of the molecules, under specific conditions. A microarray can be conceived as a small device (Greek word “mikro” for small), in which biomolecules are arranged in a systematic manner (French word “arayer” for array). In practice, a microarray is a device that has a rectangular shape and measures a few cm long – it is typically ordered, microscopic, planar, and specific.

Microarrays have been widely used by many researchers in diverse studies in biology. To mention a few facts, there have been more than 50,000 publications pertaining to microarrays, if PubMed is used as the source of information. Extending the scope to publications on “array” would list more than 120,000 abstracts in PubMed. Various databases and resources contain a wealth of microarray data for studies of gene expression, proteomics, tissue analysis, DNA methylation, genome-wide association studies, copy number variations and aberrations, and single nucleotide polymorphism, among others. The Gene Expression Omnibus database of the National Center for Integrative Biomedical Informatics of the United States contains a wealth of gene expression data, which are publicly accessible via the Internet. These databases contain millions of samples, mostly based on microarray studies.

Since the introduction of microarrays, huge amounts of data have been generated, and hence it has been imperative to create and improve specialized tools and
algorithms for the acquisition, analysis, and management and interpretation of the
data. In every aspect, microarray data analysis is invariably an essential component
of microarray studies. Roughly speaking, data analysis tools for microarrays are
derived from a variety of bioinformatics methods and techniques that bridge the gap
among molecular biology, computer science, statistics, and mathematics. Microar-
ray data analysis takes on the images that are obtained from scanning and up to the
biological and functional analysis, involving a sequence of well-established steps
aiming at identifying locations of spots or probe cells, background, and noise, and to
subsequently quantify and normalize the signals. Numerical data resulting from the
pre-processing stages are analyzed by means of different techniques for clustering,
data mining, and classification, followed by visualization, and network and func-
tional biological analysis. Data management involves storage, databases, resources,
and compression, which are important steps that complement the whole process of
analysis. Microarray image and data analysis is essential, since the aim of the whole
experimental procedure is to obtain meaningful biological conclusions, which de-
PENDS on the accuracy of the analytical steps, mainly those at the beginning of the
pre-processing stage.

This book provides a comprehensive review of the main, up-to-date methods,
tools, and techniques for microarray image and data analysis. Internationally recog-
nized experts address specific research topics and challenges in their areas of expert-
tise, providing valuable knowledge about the state-of-the-art methods in the field, and
covering the main steps for image processing, gridding, segmentation, noise treat-
ment, and normalization. Biological aspects and applications of microarrays extend
to a wide range of biological studies in cancer research, and different microarray
types and platforms such as DNA, various types of oligonucleotide and tissue mi-
croarrays, as well as the most recent tools for microarray data analysis in the Bio-
conductor suite. Machine learning methods for microarray data analysis cover the
main aspects of clustering, biclustering, multi-dimensional microarray data analysis,
and reconstruction of regulatory networks.

The book is intended for students as a graduate textbook and as a reference for
the design of graduate courses at an advanced level, while providing sufficient level
of detail for general readers as well. In addition, one of the goals of this book is to
provide a useful reference for researchers and practitioners in academia and industry
who develop tools and carry out critical analysis, management, and interpretation of
microarray data. By covering state-of-the-art methods in microarray image and data
analysis, the book provides a handy reference for professors, researchers, and gradu-
ate and advanced-level undergraduate students in multi-disciplinary fields including
bioinformatics, computer science, statistics, biology, biochemistry, genomics, and
biomedical engineering. While the readers will also learn about the current state of
the microarray technology and how data generated by the microarray experiments
are analyzed, they will understand the current landscape regarding the applications,
availability, and affordability of the research utilizing this technology.

This book has been made possible through the help of many people who partici-
pated at different stages of the production process. I would like to thank all of them
for their valuable help and support to make this happen. I am indebted to the book series editor, Rastislav Lukac, who specially invited me to edit this book and guided me through the main steps. My special thanks are for Adnan Ali for adding a self-contained biological insight to the book, which provides an excellent complement to the computing and statistical methods being presented. Special thanks are also for Iman Rezaeian for his tireless hours of work put into the edition and format conversions, as well as his contribution to a chapter. I would also like to thank Alioune Ngom for his contribution and help in reviewing, and for his constant logistic support, and Yifeng Li for his additional help in the review process.

Finally, I would like to thank the expert contributors for putting their efforts in carefully writing up-to-date and high quality chapters about their specific topic, while keeping the whole volume well synchronized, and for their help in the reviewing process. I would like to acknowledge the external reviewers, Michael Crawford, Meng Li, A. Sri Nagesh, Moysés Nascimento, Tuan Pham, Armando Pinho, Alexander Pozhitkov, Khedidja Seridi, Guifang Shao, Ruisheng Wang, Xiongwu Wu, Hong Yan, and Jie Zheng. With their solid expertise in the specific topics, they added the valuable feedback needed to establish these contributions as leading in the field.
Luis Rueda obtained his Ph.D. in computer science from Carleton University, Canada, in 2002. He joined the School of Computer Science at the University of Windsor in Canada, in 2002 as an Assistant Professor. After spending two years at the University of Concepción, Chile, he was appointed as an Associate Professor within the School of Computer Science and with the Pattern Recognition and Bioinformatics Lab at the University of Windsor. He has been recently promoted to Full Professor. His research interests are mainly focused on theoretical and applied machine learning and pattern recognition, mostly in the fields of transcriptomics, interactomics, and genomics. His research has mostly been centered on microarray and next generation sequence data analysis, including image and signal processing, pattern analysis and gene selection, and on finding relevant patterns in prediction of protein-protein interactions and stability of protein complexes, including domains, short-linear motifs, and complex types. Luis Rueda holds three patents on data encryption, secrecy, and stealth. He has published more than 100 publications in prestigious journals and conferences in machine learning and bioinformatics. He has participated in editorial and technical committees for conferences and journals. He is a Senior Member of the IEEE, and a Member of the Association for Computing Machinery and of the International Society for Computational Biology.
Contributors

José Adélaïde
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
adelaidej@ipc.unicancer.fr

Adnan Ali
School of Computer Science, University of Windsor
Windsor, Ontario, Canada
aali@uwindsor.ca

Jose Manuel Arteaga-Salas
Cooperation Group CMA, Helmholtz Center Munich
Munich, Germany
jose.arteaga@helmholtz-muenchen.de

Wassim Ayadi
LaTICE, ESSTT, Université de Tunis, Tunisia
LERIA, Université d’Angers, France
wassim.ayadi@gmail.com

Dimitris Bariamis
Dept. of Informatics and Telecommunications, University of Athens
Athens, Greece
dimitris.bariamis@gmail.com

Matthew Bashton
Bioinformatics Support Unit, Newcastle University
Newcastle upon Tyne, UK
matthew.bashton@newcastle.ac.uk

Ismahane Bekhouche
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
bekhouchei@ipc.unicancer.fr

Francis Bell
Biomedical Engineering, Science and Health Systems, Drexel University
Philadelphia, Pennsylvania, USA
fxb22@drexel.edu

François Bertucci
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
bertuccif@ipc.unicancer.fr

Ghislain Bidaut
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
ghislain.bidaut@inserm.fr

Daniel Birnbaum
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
daniel.birnbaum@inserm.fr
Max Chaffanet
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France;
Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
chaffanetm@ipc.unicancer.fr

Pascal Finetti
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France;
Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
finettip@ipc.unicancer.fr

Tai-Been Chen
Dept. of Medical Imaging and Radiological Sciences, I-Shou University
Kaohsiung, Taiwan
tb@isu.edu.tw

Caroline C. Friedel
Institute for Informatics, Ludwig-Maximilians-University Munich
Munich, Germany
Caroline.Friedel@lmu.de

Chih-wen Cheng
Dept. of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, Georgia, USA
cwcheng83@gatech.edu

Maxime Garcia
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France;
Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
maxime.u.garcia@gmail.com

Simon Cockell
Bioinformatics Support Unit, Newcastle University
Newcastle upon Tyne, UK
simon.cockell@newcastle.ac.uk

Colin S. Gillespie
School of Mathematics & Statistics, Newcastle University
Newcastle upon Tyne, UK
colin.gillespie@newcastle.ac.uk

Urška Cvek
Dept. of Computer Science, Louisiana State University Shreveport
Center for Molecular and Tumor Virology, Louisiana State University Health Sciences Center Shreveport
Shreveport, Louisiana, USA
ucvek@lsus.edu

Arnaud Guille
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France;
Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
arnaud.guille@inserm.fr

Mourad Elloumi
LaTICE, ESSTT, Université de Tunisie Tunisia
mourad.elloumi@fsegt.rnu.tn

Jin-Kao Hao
LERIA, Université d’Angers France
hao@info.univ-angers.fr
Let the contribution of the contributors have a clear and visible place in the document. Miguel Hernández-Cabronero
Dept. of Information and Communications Engineering, Universitat Autonoma de Barcelona Bellaterra, Barcelona, Spain
miguel.hernandez@uab.cat

Sonal Kothari
Dept. of Electrical and Computer Engineering, Georgia Institute of Technology Atlanta, Georgia, USA
sonalkothari86@gmail.com

Yifeng Li
School of Computer Science, University of Windsor Windsor, Ontario, Canada
li1112c@uwindsor.ca

Henry Horng-Shing Lu
Institute of Statistics, National Chiao Tung University Hsinchu, Taiwan
hslu@stat.nctu.edu.tw

Rastislav Lukac
Foveon, Inc. / Sigma Corp. San Jose, California, USA
lukac@colorimageprocessing.com

Michael W. Marcellin
Dept. of Electrical and Computer Engineering, The University of Arizona Tucson, Arizona, USA
mwm@email.arizona.edu

Dimitris Maroulis
Dept. of Informatics and Telecommunications, University of Athens Athens, Greece
dmaroulis@di.uoa.gr

Raphaële Millat-Carus
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
raphaele.millatcarus@gmail.com

Alioune Ngom
School of Computer Science, University of Windsor Windsor, Ontario, Canada
angom@uwindsor.ca

Rehman Qureshi
Biomedical Engineering, Science, and Health Systems, Drexel University Philadelphia, Pennsylvania, USA
raq22@drexel.edu

Iman Rezaeian
School of Computer Science, University of Windsor Windsor, Ontario, Canada
rezaeia@uwindsor.ca

Luis Rueda
School of Computer Science, University of Windsor Windsor, Ontario, Canada
lrueda@uwindsor.ca

Renaud Sabatier
Aix-Marseille Université; Institut Paoli-Calmettes, Marseille, France; Inserm U1068; CNRS 7258, Centre de Recherche en Cancérologie de Marseille, Marseille, France
sabatierr@ipc.unicancer.fr

Ahmet Sacan
Biomedical Engineering, Science, and Health Systems, Drexel University Philadelphia, Pennsylvania, USA
ahmet.sacan@drexel.edu
Michalis Savelonas  
Dept. of Informatics and Telecommunications, University of Athens  
Athens, Greece  
msavel@di.uoa.gr

Joan Serra-Sagristà  
Dept. of Information and Communications Engineering, Universitat Autonoma de Barcelona Bellaterra, Barcelona, Spain  
joan.serra@uab.cat

Shahram Shirani  
Dept. of Electrical and Computer Engineering, McMaster University  
Hamilton, Ontario, Canada  
shirani@mcmaster.ca

Todd H. Stokes  
Dept. of Biomedical Engineering, Emory University and Georgia Institute of Technology  
Atlanta, Georgia, USA  
todd.h.stokes@gmail.com

Marjan Trutschl  
Dept. of Computer Science, Louisiana State University Shreveport Center for Molecular and Tumor Virology, Louisiana State University Health Sciences Center Shreveport Shreveport, Louisiana, USA  
mtrutsch@lsus.edu

Meng-Yuan Tsai  
Institute of Statistics, National Chiao Tung University  
Hsinchu, Taiwan  
u9826804@stat.nctu.edu.tw

May D. Wang  
Dept. of Biomedical Engineering, Emory University and Georgia Institute of Technology  
Atlanta, Georgia, USA  
maywang@bme.gatech.edu

Yiqian Zhou  
Biomedical Engineering, Science, and Health Systems, Drexel University  
Philadelphia, Pennsylvania, USA  
yz86@drexel.edu
1 Introduction to Microarrays

Luis Rueda and Adnan Ali

CONTENTS

1.1 Introduction .........................................................2
1.2 Molecular Biology and Genomics..............................2
  1.2.1 Genome Structure and Function ......................3
  1.2.2 Macromolecules ...........................................4
  1.2.3 Nucleic Acids .............................................5
  1.2.4 Proteins .....................................................6
  1.2.5 Transcription and Translation .........................7
  1.2.6 Alternative Splicing ......................................8
  1.2.7 Reverse Transcription ...................................9
1.2.8 Translation ..................................................9
1.3 Regulation of Gene Expression ..................................10
1.4 Microarrays ....................................................13
1.5 Analysis of Microarray Data ....................................15
  1.5.1 Microarray Image Processing ..........................17
    1.5.1.1 DNA Microarrays ..................................19
    1.5.1.2 Oligonucleotide Arrays ............................23
    1.5.1.3 Protein Microarrays ...............................24
  1.5.2 Quantification and Normalization .....................24
  1.5.3 Pattern Analysis and Machine Learning ............25
  1.5.4 Time Series and Higher Dimensions ................26
  1.5.5 Visualization .............................................26
  1.5.6 Functional and Network Analysis ....................27
1.6 Resources for Microarrays ....................................27
  1.6.1 Publications .............................................28
  1.6.2 Databases and Repositories ...........................30
  1.6.3 Tools for Microarray Data Analysis ................32
1.7 Conclusions and Future Trends ............................34
Acknowledgments ..................................................35
References ..........................................................35
1.1 INTRODUCTION

During the last decade, microarrays have emerged as essential tools for many researchers to monitor biomolecules’ behaviors for thousands of genes simultaneously. Since their discovery, diverse applications and complex issues in processing and analysis of high-throughput data have been the center of modern molecular biology research, because an accurate analysis is one of the keys to provide a meaningful interpretation of the data. Microarrays are used to explore different manifestations of the genomic information in terms of cellular processes and biochemical products that ultimately dictate the behavior of a particular organism or a specific cell type and its interactions with others. Proteins, cells, tissues, and an entire organism can be studied at different levels of detail in a massive and parallel way, offering innumerable possibilities and applications in biomedical research, medical genetics, pharmacology, disease diagnosis, staging and prognosis, pharmaco-therapeutics, evolutionary studies, conservation science, and genetic polymorphic studies, to mention a few.

This introductory chapter discusses basic concepts of molecular biology, genomics and regulation of gene expression. Thereafter, an overview of the principles and fundamentals of microarrays is presented from a general perspective. In a superficial way, more emphasis is given on the means and methods for microarray production, scanning and the main steps of microarray image acquisition. Included also in this chapter are discussions, both general and specific, of the main methods for pre-processing and analyzing microarrays, including different aspects of gridding, segmentation, noise treatment, data storage and retrieval, and microarray data analysis tools available to the scientific community. The chapter also presents an overview of the existing resources for microarray data, publications, public databases and repositories, and analytical tools.

1.2 MOLECULAR BIOLOGY AND GENOMICS

This section introduces the principles of molecular biology, from the basic concepts of nucleotides, nucleic acids, amino acids, and proteins to other macromolecules of interest. A fundamental concept in molecular biology, the central dogma, defines a basic flow of genetic information from DNA to RNA and finally into a protein product. Encoded within the genomic DNA molecule is all of the genetic information in form of its specific nucleotide sequence, which is ultimately transcribed into RNA. In turn, RNA encodes the information (codons) for translation into a protein molecule that contains a specific sequence of amino acid residues. Since the discovery of nucleic acid molecules, the central dogma has evolved into its present form where other aspects, such as reverse-transcription of RNA into cDNA molecules, are also considered to be part of the basic concept (Figure 1.1). Other forms of heterogeneous RNA molecules, tRNA, rRNA, siRNA, miRNA, and lincRNA do not encode protein products but rather have specific functions in other important cellular activities, including the process of translation itself. In addition, protein folding, structural modifications and molecular interactions are required for the proper function or activity of some proteins and their domains (with different gray scales in Figure 1.1), conforming net-
FIGURE 1.1  The basic concept of the flow of genetic information in biological systems stems from the central dogma in biology. DNA is transcribed into RNA molecules, which in turn encodes a specific amino acid sequence in a protein molecule. After folding, most proteins adopt a structure and interact with other molecules.

works of protein and domain interactions. For more details, the reader should refer to more specific textbooks in the subject [1, 2]. In the following sections, a basic understanding of the molecular structure of macromolecules and their interactions with each other are discussed with respect to the concept of microarray hybridization.

1.2.1 GENOME STRUCTURE AND FUNCTION

In eukaryotic cells, the genomic DNA is present in a discrete nucleus in the form of chromosomes wherein the double stranded DNA is complexed with a number of protein molecules, such as histones. In contrast, the genomic DNA in prokaryotic cells is a circular nucleoid structure present in the cell without any nuclear envelope. During the last two decades, genome sequencing projects of several organisms, both eukaryotic and prokaryotic, have been completed where the entire genome sequences of the genomic DNA have been identified and comparative analyses have been performed.

The genome of an organism presents a complete set of genes and is maintained generation after generation. It is important to note that the eukaryotic genomic DNA comprises coding as well as non-coding regions. Following the transcription of genes using genomic DNA as a template, the coding regions are eventually spliced together and then ultimately translated into a final protein product in the cytoplasm of the cell. The identity of a gene in the genome is specified by the presence of open reading frames (ORF) containing the coding regions, and regulatory regions that may be present upstream or downstream of the open reading frame (coding region). In the context of microarrays, for the study of gene expression, we are mainly concerned with the coding regions of the genomic DNA. Therefore, it is important to define
the concept of transcriptome, which represents the entire repertoire of genes that are expressed in a cell at a given time under defined conditions. Since the RNA molecules transcribed from coding regions may be non-coding RNA molecules or messenger RNA (mRNA) coding molecules, the transcription of the genes results in mRNA molecules, which are more relevant when one considers gene expression to study the production of protein products in a cell.

Primarily, the production of proteins in the cell ultimately renders certain cell types to acquire their desired function or phenotype. This is the property of a protein molecule that gives the proper morphology and physiology of each cell type. Certainly, the role of other types of important molecules, like carbohydrates, polysaccharides, and lipids in this regard cannot be ignored either. Since higher eukaryotic organisms comprise multiple types of cells, each cell type has its own structure and function, and hence is able to form part of a specific tissue type by cellular differentiation leading to the division of labor. In this sense, a proteome represents a complete set of proteins being produced by specific mRNA molecules expressed in a specific transcriptome. Different proteins may function independently or in combination and coordination with their other protein or carbohydrate counterparts. Thus, a diversity of interactions of macromolecules is achieved in a very complex manner, and is responsible for the presence of multiple types of cells, and hence a variety of tissues within a single higher eukaryotic organism.

The cells in an organism are exposed to a number of external and internal stimuli that can initiate a cascade of signaling mechanisms that eventually affect the regulation of gene expression from the genomic DNA that resides in the nucleus in case of eukaryotic organisms. The process of these signaling mechanisms is commonly termed as signal transduction. There are numerous signaling cascades in cells that run in parallel under normal or stress conditions. Cells respond to these signals in an appropriate manner to carry out normal cellular functioning. Any malfunctioning in such signaling mechanisms causes cells to behave in an otherwise inappropriate manner. Such malfunctioning in signal transduction mechanisms has been shown to cause severe disorders, including cancer. This topic is further discussed in Chapter 2.

1.2.2 MACROMOLECULES

All living organisms are composed of cells, which present a wide range of variations from organism to organism and from cell to cell within the same organism. The molecules that form the cellular structures are generally divided into (i) small molecules, such as fatty acids, amino acids, nucleotides, and sugar molecules, and (ii) large molecules or macromolecules, such as lipids, proteins, nucleic acids, and polysaccharides [3]. Primarily, small molecules interact to make polymers that give rise to different polymeric structures of macromolecules. Here, we briefly consider the molecular structure and function of nucleic acid and protein molecules. Since these macromolecules are produced by aggregation of nucleotide and amino acid molecules, respectively, it is necessary to understand how it is possible that small molecules polymerize to form macromolecular structures that contain unique properties and functions, and contain all the genetic information.
1.2.3 NUCLEIC ACIDS

A nitrogenous base linked to a sugar is called *nucleoside*, and a nucleoside linked to one or more phosphate groups is called *nucleotide*. Nucleotides make the primary building blocks of nucleic acid molecules, which can be *deoxyribonucleic acid* (DNA) or *ribonucleic acid* (RNA) molecules. The DNA has a 2'-deoxyribose, whereas an RNA molecule has a ribose instead. In summary, a typical nucleotide molecule comprises three components: a nitrogenous base, a sugar, and a phosphate molecule. In any nucleic acid molecule, the nitrogenous base is either a purine (adenine and guanine) or a pyrimidine (cytosine and thymine). A nucleic acid molecule comprises a chain of nucleotide molecules linked together by phosphodiester linkages. The backbone of the resulting polynucleotide is composed of alternating sugar and phosphate residues wherein the 5' end of one pentose ring is linked to the 3' end of the next pentose ring via a phosphate group in between. Each nucleic acid molecule contains the four types of bases (A, G, C, and T) in a defined order or a sequence. In both DNA and RNA molecules, the same type of purines are present. However, a difference exists with respect to the pyrimidines. In a DNA molecule, cytosine and thymine are present, whereas cytosine and uracil are present in an RNA molecule. Therefore, when writing the nucleic acid sequence of a DNA molecule with sequence ACGT, the corresponding identical sequence is written as ACGU for an RNA molecule. The terminal ends of nucleic acid molecules contain a 5' terminal free phosphate group and a 3' most OH group. Traditionally, the sequence of a nucleic acid molecule is written in 5' to 3' direction. The terminal ends are often used to tag the nucleic acid molecules with different labels, which aid in their detection using conventional biological assays. The synthesis of RNA molecules from a genomic DNA template within the cell is called *transcription*.

By virtue of their structure, there is a specific interaction between purines and pyrimidines. So far, we have considered only a single linear polynucleotide molecule, which is produced by polymerizing individual nucleotides. When considering a double stranded DNA molecule, which occurs in nature, the specific interactions between purines and pyrimidines become crucial. In this regard, adenine can form bonding with thymine; and guanine can interact with cytosine. This hydrogen bonding between nitrogenous bases is termed as *base pairing* and is said to be complementary in nature. Therefore, the sequence of a given DNA molecule contains two polynucleotide molecules held together by *hydrogen bonding* in an antiparallel fashion, i.e., one strand or polynucleotide in 5'-3' orientation will face its complementary strand in 3'-5' direction.

It is also important to note that, in nature, RNA molecules exist in single stranded conformation. The same property held for nucleotides to form hydrogen bonding with their respective bases allows us to make *in vitro* assays, where one polynucleotide is used as a probe to detect its corresponding complementary strand in a given biological sample, by an assay commonly called *hybridization*. Hybridization can occur between DNA-DNA, RNA-RNA, and DNA-RNA. The ability of single stranded nucleic acid molecules to hybridize can be used as a measure of complementarity between the two nucleic acid sequences. In fact, this is the basic principle
used in gene expression studies using microarray platforms. Hybridization between nucleic acid molecules can be controlled by determining the \textit{in vitro} conditions under which they can denature, renature or hybridize. The detection of hybridization is a common routine and various detection methods are used in a biological laboratory setup. Traditionally, Southern blots have been used to detect DNA-DNA hybridization, and Northern blots for the detection of RNA molecules using a labeled DNA probe (DNA-RNA hybridization). Microarrays primarily provide a method for detection of RNA molecules using immobilized DNA probes. A variety of techniques exist for DNA sequencing in which the sequence of individual nucleotides is determined using chemical reactions for the detection of the nucleotides.

1.2.4 PROTEINS

Protein molecules form major components in a cell. Amino acid residues are linked to each other by peptide bonds and form small continuous chains called \textit{peptides}. Within the cytoplasm of a cell, the synthesis of a polypeptide, encoded by an mRNA strand, is called \textit{translation}. A protein molecule consists of one or more linear polymers and comprises a specific sequence or order of amino acid residues, where each residue contains one of the 20 amino acids that exist in nature—amino acid molecules are called \textit{residues} when they are incorporated into a polypeptide. A protein molecule may contain several polypeptides, where each polypeptide is synthesized separately. Also, a polypeptide chain may be truncated into several shorter peptides, which may be part of a final structure of a protein. Following the synthesis of polypeptides, they may undergo several modifications within the cell, achieving specific properties for the proper functioning of the final protein product. The sequence of amino acids in a polypeptide can be determined \textit{in vitro} by a chemical reaction termed as \textit{Edman degradation}.

Each amino acid consists of an amino group that is linked to a carbon, which is ultimately linked to a carboxyl group. In addition to carbon, each residue has a distinctive side chain or \textit{R} group that determines the specific properties of the amino acid contained in it. In a polypeptide, the amino acids are arranged in an amino (N-terminus) to carboxyl (C-terminus) direction. The primary structure of a protein molecule is determined by the composition of a variety of amino acid residues and their specific order in the polypeptide. Different types of amino acid residues present unique properties with respect to polarity, charge, acidity and basicity due to the nature of their form and structure. Following the synthesis of a polypeptide, further post-translational modifications provide specific properties to the protein structure and function. Each protein has a unique structure, which is achieved through the intracellular modifications of the polypeptide molecules and via its interaction(s) with a number of other proteins or carbohydrate molecules.

A proper folding of some protein molecules is crucial to their function within the cell or outside the cell when released. During the process of folding, the chain of amino acid residues can fold into secondary or tertiary structures with the help of chaperones within the cytoplasm of the cell. Proteins can be structural proteins, such as the fibrous protein keratin, or globular proteins, such as enzymes that can
catalyze specific enzymatic reactions. Also, proteins can be retained within the cell or released from the cell where they perform their extracellular functions. Secreted enzymes and hormones are examples of some secreted proteins.

Historically, the identification of antibodies has had a significant impact in protein biology. Antibodies are also protein molecules that bind with specific antigenic sites present within a target protein molecule. The specificity of binding of antibodies with their specific domains in a polypeptide molecule makes them very useful in polypeptide or protein detection assays. For example, an antibody generated against a specific antigenic site (epitope) in a protein will always bind to its specific epitope when exposed to that target protein molecule. Thus, the specificity of antibody-antigen interaction has allowed the development of a number of protein detection assays, such as Western blots, immunoprecipitation, enzyme-linked immunosorbent assays (ELISA), as well as protein and antibody microarrays. Antibodies can be labeled with different radioactive or non-radioactive labeling molecules. The labeled specific antibodies are also used to isolate, purify and detect specific proteins. Western blots are commonly used to detect specific protein molecules using labeled antibodies. One of such assays is termed as immunoprecipitation, where a specific antibody is used to isolate and purify a specific protein molecule from a mixture of proteins in a given biological sample. The use of antibodies to detect proteins has also been widely used in microarrays, where a variety of protein molecules can be detected simultaneously.

1.2.5 TRANSCRIPTION AND TRANSLATION

As indicated earlier in this chapter, the synthesis of RNA molecules using DNA as a template is called transcription. There are multiple protein factors that are involved in the initiation, elongation and termination of the transcription in the cell. In eukaryotic cells, this phenomenon occurs within the nucleus of the cell wherein the genomic DNA present in the chromosomes acts as a template. The RNA produced in the cell is identical to one strand of the double-stranded genomic DNA, which is called coding strand. Therefore, the complementary strand or the non-coding strand of the double-stranded genomic DNA acts as the actual template for the synthesis of new RNA. Transcription reaction is catalyzed by an enzyme called RNA polymerase. Transcription starts with the binding of the RNA polymerase enzyme to the promoter region of a gene that is present upstream of the coding region of the genomic DNA. This binding is facilitated by several factors without which the transcription would not proceed. In the gene, a starting point marks the start site for the synthesis of RNA.

Transcription proceeds with the movement of the polymerase complex along the template strand until it reaches a terminator sequence in the template strand. The double stranded DNA is uncoiled by the action of other protein factors making the template available for accessing the polymerase complex, which moves forward. During this process, the polymerase enzyme is capable of reading the template and adding an appropriate complementary ribonucleotide in 5′ to 3′ direction. Every new ribonucleotide is added to the 3′ OH end of the previously incorporated ribonucleotide.
Once the polymerase complex reaches the terminator sequence, the complex falls off and the single-stranded newly synthesized RNA is released and is then free to move into the cytoplasm of the cell.

The region containing sequences before the starting point of the gene is called upstream, whereas the region following the terminator region is termed as downstream. The first base being transcribed into RNA is numbered as +1, whereas the base prior to it is −1. There is no base with a number zero in a gene. There are also other specific regions upstream of a promoter or downstream of a terminator region, which are also involved in the regulation of gene expression, e.g., enhancers. However, the discussion of such regions is not intended for the present chapter. Three different types of RNA molecules are produced in the cell through the process of transcription. These are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). The tRNA and rRNA molecules are mainly involved in the process of synthesis of protein molecules during translation, whereas the mRNA actually provides a template for the protein synthesis. The mRNA, as apparent from its name, represents the coding region of a gene and will be translated into the final polypeptide molecule. The polypeptide molecules may further become modified or interact with other polypeptide molecules to produce the final protein molecule. Here, we limit our discussion to mRNA molecules as they are the main vehicles for the active gene expression at the transcriptional level.

**1.2.6 ALTERNATIVE SPLICING**

Once the pre-mRNA (nascent RNA molecule) has been synthesized, it may also be subjected to further modifications. For example, one of such modifications is termed as *RNA splicing* in eukaryotic cells. During RNA splicing, the regions in the RNA molecules that may not represent coding regions are spliced out, and a mature mRNA molecule is produced. The genomic DNA contains *exons* (coding regions) and *introns* (non-coding regions) that are also transcribed into mRNA. The transcription mechanism by itself cannot discriminate the introns from exons. Therefore, the RNA splicing mechanism is responsible for ensuring the synthesis of mature mRNA, which does not include any non-coding sequences.

There is also another layer of complexity here, wherein the splicing mechanism is not always used only to splice out non-coding regions. Sometimes, the splicing of some exons is also observed wherein a distinct species of mRNA molecules can be generated from a single type of a newly synthesized full length RNA molecule. For example, if a gene consists of three different exons, exon 1, exon 2, and exon 3, an alternative splicing mechanism can produce a mature mRNA molecule that may contain any combination of these three exons, e.g., exon 1-exon 2-exon 3; or exon 1-exon 2; or exon 1-exon 3, and so on. The presence of alternative splicing, which is unique to eukaryotic cells, provides another level of regulation of gene expression at the post-transcriptional level. There are many genes within the cell that are regulated by alternative splicing. Such alternative splicing results in the biodiversity of protein molecules being synthesized in the cells in which one gene can code for multiple isoforms of protein molecules with different functions. In the context of al-
ternative splicing in the regulation of cellular activities, the use of microarrays has proven to be an excellent tool. The presence of specific exons/introns as probes on a microarray enables the detection of transcribed regions of a gene by detecting the presence of the corresponding RNA transcripts in a given sample using microarray hybridization assays. This topic is further discussed in Chapter 2 with reference to the post-transcriptional regulation of gene expression.

1.2.7 REVERSE TRANSCRIPTION

In certain RNA viruses, called retroviruses, some special enzymes are present for genetic modifications such as transposition of genetic elements. These enzymes are termed as reverse transcriptases. These enzymes from viral particles have been isolated and purified. Under suitable conditions, reverse transcriptase enzymes can be used to reverse transcribe an RNA molecule to obtain a complementary DNA strand, termed as cDNA. The newly synthesized cDNA molecule can be further made double stranded using the same reverse transcriptase enzyme in the same reaction, since the enzyme can read RNA or DNA templates for the synthesis of a complementary DNA strand. In fact, the reverse transcription reaction has made a significant contribution towards understanding molecular mechanisms within eukaryotic cells.

The genes that are expressed in a cell are actively transcribed and mRNA from the cellular mixture can be isolated using standard techniques. The mRNA isolated from the cell can be reverse transcribed and the resulting double stranded cDNA can be cloned into DNA vectors for sequencing, amplification by polymerase chain reaction, or further genomic studies as required. In a microarray experiment, detection of gene expression is carried out using the same reverse transcription reaction during which different labeling molecules, such as Cy3 or Cy5, can be used to label the resulting cDNA, which is then used to hybridize against the immobilized DNA probes on a solid support.

1.2.8 TRANSLATION

Following the synthesis of mRNA via gene transcription, the mechanism of translation into protein takes place in the cytoplasm, which again involves a number of factors including ribosomal machinery of the cell. This process of biochemical synthesis of protein molecules is dependent upon the specific sequence of the template mRNA transcribed from a specific gene. Within the protein coding region in the mRNA molecule each triplet (three consecutive nucleotides) is called codon. With a few exceptions, the genetic code related to the nucleotide triplets is considered as universal code. Each amino acid is encoded by a specific triplet codon. For example, a standard codon AUG specifies the incorporation of a methionine residue in a nascent polypeptide molecule. This codon is also termed as initiation codon because usually methionine is a standard first amino acid residue in eukaryotic polypeptides. In eukaryotes, the termination codon is one of the three codons UAA, UGA and UAG, which do not encode any amino acid residue. Translation of an mRNA molecule requires an mRNA template, tRNA molecules linked to amino acid residues, and
ribosomes. Following the synthesis of polypeptide molecules, a post-translational modification of the newly synthesized polypeptide molecule occurs with the help of chaperone proteins, which allow the proper folding and conformation of the final protein structure. Other processes, such as glycosylation, acetylation, and acylation may also occur with specific polypeptides as needed before they can enter into their final secondary or tertiary conformation.

As indicated earlier, antibodies are also protein molecules that are primarily host molecules produced in response to foreign molecules (antigens) and are also protein molecules themselves. Antibodies are produced by plasma cells and circulate in the body where they interact with foreign antigens and make an antibody-antigen complex, which is ultimately removed through phagocytosis. Fundamentally, antibodies are a family of glycoproteins that share structural and functional features in some aspects. The specific interaction of antibodies with their respective antigenic sites in the target protein has allowed their frequent use in the protein detection assays. In protein microarrays, the use of known specific antibodies is employed to detect the presence of their respective binding proteins in a mixture of proteins in a given biological sample.

1.3 REGULATION OF GENE EXPRESSION

Regulation of gene expression can occur at various levels within the cell. A given gene may be considered to be regulated for its expression, for example synthesis of mRNA or protein molecules, or for its associated activity. In view of the microarrays, we consider only the expression of the genes at the level of transcription or translation. In this regard, genes may be regulated at the level of genomic DNA, transcriptional level, post-transcriptional level, translational level, or even at post-translational level. Living cells are continuously exposed to internal or external stimuli, and they are well equipped with an intracellular infrastructure to respond to these signals by changing the internal cellular activities in a number of ways (Figure 1.2). For example, in response to a chemical exposure, a specific signaling mechanism is activated that can ultimately affect the regulation of gene expression by regulating the transcription, translation, or even by modifying the post-translation structure or function of the protein molecules. The study of genetic changes in the genomic DNA on a large scale is termed as genomics. Similarly, obtaining a global picture of the active transcriptional state of a cell, the transcriptomics, and the entire protein expression and interactive architecture, the proteomics, are studied on a genome-wide scale (Figure 1.2).

In genomic DNA, the variation in the nucleotides may occur within the member of a given biological species. The variation may be at a single nucleotide level, termed as single nucleotide polymorphism (SNP). Such variation in a single nucleotide may also be present within the paired chromosomes giving two kinds of alleles. Allelic variations due to the presence of SNPs can occur within the coding as well as in non-coding regions. Such variations may or may not have any effect on gene expression. For example, a specific SNP in a promoter region may present a mutation that impedes binding of the RNA polymerase complex, and hence inhibiting the ini-
FIGURE 1.2  Analysis of complex biological processes on a genome-wide scale. The genetic changes at large scale (genomics), and the expressed genes at the RNA level (transcriptomics) or at the protein level (proteomics), and their molecular interactions can be visualized using a single microarray representative of the entire genome.

tiation of transcription. However, a greater number of SNPs appears to be present within the non-coding region in the genomic DNA, which is presumed to have some type of evolutionary significance. A number of SNPs have been associated with the presence of a range of human disorders, such as cancer, infectious diseases, and autoimmune diseases, which appear to result from particular SNPs. Even if the SNPs themselves do not cause a certain disorder, their use in the linkage analysis has allowed the identification of candidate genes that are directly associated with certain diseases. Therefore, SNPs provide important biomarkers in clinical diagnosis for certain genetic diseases. Microarrays have also provided scientists with a way to use high-throughput identification of SNP markers from a number of biological samples simultaneously. Such polymorphic analysis using either target SNP sites, or their genome-wide association studies, for certain genetic diseases using microarrays are discussed in more detail in Chapter 2.

It has been shown, however, that some modifications of the genomic DNA do not involve mutations or modifications of the DNA sequence itself. Such modifications can occur in the chromatin conformation or the biochemical modification of certain nucleotides in the DNA. One of these modifications involves methylation of cytosine nucleotides in the promoter region of some genes. Such heritable changes in the DNA, which affect gene expression or cellular phenotype, are caused by mechanisms
other than changes in the DNA sequence itself, and are called epigenetic regulation. There are certain enzymes present within the nucleus of eukaryotic cells, which can methylate cytosine nucleotide when present in the promoter region as a dinucleotide CpG (cytosine-phosphate-guanine). Such methylation may impede the RNA polymerase complex to access the promoter to initiate the transcription of the coding region, which is downstream of the promoter region. A number of eukaryotic genes have been shown to be regulated by such methylation process. Methylation-sensitive microarrays are discussed in Chapter 2.

The next level of regulation of gene expression is at the level of transcription. Here, a number of regulatory proteins, the trans-acting factors (such as transcription factor), the regulatory elements within the gene, and cis-acting elements (such as a promoter region) interact with each other to activate or deactivate transcription. The cells can be stimulated by external factors or internal cues, which may direct the activation of transcription factors to interact with their corresponding cis-acting elements and allow the RNA polymerase to initiate the transcription of the gene. Certain genes are always expressed within the cell without any stimulation—such genes are called house-keeping genes or constitutively expressed genes. Examples of inducible genes are genes regulated by growth factors, hormones, other physiological signals, external stimuli, toxicants, mutagens, or high temperatures. Similarly, certain genes are expressed abnormally in cancer cells due to malfunctioning of signaling mechanisms in a cell. Such abnormal expression of certain genes can initiate a cascade of events in the cell, which ultimately modify the cellular behavior and morphology. This phenomenon may cause formation of lumps of cells due to uncontrolled growth and results in the formation of a tumor. Therefore, scientists have always been keen to understand the regulation of gene expression using traditional techniques, such as Northern blots, polymerase chain reaction, primer extension, RNase protection assays, and serial analysis of gene expression (SAGE), to name a few. The advent of microarrays has played an important role in high-throughput gene expression studies, since they can provide great detail about the transcription of a wide selection of genes in a cell at the same time under the same conditions.

One of the most important applications of microarrays is in post-transcriptional regulation studies. As discussed earlier, the molecular mechanisms, such as RNA splicing, also occur in the cell and may regulate the synthesis of specific multiple types of mRNA molecules from a single gene. It is possible to detect alternative splicing using various molecular techniques such as polymerase chain reaction and primer extension assays. Microarrays have also been employed to study the alternative splicing in the regulation of gene expression.

Translational regulation of gene expression is another area of cellular research where the application of microarrays has been widely used in recent years. Translation of mRNA templates to produce polypeptide molecules finally determines the cellular phenotype. The presence of specific proteins within the cell can be examined by isolating a mixture of the proteins of the cell employing techniques such as Western blots using specific antibodies. However, the use of microarrays has allowed scientists to take benefit of multiple antibodies, and detect a large number of protein
molecules simultaneously in a single experiment. The use of protein microarrays is also discussed in Chapter 2.

1.4 MICROARRAYS

Microarray technology allows scientists to study and examine key biological questions on a genomic scale. This has influenced the way in which transcriptional, proteomic and cell developmental studies are carried out, providing a systematic way to study gene expression, elucidation of the protein expression, discovering molecular interactions and genetic studies, independently or in a synchronized fashion. Transcriptional studies are one of the most important applications of microarrays. Measuring the entire repertoire of transcripts in an organism or cell lines provides a wealth of information about the link between the DNA and the cellular phenotype. The development of gene expression arrays has helped enormously in this direction. To study the transcriptome comprising multiple transcripts expressed simultaneously in a cell or tissue, for example, about 300,000 RNA molecules in a human cell, is a challenge for biologists [4]. In this regard, microarrays, whether using full-length genes or short oligonucleotide arrays, constitute an extraordinary tool for parallel and systematic analysis.

Protein microarrays are also important in studying protein expression quantification. In this regard, tissue microarrays are one type of technology being used for protein expression patterns in situ (more details in Chapter 2). In addition to gene and protein expression studies, the use of microarrays in clinical studies has explored novel uses of this technology in contemporary medical sciences. Identification of single-nucleotide polymorphism (SNP), disease staging and diagnosis, drug screening, and toxicological studies are only a few to name. In particular, genes have been successfully used for detecting driver genes in cancerology (more details in Chapter 12). Other applications are discussed in detail in Chapter 2.

Although different shapes, resolutions and features of microarrays vary substantially across platforms and in respect of techniques and applications, a typical microarray can be defined as a small glass slide with a plurality of spots or elements in which biochemical reactions take place. The microarray reveals qualitative information about gene or protein expression, or other biological activity, under specific conditions of an experiment. The word microarray comes from mikro (small) and arayer (arranged). A microarray has a rectangular shape and measures a few cm long—it is typically ordered, microscopic, planar and specifically coated with a suitable substrate [5].

There are two main technologies for producing microarrays. The first one is known as spotted arrays, which are produced by robotic spotting or by an inkjet printer. This technology is referred to as cDNA\(^1\) microarrays or DNA microarrays. Oligonucleotide arrays, on the other hand, are produced by employing mainly

---

\(^1\)cDNA refers to “complementary” or “copy” DNA produced during the reverse transcription of RNA in the samples and which is aimed to hybridize the DNA probes printed on the slides. The term “DNA microarrays” used through this book includes cDNA, EST and genomic DNA microarrays.
one of the following approaches [6, 7, 8]: photolithographic (Affymetrix array, aka Affymetrix GeneChip), inkjet technology (Agilent), electrochemical synthesis (CombiMatrix), solid state (NimbleGen), and silica beads in microwells (Illumina arrays, aka Illumina BeadArrays). More details about these technologies are discussed in Chapter 2.

In a typical DNA microarray, the DNA molecules are selected from a library of DNA clones in the case of spotted arrays, whereas in the case of oligonucleotide arrays, the sequences are selected from known genes and then synthesized in the form of short oligonucleotides on the slides. The DNA printed on a spot of the array is known as a probe, since each spot has a known DNA sequence. The precise location of each spot on an array is known and is recorded for image processing. The detection is carried out by hybridization reaction using a labeled sample, usually by Cy3 and Cy5 fluorophores, green and red channels respectively. The labeled nucleic acid molecules (cDNA) are produced by reverse transcription from the mRNA isolated from cells, tissues or organisms. The hybridization with the probes on a microarray helps identify the presence of target complementary sequences in a given sample. Washing the array after this process eliminates any irrelevant and unhybridized molecules and prepares the slide for scanning. High resolution scanning allows devices to detect fluorescent dyes in different channels, typically green and red, which produces images for different channels or conditions, namely experiment versus control, or even normal versus diseased sample. A typical microarray scanner uses two lasers to detect Cy3 and Cy5 wavelengths to produce green and red images of the same array. In addition, the nucleic acid molecules are usually printed in duplicates or triplicates so that any differences during the hybridization procedure can be detected and normalized.

Generally speaking, microarrays are prepared by spot printing DNA fragments or cDNA clones, which are selected to represent the expressed genome of a given cell. The clones may be available as expressed sequence tags (EST) from any specific expression library. EST data is available from the widely used National Center for Biotechnology Information (NCBI). Microarrays are prepared by printing amplified PCR products using ESTs or from nucleic acid molecules isolated from the cells or tissues. The amplified products are suspended in appropriate buffers and are spot printed onto poly-L-lysine- or aminosilane-coated glass microscopic slides using a high-speed robotic system. Alternatively, microarrays are prepared by synthesizing short oligonucleotides on the coated glass slides themselves by chemical reactions on the substrate present on the slides.

Throughout the design and execution of gene expression profiling via a microarray experiment, the key element in the expression analysis is the RNA isolated from a cell, tissue or an organism (Figure 1.3). This isolated RNA is used as a template for reverse transcription to make cDNA for comparative hybridization against the probes that are present on the microarrays. Total RNA is extracted from the biological samples, experimental and control, which may further be fractionated to purify poly A+ mRNA if needed. The use of poly A+ mRNA further adds to the efficiency of reverse transcription and hence, leads to an improved quality and specificity of
the hybridization reaction. Reverse transcription yields cDNA, which is labeled with two fluorescent labels, such as Cy3 and Cy5. Both of these fluorophores emit fluorescence at different wavelengths and can be detected during the scanning of the hybridized microarrays. The difference in the hybridization between the two fluorophores is representative of the prevalence of the transcripts in the hybridizing samples, one being the control and the other being the experimental. A difference in the intensity of specific fluorescence is due to relative amounts of a specific transcript in the two samples. Therefore, each probe representing a different gene is evaluated for its relative expression under control and experimental conditions in a comparative manner.

1.5 ANALYSIS OF MICROARRAY DATA

Microarray data analysis is an essential part in microarray studies, since it provides the right steps toward an interpretation of the experiments and derivation of meaningful, and hopefully the best possible, biological conclusions. Data analysis tools for microarrays are derived from a variety of bioinformatics methods and techniques that bridge the gap among molecular biology, computer science, statistics, and mathematics. It is not just as simple as running a software tool on a computer, but a well-designed, thoroughly-validated technique using algorithms and principles of the aforementioned disciplines. In particular, computer science does not study the way in which a particular software tool works, but the way in which an abstract computing model is designed, implemented and validated on a real computer to deliver the desired results as efficiently as possible. This section, and subsequently this book, provides a variety of well-studied and optimally designed methods for such analyses.

Although the approaches for analyses of different microarray technologies and types differ from each other, a generic approach for microarray data analysis is discussed here. The particulars for each type of technology are discussed for each case separately and with a sufficient level of specificity, as appropriate. The main process of microarray data analysis starts with the result of the scanning of the glass slides or chips, which produce images in different formats.

Scanning the slides at a very high resolution produces images, which depending on the type are composed of sub-grids of spots in DNA microarrays or squared probe cells in Affymetrix arrays [6]. Image processing and data analysis are the most important aspects in microarray studies. This is so, because the aim of the whole experimental procedure is to obtain meaningful biological conclusions. The analysis is carried out in a sequence of well-defined steps, which attempt to produce the most reliable results. Thus, any error in one of the stages will propagate to the subsequent stages yielding inaccuracies, and possibly misconceptions, about the answers to the relevant biological questions.

A schematic view of the process of microarray data analysis, wherein the images are obtained from scanning to obtain biological and functional analysis, is depicted in Figure 1.4. Scanning a microarray at very high resolution delivers a microarray image in a specific format. The image processing step aims at transforming the image by a series of steps into a more meaningful image with various parameters identified,
such as locations of spots or probe cells, background, signal and noise. Processed images are analyzed further in order to quantify the signals corresponding to gene expression. Noise, signal intensity, printing artifacts, shape and size of the spots, and lightning conditions may impose some bias in reading the gene expression pattern at different parts of the image, at different times or across different cell lines. The first two stages of the analysis (image processing and quantification) are different for

FIGURE 1.3 (SEE COLOR INSERT.) Schematic view of microarray hybridization assay and analysis for comparison of two samples (e.g., experiment versus control).
DNA microarrays, and oligonucleotide arrays or Affymetrix. The DNA microarray images are first processed by a segmentation algorithm in order to separate spots from background noise. For the sake of simplicity, a single DNA microarray image is shown in Figure 1.5a, and it must be noted that typical experiments generate at least two images for the green and red channels, which are then combined in the quantification and analysis steps. Oligonucleotide arrays are typically processed by proprietary algorithms associated with the technology being used. The normalization step aims at “erasing” this bias by bringing the data into “normal” conditions. Normalized data can be analyzed in many different ways, and in this sense, machine learning algorithms for clustering, data mining, or supervised classification aim at discovering relevant patterns in the data in such a way that these patterns are more easily understood by scientists when interpreting the data. An important step in the analysis is how to visualize those patterns or even the raw data—visualization techniques help obtain a better insight of the data and analyze it from different perspectives. The final step in the analysis involves functional and network analysis in terms of gene ontology, genomic features and inferring different types of biological networks.

1.5.1 MICROARRAY IMAGE PROCESSING

During the image processing and analysis, an image is considered to be a matrix of values that contains the pixel intensities. Two fundamental properties of an image are important in image processing: resolution and color depth or pixel depth [7, 9]. The resolution of an image is the number of pixels per unit of measurement. Typically, the term resolution is referred to as the number of horizontal and vertical pixels of an image, disregarding the unit of measurement. For example, a typical picture whose resolution is $640 \times 480$ may show an entire city or a single house—in the latter case the image shows much more details about the picture. In microarray images, the resolution refers to the number of pixels per spot in a DNA microarray image, or pixels per square probe cell in an oligonucleotide array. The color depth of an image is defined as the number of possible values (colors or grayscale values) the intensity of a pixel can take, and is usually given by the number of bits used to store each pixel. Thus, an 8-bit grayscale image will use 8 bits for each pixel, which contains one of 256 possible values (0 to 255), whereas a 24-bit color image uses 24 bits (8 bits for each channel of the RGB palette), which contains one of 16,777,216 possible values. The latter images are also known as true color images [9].

The format of an image is an important aspect in its representation and storage. Due to the size of the images in their primitive format, they are usually compressed before being stored. Compression can be lossless or lossy, in which case the original image may or may not be completely recovered. In the latter case, although there is some loss of information, the retrieved image has an acceptable quality. Most microarray images representing one channel are stored in tagged image file format (TIFF), which uses a lossless compression algorithm, while composite or two-channel images are stored in JPEG file interchange format (JFIF, aka JPEG), which involves lossy compression [9]. Compression of microarray images is a grow-
FIGURE 1.4 (SEE COLOR INSERT.) Schematic view of processing the microarray data and its analysis.
ing and exciting field that allows efficient storage and management of microarray images and data. A detailed discussion about the most recent methods is presented in Chapter 8.

The format, resolution, and steps involved in the microarray image processing stage vary, depending on the technology being used. The DNA microarray images have typically lower resolution than the Affymetrix or Illumina BeadArray images, which are produced by more powerful scanners. In terms of layout, a DNA microarray image contains rows and columns of spots separated from each other by some gap, while Affymetrix microarray images contain square probe cells adjacent to each other, each represented with a few pixels. Discussions on two types of technologies, DNA and Affymetrix, are given in the next subsections.

1.5.1.1 DNA Microarrays

DNA microarray images contain spots arranged in a grid with a certain number of rows and columns. Most images contain a certain number of sub-grids, which facilitates the location of the spots and the overall analysis. DNA array experiments are usually carried out in two channels, experimental and control, typically associated with the green and red channels. The array is scanned for each of these channels separately. In the first pass the scanner uses a laser light that detects the green dye, and in the second pass the scanner does so for the red. Since these two phases are independent from each other, two grayscale (typically 16-bit TIFF) images are produced. The example of Figure 1.5 depicts a real DNA microarray image downloaded from the Stanford Microarray Database (SMD) [10], which corresponds to a study of the global transcriptional factors for hormone treatment of Arabidopsis thaliana samples. The full image, Figure 1.5a, contains $12 \times 4 = 48$ sub-grids and was scanned at a $1,910 \times 5,550$ resolution. The spacing among different sub-grids allows specialized algorithms to identify each sub-grid, which contains $18 \times 18 = 324$ spots; the spot resolution is $24 \times 24$ pixels. One of the sub-grids is shown in Figure 1.5b. However, as shown later, noise present and different shapes for the spots make the image processing steps far from trivial.

The images from the two channels may be combined into a single color image, called the composite image. This can be achieved by superimposing the two grayscale images creating the composite image. More specifically, this can be done by simply taking the intensity of each pixel from the red channel image and placing it in the first component of the RGB triplet of the composite image. The same procedure can be applied for the green channel. The composite image will contain variations of green and red in different intensities and combinations. Portions of two DNA microarray images (green and red) are shown in Figure 1.6, along with the corresponding composite image. A green-like spot will indicate that the gene is more expressed in the experiment, while a red-like spot will indicate the gene is more expressed in the control. A black spot will represent no expression in either channel, while a yellow spot will indicate that the gene is expressed in both channels equally. Different variations of yellow (orange, more “greenish” or “reddish”) will show variations of expressions between the two channels. It is important to know
FIGURE 1.5 (a) Original DNA microarray image, 20385-ch1 (green channel), from the SMD; (b) sub-grid extracted from the 9th row and 2nd column.

that the composite image is usually used for visualization purposes only. The computational analysis is done by processing each grayscale image individually.

Image processing of DNA microarrays has its own particulars. Roughly speaking, the aim is to find the positions of the spots and then identify the pixels that represent gene expression, separating them from the background and noise. In a nutshell, the
main steps involved in processing a DNA microarray image are the following: spot addressing or gridding, segmentation, noise treatment and removal and background correction, which are discussed in more detail below.

1.5.1.1.1 Gridding

When producing DNA microarrays, many parameters are specified, such as the number and size of spots, number of sub-grids, and even their exact locations. However, many physicochemical factors produce noise, misalignment, and even deformations in the sub-grid template so that it is virtually impossible to know the exact locations of the spots after scanning, at least with the current technology, without performing complex procedures. Prior to applying the gridding process to find the locations of the spots, the sub-grids must be identified, a process that is also known as sub-gridding. Once the sub-grids are identified, the gridding step takes a sub-grid as input and aims at finding the exact location of each spot. Depending on how complex the mechanisms are, the gridding method may or may not require some parameters about the sub-grids, namely the number of rows and columns of spots, and the size of the spots in pixels, among others. Various methods have been proposed for solving this problem with some variations in terms of the amount of computer processing time, user intervention and parameters required. Among these, there are methods based on multi-level thresholding [11], machine learning [12], mathematical morphology [13], Bayesian models [14], and hill-climbing [15]. Chapters 3 and 4 present a detailed review of the existing methods and comparisons.

1.5.1.1.2 Segmentation

Image segmentation, in general, can be seen as the process of dividing the image into a set of non-overlapping regions. This process can also be seen as a classification problem in which each pixel is assigned to a specific group or class. In microarray images, pixels can be assigned to spot, background, noise, or possibly another group. Traditional segmentation techniques aimed to define a simple shape for the spot, namely circular or elliptical, are used to separate the spot from the background
or noise. Other techniques such as seeded region growing, mathematical morphology [16], clustering and pattern recognition, and Markov random fields [13] have improved these approaches by adding more features using spatial information and intensities combined in order to define arbitrary shapes for the spots, including circular, elliptical, donut-like, half-moon-like and even free-form spots. Several techniques have been proposed to counteract this problem. A detailed description of the problem and the most up to date methods of segmentation of DNA microarray images are discussed in Chapters 5 and 6.

1.5.1.1.3 Noise and Background Correction

Noise is one of the main issues in microarray experiments. It can be present even when noise in microarrays can be due to the fluctuations in a probe, target and array preparation, in the hybridization process, and in background artifacts and lightning effects in the scanning and the resulting image processing. Most of the sources of noise present in microarray images, as described in [17], can be summarized as follows:

- Probes may vary from sample to sample during mRNA preparation.
- cDNA from reverse transcription may have different lengths.
- Fluctuations may exist in labeling of probes.
- Difficulties in quantifying PCR amplification of clones.
- Pins may be affected by surface properties.
- Amounts of transported target DNA fluctuates randomly.
- Hybridization is subject to different experimental parameters.
- Cross-hybridization or nonspecificity in hybridization.
- Hybridization may vary due to physical conditions in the slide.
- Stringency conditions of post-hybridization washes.
- The amount of probe DNA attached to the slide is unknown.
- Background, lightning and radiation conditions vary at different places in the slide.

Noise present in microarray images may affect the overall process of analysis if caution is not taken mainly in those stages at the beginning of the process. Lightning variations and artifacts present in the images can hopefully be detected and eliminated by sophisticated techniques for noise detection, treatment and removal. An overview of these problems as well as a detailed discussion of the latest techniques available to date are presented in Chapter 7. Most of the noise factors discussed above, especially variations in amounts of target DNA and experimental parameters, and lightning differences across the slide, can be detected and corrected by means of well-established normalization techniques. Some of these techniques are discussed in Chapters 9 and 10. Variations in amounts of target DNA and noisy features affecting the regularity and shape of spots are usually detected by well-known segmentation mechanisms such as those discussed in Chapters 5 and 6.
1.5.1.2 Oligonucleotide Arrays

Image processing and quantification of oligonucleotide arrays depend on the technology being used. Since Affymetrix and Illumina technologies are the most widely used at present, the underlying pre-processing stages for these two technologies are briefly discussed here. Unlike DNA microarrays, Affymetrix arrays use probes that are 25-nucleotide oligomers [6]. Probes are designed as pairs of perfect match (PM) and mismatch (MM). The probes for a gene transcript are grouped into a probe set, each of them containing between 11 and 20 PM and MM. An Affymetrix GeneChip is a very small (2.54 cm × 2.54 cm) high-density chip. Scanning the chip produces an image that is composed of a matrix of cells, each cell represented by an array of pixels, typically 5 × 5 pixels. The image, stored in a DAT file, contains 4 × 4 checker-boards on the four corners, used for alignment purpose. The DAT file is a 16-bit grayscale image stored in TIFF format. The DAT file is also accompanied by a CDF file, which provides a description of the chip layout.

Image processing of an array proceeds by locating the 4 × 4 checker-boards by means of bi-linear interpolation. Once aligned, a refinement procedure allows for a more precise location of the probe cells. The segmentation proceeds by locating each probe cell, typically an array of 5 × 5 pixels, and estimating the intensity of the probe cell by using the 75th percentile of its 3 × 3 central pixels. This information and the standard deviation are stored in a CEL file. Since the Affymetrix technology is proprietary, most of the image processing and quantification algorithms are not made publicly available. Once an array is processed, the CEL file contains a summary of the actual array image. As the CEL file is normally made publicly available, many methods for post-processing of Affymetrix data have been proposed. The Bioconductor tools contain an open-source suite of tools in R [18, 19]—more details are discussed in Section 1.6.3 and in Chapter 17.

Although Affymetrix arrays are high-density and more reliable, they are not free of noisy factors and artifacts. One of the main sources of errors in these arrays is called blur, which implies that the pixels do not correspond to the actual region of hybridization. This is mainly due to the fact that the physical size of probe cells has decreased to 5 × 5 microns, increasing the resolution of the image, while being more prone to errors. There are also other errors that may be present in Affymetrix arrays, including dirt, dark and bright spots, dark clouds and shadowy circles [20], as well as blobs, lines and coffee rings, among others [21]. Other possible sources of errors that are typical in oligonucleotide arrays include cross-hybridization (probe to a transcript of another gene or mapping of the probe to an intron), alternative splicing, and single nucleotide polymorphisms [3]. More discussion and details about this technology, its image processing as well as dealing with the improvement of signal to noise ratio can be found in the references of this paragraph and in Chapters 9 and 10.

Illumina arrays differ from other types of technology mainly because they randomly arrange the probes in a hexagonal grid [8]—a typical array contains one million beads. Thus, the probes will occur a random number of times and at different random locations of the array. Although raw bead-level data from the array provides information on the locations of the beads, spatial artifacts make the image processing
step difficult. The main process involves identifying the bead locations, followed by finding the bead types, to finally obtain the intensities. Once the images are scanned, pre-processing usually proceeds in three steps: registration of beads (using a bead detection algorithm), interpolation of the remaining beads, and centering the grid over the array. A final decoding procedure allows the user to identify the probe attached to each bead [22]. After the image processing step is completed, the foreground is well separated from the background.

1.5.1.3 Protein Microarrays

Protein microarrays are rather a new addition to the technology, and do not offer many advances that could have been made in regards to specific methods for image processing and analysis. The format and layout of the protein array images are similar to those of DNA microarrays. That is, they contain spots arranged in one or more sub-grids. Currently, the tools used for spot finding, segmentation and quantification of protein array images are essentially those used in DNA microarray image processing [23]. While it is an emerging field, new studies are being conducted so that more specific analytical techniques can be used for this type of technology.

1.5.2 QUANTIFICATION AND NORMALIZATION

The main purpose of quantification is to obtain a numeric value representing the expression of a particular gene transcript as acquired from the signal from a hybridized spot during the scanning of the microarray. In the case of DNA microarrays, each spot is usually associated with a gene, and hence the pixels identified in the segmentation procedure are combined in order to obtain a unique value for the expression. There are various methods for combining these pixel intensities, offering their own advantages and disadvantages. Mean, median, and mode of pixel intensities are the most widely used ones, the median being the most preferred one, since it is more resistant to the effect of outliers [7]. Other measures are volume of intensities and total signal intensities. But they may over/under estimate the actual expression due to the differences in image resolutions or brightness conditions.

Quantification of Affymetrix array images, and in general in oligonucleotide microarrays, differs from DNA microarrays in the sense that each gene transcript is represented by a probe set. Since each probe set contains various probes in terms of PM and MM, these quantities are combined in order to obtain a single value representing the expression of a gene. A commonly used value for this is the average of the differences between PM and MM over all probes in the probe set [6, 7]. Since there may be substantial variations among probes in the cell, more information can help quantification and hence analysis, such as the standard deviation.

As discussed earlier, noise, artifacts and lightning conditions may impose some bias in reading the gene expression levels at different parts of the image, at different times or across cell lines. Normalization is an important step in pre-processing microarray data, and its main purpose is to erase these biases by bringing the data into normal conditions. There are a large number of methods that have been ap-
plied to normalization of microarray data, some of them being specific to the type of microarray or any other technology being used. The most well-known methods for normalization include LOESS/LOWESS regression, B-spline smoothing, wavelet smoothing, kernel regression and support vector regression [24].

Normalization of oligonucleotide arrays has its own particulars. More discussion, methods and comparisons for Affymetrix arrays and Illumina arrays can be found in [8, 25, 26, 27, 28] and in Chapters 9 and 10. On the other hand, quantification normalization of protein array data has its own intricacies. One of the problems present in protein arrays is the variability in the quantity of material deposited on the chip. Then, higher concentrations in some spots would produce higher absolute signals. Using specific quantification and normalization approaches such as concentration dependent analysis helps alleviate these drawbacks [23]. The tendencies of the antibodies to interact with protein molecules non-specifically is also another major factor in the normalization process to account for any difference between the actual and the deduced results.

1.5.3 PATTERN ANALYSIS AND MACHINE LEARNING

Machine learning is one of the most exciting fields for data analysis that gathers knowledge from computer science, mathematics, statistics, and engineering. The main goal of machine learning is to learn patterns from data, observations and experience, and design efficient mechanisms for prediction, pattern analysis or decision making [29]. Once microarray data has been obtained as a result of image processing, quantification and normalization, the aim of this stage is to find relevant patterns of behavior in differentially expressed genes under different conditions, cell lines or organisms.

Machine learning for pattern analysis and recognition can be subdivided into two main areas, supervised or unsupervised learning [30]. In supervised learning, patterns, categories or classes are already known, and the aim is to find relevant features or descriptors and classification mechanisms to improve predictions, which help in disease classification, diagnosis, prognosis, staging, treatment and therapeutic responses [31].

Pattern classification and prediction have been successfully applied to microarray data analysis since their early stages. The most widely used classification techniques for microarrays include the well-known and widely used support vector machine [32], linear classifiers and dimensionality reduction [30, 33], distance-based classifiers or nearest neighbor approaches [30], and neural networks [34]. In the pattern classification context, feature selection, in particular gene selection, is a problem that has been applied to microarray data analysis. Gene expression studies typically involve thousands of genes, e.g., more than ten thousand for human transcriptome studies. Gene selection aims to find a small subset of genes that are differentially expressed and provide relevant information on disease classification and other tasks.

Feature selection methods can be grouped into two categories, namely wrapper methods and filter-based approaches [30]. Some of the recent methods include correlation-based filters [34], combination of multiple filters and wrappers [35], gene
selection via sample weighting [36], and mutual information theoretic approaches [37], to name a few.

In an unsupervised learning scenario, classes and patterns are not known and have to be discovered. In the microarray analysis context, it mostly includes the application of clustering techniques for pattern finding. Clustering techniques have been used in microarray data analysis for a long time. Roughly speaking, the aim of clustering is to group the data into clusters or groups in such a way that genes or samples are similar or close to others in the same group and dissimilar to or far from samples in a different group [38]. Clustering is done in two ways, flat or hierarchical. In the latter, a hierarchy of classes is discovered and can be clearly visualized by means of a dendrogram. Various methods have been proposed for clustering microarray data, including hierarchical models [39], spectral clustering [40], $k$-means and expectation maximization [40, 41, 42], and graph and spectral clustering [43]. Advanced clustering approaches on more challenging problems include biclustering and co-clustering techniques for microarray data [44]. An extended coverage of existing approaches for clustering microarray data is given in Chapters 13 and 14.

1.5.4 TIME SERIES AND HIGHER DIMENSIONS

Time-series microarray data involves gene expression measurements at different time points. These points can occur at fixed or variable interval in times. The aim of these studies is to observe the trends of gene expression patterns across time, which is useful for clinical studies, embryogenesis, development, and gene-gene interactions. Clustering time-series microarray data is a challenging problem that involves its own particular constraints. Various methods have been proposed recently for clustering microarray time series, which either define a similarity measure appropriate for temporal expression data or pre-process the data in such a way that the temporal relationships are taken into account. The most well-known methods for clustering time-series microarray data include pairwise and multiple profile alignment [40], variation-based co-expression detection [45], Granger causality test analysis [43], and others.

Temporal information about gene expression data can be considered as a second dimensional feature, while the first dimension represents the genes themselves. Adding more information and data, such as different patients in a microarray study, cell lines or different conditions of analysis and treatment involves a higher dimensional representation. Thus, three (or higher) dimensional microarray data analysis is an evolving field that formalizes such a model [46]. More details and discussions about the models and techniques for analysis of time-series and high-dimensional microarray data are presented in Chapter 13.

1.5.5 VISUALIZATION

Visualization of microarray data and patterns discovered are very important in data analysis and mostly in interpretation of results. The intricacies of these problems lie in the fact that the data is typically high-dimensional and involves a large number of genes, typically in thousands. Many visualization techniques have been used and
proposed for microarray data. Of these, heatmaps are the most widely used tool for visualization [47], which aims to show data in a color-based matrix using colors typically varying in a scale from red to green for up- or down-regulated genes, respectively. Other techniques for visualization used for microarray data are profile plots, mainly used for time-series microarray data, and scatter plots, which aim at plotting data points in a two or three dimensional space showing expression profiles and revealing useful visual information for clustering. Scatter plots can be obtained using different techniques for data and pattern analysis such as principal and independent component analysis (PCA) [30], linear discriminant analysis and dimensionality reduction [48], independent component analysis, multi-dimensional scaling, and self-organizing maps [30]. Other statistical data analysis and visualization tools that are useful for microarray data visualization are box plots, and pie charts, among others [7]. In addition, the integration of visualization techniques and data clustering poses various challenges and shows its own advantages, especially for three dimensional data. Points, clouds and explore are useful tools for microarray data visualization [48]. More details about tools and techniques for visualization of microarray data are discussed in Chapter 16.

1.5.6 FUNCTIONAL AND NETWORK ANALYSIS

Once analyzed and visualized, microarray data provides a wealth of information for further analysis and discovery of new relationships among genes, transcripts, proteins and metabolites. One of the next tasks involves finding pathways or a network in which the differentially expressed genes may have important roles in growth and development, embryogenesis, normal cellular processes, diseases, and other related conditions. Network analysis combined with visualization may also show how up-regulated transcripts explain the up- or down-regulation of other genes, which may or may not actually interact [47]. Detecting such relationships at a large scale involves the study of gene regulatory networks, protein-protein interaction networks, pathway analysis, and metabolic network modeling. These relationships also arise from studying gene expression relationships under different experimental conditions, such as time points or variable cell lines. In this regard, relevant patterns may reveal new links among genes at different conditions, and with specific interacting proteins, RNA, carbohydrates, and other molecules. Starting from a list of relevant genes under different conditions arranged in terms of some patterns, the challenge is to infer a network that represents these relationships. Various techniques have been proposed for inferring gene regulatory networks from microarray time-series and other data [49]. An extensive coverage of methods for discovering gene regulatory networks from microarray data is presented in Chapter 15.

1.6 RESOURCES FOR MICROARRAYS

Since the microarray technology was introduced, a large number of resources have been made available. These resources help researchers in a number of ways and forms by facilitating the availability of microarray data for further analysis and pro-
viding tools for storage, retrieval, annotation, pre-processing, analysis and visualization. This section discusses the most important resources available for microarray data, including publications, databases and repositories, Web applications, and software tools.

1.6.1 PUBLICATIONS

Microarrays have been used by many researchers in many studies within the area of natural sciences. To mention a few facts, as of January 2013, searching the word “microarray” in PubMed would bring more than 51,000 publications, while typing the key word “array” lists more than 120,000 abstracts. Various databases and resources contain a wealth of microarray data for studies of gene expression, proteomics, and tissue analysis, among others. The Gene Expression Omnibus (GEO) database of the National Center for Integrative Biomedical Informatics (NCBI) of the United States contains a wealth of gene expression data and is publicly accessible via the Internet [50]. As of January 2013, the GEO contained more than 870,000 samples, the majority of which are associated with the use of microarray technology. There are many other databases and resources for microarrays as well—a more detailed discussion follows at the end of this chapter.

Presented here are the results of some queries performed in PubMed, which comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books [51], as of January 2013. Figure 1.7 shows a chart of the number of publications per year. These statistics were gathered by simply searching the database using “microarray” as a keyword. Sorting the publications by year shows a steady, small number at earlier stages and a rapidly growing trend after year 2000, with a steady growth in the past few years. This shows that although newer technologies have been recently introduced, microarray technology continues to be a growing field of research.

The list of these publications by considering the type of organism involved in the microarray study is shown in Table 1.1. These publications have also been searched in PubMed using “microarray” as a keyword and the scientific or common name for the organism. As observed in the table, microarray studies on human beings take up more than 63% of the publications. Studies on rat and mouse are reported in more than 25% of the publications. S. cerevisiae and E. coli are also two very well studied organisms, while others are studied to a minor extent, such as A. thaliana, C. elegans, Drosophila, and D. rerio.

Searching publications from another perspective shows interesting trends, too. The charts in Figure 1.8 show how publications group by microarray type and manufacturer. Since microarrays have been invented, DNA arrays were the dominant technology. However, at present, oligonucleotide arrays are dominating in the microarray studies, being employed in more than 64% of the publications, while DNA microarrays have been losing terrain and now have 23% of the share. The use of protein and tissue arrays has also grown in the past few years, accumulating 7% and 5% of the publications, respectively. Inspecting the publications by manufacturers shows that the Affymetrix platform continues to be the most widely used technology with
FIGURE 1.7  Microarray publications in PubMed, grouped by year of publication.

TABLE 1.1
Microarray publications in PubMed, grouped by organism.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of publications</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em></td>
<td>31,266</td>
<td>63.26%</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>9,876</td>
<td>19.98%</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>3,618</td>
<td>7.32%</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>1,930</td>
<td>3.91%</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>1,071</td>
<td>2.17%</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>669</td>
<td>1.35%</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>344</td>
<td>0.70%</td>
</tr>
<tr>
<td><em>Drosophila</em></td>
<td>326</td>
<td>0.66%</td>
</tr>
<tr>
<td><em>Danio rerio</em></td>
<td>321</td>
<td>0.65%</td>
</tr>
<tr>
<td>Total</td>
<td>49,556</td>
<td></td>
</tr>
</tbody>
</table>
FIGURE 1.8 Microarray publications per type and technology being used.

73% of the share. To some extent, the use of Agilent arrays has also been growing in the past few years, and by the time of this survey with a share of 18%, followed by Illumina arrays with 8% of the share. To a minor extent, less than 1% of the publications involve Combimatrix and Nimblegen arrays. It is clear that Affymetrix, which is one of the pioneering technologies in microarray studies, will continue to be the dominant technology used in microarray studies.

1.6.2 DATABASES AND REPOSITORIES

With the advent of microarray technology, large amounts of data are being generated. The increase in the volume of data and the need for making these data publicly available have made possible the creation of a large number of databases and public repositories. Of the many repositories that have been created and maintained, the most widely used ones are discussed below. The corresponding Web sites and references for all these resources are listed in Table 1.2.

SMD/PUMAdb: The Stanford Microarray Database (SMD) is the most traditional databases for microarray and gene expression data [10]. It started as an effort to maintain microarray data from ongoing research at Stanford University, and to facilitate the public dissemination of that data once published, or released by the researchers. The SMD has been one of the most important public repositories for many years. As of October 2011, the SMD contained microarray data for more than 82,000 experiments. It has recently been transferred to the Princeton University Microarray Database (PUMAdb). The PUMAdb contains microarray data in raw, pre-processed and normalized format, as well as tools for data storage, retrieval, transformation, quality assessment, analysis, and visualization. As of January 2013, the PUMAdb contained microarray data for 36,592 hybridization experiments for 48 different organisms.

GEO: The Gene Expression Omnibus (GEO) is a public repository that was established by the National Center for Biotechnology Information (NCBI) of the United States, in the early 2000s, and it is still one of the most important sources for gene
### TABLE 1.2
Resources and tools for analysis of microarray data.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Web site</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Databases and Repositories</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUMAdb</td>
<td>puma.princeton.edu/</td>
<td>[10]</td>
</tr>
<tr>
<td>ArrayExpress</td>
<td><a href="http://www.ebi.ac.uk/arrayexpress/">www.ebi.ac.uk/arrayexpress/</a></td>
<td>[52]</td>
</tr>
<tr>
<td>PLEXdb</td>
<td><a href="http://www.plexdb.org/">www.plexdb.org/</a></td>
<td>[53]</td>
</tr>
<tr>
<td>GXD</td>
<td><a href="http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml">www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml</a></td>
<td>[54]</td>
</tr>
<tr>
<td>PRKdb</td>
<td>www2.cancer.ucl.ac.uk/Parkinson_Db2/</td>
<td>[55]</td>
</tr>
<tr>
<td>GENT</td>
<td>medical-genome.kribb.re.kr/GENT/</td>
<td>[56]</td>
</tr>
<tr>
<td>FlyTED</td>
<td><a href="http://www.fly-ted.org/">www.fly-ted.org/</a></td>
<td>[57]</td>
</tr>
<tr>
<td>EMAGE</td>
<td><a href="http://www.emouseatlas.org/emage/">www.emouseatlas.org/emage/</a></td>
<td>[58]</td>
</tr>
<tr>
<td>FFGED</td>
<td>bioinfo.townsend.yale.edu/</td>
<td>[59]</td>
</tr>
<tr>
<td><strong>Tools for Microarray Data Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bioconductor</td>
<td><a href="http://www.bioconductor.org/">www.bioconductor.org/</a></td>
<td>[19]</td>
</tr>
<tr>
<td>geNorm</td>
<td><a href="http://www.proteinatlas.org/">www.proteinatlas.org/</a></td>
<td>[60]</td>
</tr>
<tr>
<td>GIP</td>
<td>compbio.dfci.harvard.edu/tgi/</td>
<td>[5]</td>
</tr>
<tr>
<td>ArrayTrack</td>
<td><a href="http://www.fda.gov/ArrayTrack/">www.fda.gov/ArrayTrack/</a></td>
<td>[61]</td>
</tr>
<tr>
<td>chipD</td>
<td>chipd.uwbacter.org/</td>
<td>[62]</td>
</tr>
<tr>
<td>cancerMA</td>
<td><a href="http://www.cancerma.org/uk/">www.cancerma.org/uk/</a></td>
<td>[63]</td>
</tr>
<tr>
<td>inCroMap</td>
<td><a href="http://www.cogsys.cs.uni-tuebingen.de/software/InCroMAP/">www.cogsys.cs.uni-tuebingen.de/software/InCroMAP/</a></td>
<td>[64]</td>
</tr>
<tr>
<td>MADTools</td>
<td><a href="http://www.madtools.org/">www.madtools.org/</a></td>
<td>[65]</td>
</tr>
<tr>
<td>GSEA</td>
<td><a href="http://www.broadinstitute.org/gsea/index.jsp/">www.broadinstitute.org/gsea/index.jsp/</a></td>
<td>[66]</td>
</tr>
</tbody>
</table>
expression data in different forms [50]. The GEO archives and provides access to microarray, next-generation sequencing and other forms of high-throughput functional genomic data, which are submitted by the scientific community via its Web site. The GEO also provides various tools for data retrieval, analysis and means for finding gene expression patterns. Due to the increasing use and versatility of applications of microarrays for different types of research projects, the GEO also stores microarray data for other than gene-expression studies, including examination of genome copy number variation and genome-wide profiling of DNA-binding proteins. As of January 2013, the GEO contained public data for 11,018 platforms, 872,170 samples in 2,720 different datasets.

ArrayExpress is one of the main public repositories for gene expression data, and has existed since 2003 at the European Bioinformatics Institute [52]. It contains gene expression data generated by microarray or sequencing technologies. The ArrayExpress database is currently integrated with the GEO, by allowing users to import GEO-formatted data into ArrayExpress. In addition, most GEO-deposited experimental data have been imported into ArrayExpress by means of high-throughput algorithms. In addition, ArrayExpress is integrated with the Gene Expression Atlas and the sequence databases at the European Bioinformatics Institute. As of January 2013, ArrayExpress contained data for 35,682 experiments on 1,030,151 assays.

In addition to the three main repositories for microarray and gene expression data, a number of smaller databases have been created and maintained. Most of these repositories target a specific organism, disease or taxonomy. While the list of these public and private repositories is rather extensive, the most well-known and recently introduced specific repositories are include here (more information in Table 1.2). PLEXdb is a public repository that contains gene expression resources for plants and plant pathogens, mostly rice and Arabidopsis [53], while the gene expression database (GXD) is a resource for mouse developmental gene expression information [54]. The Parkinson’s disease gene expression database [55], ParkDB, contains datasets of previously-analyzed, curated and annotated microarray data for analysis. The gene expression database of normal and tumor tissues (GENT) is a repository that contains more than 40,000 samples for gene expression patterns for human cancer and normal tissues [56]. FlyTED is the Drosophila testis gene expression database that contains more than 2,700 gene expression images for studies of the fruit fly, Drosophila melanogaster [57]. EMAGE, the mouse embryo spatial gene expression database, is a freely available online repository that contains gene expression patterns for studies of developing mouse embryo [58]. The filamentous fungal gene expression database (FFGED) is a public repository that contains gene expression data for filamentous fungal studies [59].

1.6.3 TOOLS FOR MICROARRAY DATA ANALYSIS
Since microarrays were introduced, a large number of analytical software tools for pre-processing, data analysis and visualization of gene expression profiles have been proposed. Many of these tools are freely available for downloading and/or use in Web-based servers, while some others are licensed or associated with specific types
of microarray platforms. Some tools are combined with databases and repositories in the aim of facilitating the retrieval, analysis and visualization of microarray data. The most widely used tools for microarray data analysis are discussed below.

Bioconductor happens to be one of the most widely used tool suites in bioinformatics [19]. It is an open source platform of tools developed in the R statistical programming language, whose main aim is to provide solutions for analysis and understanding of high-throughput genomic data. Bioconductor has a special set of tools for analysis of microarray data including various platforms, such as Affymetrix, Illumina, Nimblegen, and Agilent, amongst the rest. It provides packages for analysis of expression arrays at different stages, including pre-processing, quality assessment, differential expression, clustering, classification, and visualization. Since its introduction, a number of new Bioconductor tools and packages for microarray data analysis have been introduced. The list would be so extensive that it would not be possible to enumerate all the tools in the present discussion. As a matter of fact, as of January 2013, there are more than 300 publications associated with Bioconductor studies on microarray or gene expression data. Discussions on specific Bioconductor packages for microarray data analysis are presented in Chapter 17.

The Protein Atlas and geNorm is another tool that uses an algorithm to compare the level of expression of housekeeping genes from a set of tested reference genes to provide optimal normalization for the microarray data [60]. The Protein Atlas, which is also a Web-based tool, permits viewing of protein expression profiles based on immunohistochemistry from a large number of human tissues, cancers, and cell lines. The enormous amount of protein data allows researchers to interpret their microarray data in view of information obtained from the subcellular localization of proteins. Further, it is possible to associate the data with certain tissue types, or cancer cells, thereby identifying the complete profile of the cells used in the study. More focused approaches can help in determining the role of subcellular components of the cell in the development and progression of specific tumor types.

The Gene Index Project (GIP) was designed to use the DNA sequence databases (EST, genomic DNA, and gene encoding sequences) to construct catalogues of genomes of reference model systems [5]. The isoforms, variants, mutants, alleles and derivatives are annotated with respect to their functional roles and with regard to their protein products. In addition, the catalogued genes are also associated with their potential cellular pathways in some model organisms. This project assists researchers in finding genes and their variants in genomic sequences with their possible functional significance, using a comparative approach.

ArrayTrack is a bioinformatics tool developed at the Food and Drug Administration (FDA) of the United States [61]. It includes an environment for microarray data processing, analysis and visualization, and incorporates statistical, pathway and gene ontology analysis. In addition, the Gene Expression Omnibus (GEO) also provides a set of tools for querying, analyzing and visualizing gene expression datasets in its database [50].

Other Web tools for analysis of microarray data include chipD, a tool for designing oligonucleotide probes in high-density tiling arrays [62], cancerMA, a Web tool
for analysis of cancer microarray data [63], inCroMap, an integrated tool for analysis of microarray and pathway data [64], MADTools, a set of tools for storage, analysis and annotation of DNA microarray data [65], and GSEA, a computational method that allows the determination of statistical significance of the set of genes whose expression has been characterized on a genomic scale [66].

1.7 CONCLUSIONS AND FUTURE TRENDS

Microarrays have proven to be essential biotechnological tools for monitoring biomolecules’ behaviors in cells, tissues, and organisms in a massive and parallel way. Since their introduction, many different types of technologies have been developed to introduce the versatile approaches that can be employed in biological studies at the molecular level, including gene and protein expression, genetic and epigenetic research, tissue analysis and genotyping for paternity or forensic analysis. The ability of the microarray technology to be used in diverse types of applications allows scientists to focus on many aspects of biological and medical sciences with a wide range of applications in pharmacology, medical genetics, disease diagnosis/prognosis, detection, classification, staging and treatment of diseases, drug efficacy, evolutionary studies, and genotypic characterization of selected organisms.

This chapter reviews the main aspects of microarray technologies, principles, fundamentals, production, data analysis, as well as an introduction to a wide range of resources available. Also included in this chapter are discussions of the modern concept of central dogma in view of its analysis through microarray methodology. Therefore, the first introductory part reviewed the main aspects of molecular biology and genomics and the main building blocks, nucleotides, amino acids, DNA, RNA, and proteins. The fundamental processes in the cellular processes, namely transcription and translation, in which biochemical transformations from DNA to RNA, and ultimately to protein molecules and the characterization of their molecular interactions are explored by the use of this high throughput technology. Microarray studies have had a rich and well-established history since their introduction in the early 1990s, starting from simple arrays with a few genomic markers, evolving to extremely high resolution slides with more than two million biomarkers available in today’s technology.

Pre-processing and analysis of the underlying data are fundamental aspects in microarray studies, while at the same time, they offer big challenges to scientists in computational sciences. Although some of the initial stages, e.g., image processing and quantification, are highly dependent upon the type of technology being used, the main streamline for the analysis is common for most types of microarrays. Initial stages in the analysis, namely addressing, segmentation and quantification, involves spot or cell localization and identification of true signals, separating them from the slide background or noise. The efficiency of early stages in the analysis is crucial in the overall process, since any error at any of these stages is propagated to subsequent steps. Normalization, pattern analysis and machine learning, visualization, as well as functional and network analysis, provide a wealth of tools and techniques with sufficient flexibility for different types of microarray technologies.
A wide range of resources are available for microarray studies providing free online access to data, functional annotation, and data analysis. At present, oligonucleotide arrays have become the most preferred and affordable technology, while protein and tissue microarrays are still emerging tools for microarray studies. The survey presented in this chapter shows a consistent and increasing interest in microarray studies in terms of publications, databases and repositories, as well as software and Web-based tools for efficient analysis. Although there are three main repositories for microarray data, there are a large number of new databases for microarray studies, most of them focusing on specific organisms or diseases.

Overall, the chapter presents the basic aspects of microarray technology in a superficial way, targeting the general reader. The subsequent chapters cover a wide range of topics in microarray studies with sufficient level of detail for general and advanced readers, providing an up-to-date coverage outlined by worldwide experts in the respective fields. The subsequent chapters are aimed to achieve a wide-range coverage of the most important aspects in contemporary microarray studies and their applications, with detailed discussions on image processing, quantification, normalization, data management and storage, machine learning, visualization, software tools, and functional analysis, all of them independently addressing particular technologies.

ACKNOWLEDGMENTS

The authors would like to thank Iman Rezaeian for his help in preparing the LATEX version of this chapter, and Manoj Gajjarapu for his help in gathering the statistics for the plots and tables.

REFERENCES


27. R. Moffitt, Q. Yin-Goen, T. Stokes, et al., “caCORRECT2: improving the accuracy and


45. Z.-X. Yin and J.-H. Chiang, “Novel algorithm for coexpression detection in time-varying microarray datasets,” *IEEE/ACM Transactions on Computational Biology and Bioinform-


defective intracellular processes in a very quick and elegant manner. Fingerprints or gene signatures can also help diagnosis and staging of particular epidemic diseases, across animal and plant species. The applications are limited only by imagination, cost, and accessibility. As the latter two improve, adoption of microarrays will no doubt play a role in the emergence of new paradigms, innovations, and technologies.

ACKNOWLEDGMENTS

The author would like to thank Iman Rezaeian for his help in preparing the \LaTeX\ version of this chapter, and the anonymous reviewers for their positive input that helped enhance the chapter. The author would also like to thank Dr. Mark Schena, Arrayit Corporation (ARYC), and Affymetrix for providing the images for Figures 2.1 and 2.2.

REFERENCES


Biological Aspects: Types and Applications of Microarrays


82. T. Lettieri, “Recent applications of dna microarray technology to toxicology and eco-


ate the corresponding grid for each sub-grid, while other methods use sophisticated mechanisms such as a genetic algorithm or the hill-climbing approach to identify spot locations.

Most of these methods use a few parameters that the user can adjust to obtain the best results for each dataset. While this flexibility may be a benefit in some cases, generally, most methods depend on a particular dataset. This makes these models rather difficult to be used as a general method for gridding microarray images regardless of the specific properties of each microarray image. OMTG is among the few methods that uses almost no parameters and is able to perform gridding on various DNA microarray images, achieving a high accuracy and without setting parameters individually for each image or dataset. The $M^3G$ method also performs the gridding accurately even in the presence of noise and artifacts, while taking into account rotations in the input image.

While most of the methods presented in this chapter perform very well, there are several aspects that can be improved, and which open various avenues for future research. Using more sophisticated methods with a minimum number of parameters can make the gridding phase even more accurate and flexible. Some of the modules in these methods can be replaced with more sophisticated and efficient algorithms that can boost the overall performance and accuracy of the gridding step and, in turn, the biological analysis.

ACKNOWLEDGMENTS

The authors would like to thank the support from NSERC, the Natural Sciences and Engineering Research Council of Canada, and the anonymous reviewers for their valuable feedback on the chapter. The authors would also like to thank Dr. L. Ramdas for providing the cDNA microarray images for the dilution experiments, the DILN dataset.

REFERENCES

27. D. Bariamis, D. Maroulis, and D. Iakovidis, “$M^3$G: maximum margin microarray grid-


FIGURE 4.13 (a) A microarray image rotated by 15°, (b) the counter-rotated image ($\Delta \theta = 0.9^\circ$), and (c) the resulting gridding for this image [2].

$M^3G$ involves several preprocessing components, including Radon-based rotation estimation for the microarray image, as well as spot detection and selection. The distance between rows and columns of spots is then estimated and the positions of the selected spots are used to train a set of linear soft-margin SVMs.

The use of soft-margin SVMs reinforces robustness to outliers that result from artifacts and noise, whereas the use of redundant vectors in the SVM training set and the automatic determination of the operating parameters facilitate a substantial increase in gridding accuracy. In addition, $M^3G$ copes with the presence of irregular and weakly expressed spots, as well as rotation. A potential weakness of $M^3G$ is that the SVM classifiers require several detected spots in each row and column of spots. Rarely, most of the spots in a row or column might be weakly expressed and not detected. In such cases, which account for less than 0.1% of the rows and columns in the data set, the grid line positioning is determined by the nearest grid lines. The experimental evaluation on standard DNA microarray images showed that $M^3G$ outperforms state-of-the-art gridding methods, such as [11] and [14], providing the potential of achieving nearly optimum gridding. A future perspective is the integration of $M^3G$ with subsequent segmentation and quantification methods in the context of a user-friendly GUI, in order to facilitate everyday use by biologists. Another research path is to study the use of alternative filters for spot detection, aiming to enhance detection accuracy. Furthermore, the effect of alternative SVM kernels is worth investigating.

REFERENCES

TABLE 5.2
Comparison of DNA microarray image segmentation approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Segmentation Quality</th>
<th>Level of automation</th>
<th>Features/Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circular-Shape</td>
<td>Low</td>
<td>High</td>
<td>Limited to circular spots</td>
</tr>
<tr>
<td>Region-growing</td>
<td>High</td>
<td>Low</td>
<td>Requires foreground and background seed points</td>
</tr>
<tr>
<td>Thresholding</td>
<td>Low</td>
<td>Requires a mask</td>
<td>Independent of the shape of a spot</td>
</tr>
<tr>
<td>Clustering</td>
<td>High</td>
<td>High</td>
<td>Pixels are divided into three groups: foreground, background, and deletions</td>
</tr>
<tr>
<td>Snake</td>
<td>High</td>
<td>Low</td>
<td>Computationally expensive</td>
</tr>
<tr>
<td>Shape</td>
<td>High</td>
<td>High</td>
<td>Assumes that the spots are symmetric</td>
</tr>
<tr>
<td>Graph</td>
<td>High</td>
<td>High</td>
<td>Computationally expensive</td>
</tr>
<tr>
<td>Morphology</td>
<td>High</td>
<td>High</td>
<td>Insensitive to changes in imaging and calibration</td>
</tr>
<tr>
<td>Model</td>
<td>High</td>
<td>High</td>
<td>Assumes that the spots are symmetric</td>
</tr>
<tr>
<td>Watersheds</td>
<td>High</td>
<td>High</td>
<td>Might over-segment the image</td>
</tr>
<tr>
<td>Supervised</td>
<td>High</td>
<td>High</td>
<td>Requires already segmented images for training</td>
</tr>
<tr>
<td>Learning-based</td>
<td>High</td>
<td>High</td>
<td></td>
</tr>
</tbody>
</table>

In Table 5.2, we summarize important information for each method, including: 1) segmentation quality reported (might be based on different measures and datasets), 2) the level of automation and the need of parameter adjustment, and 3) associated features and/or limitations.

With the availability of more powerful computational resources, more advanced techniques will be developed for DNA image segmentation. Learning-based segmentation methods should also be considered for DNA image segmentation. In [29] a preliminary work based on neural networks and used for segmentation of DNA microarray images was proposed. However, more research is necessary in the field to assess the efficacy of the method.

ACKNOWLEDGMENTS

The author would like to thank Dr. Nader Karimi from Isfahan University of Technology for contributing some of the images.

REFERENCES


REFERENCES

40. A. Chao, T. H. Wang, Y. S. Lee, et al., “Molecular characterization of adenocarcinoma and squamous carcinoma of the uterine cervix using microarray analysis of gene expres-


REFERENCES

39. V. Aurich and J. Weule, “Non-linear Gaussian filters performing edge preserving diffu-
57. J. Astola and P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering*. Boca Raton,


modified JPEG2000-compliant coder. These performance enhancements range from 0.213 bpp to 0.918 bpp, i.e., from 1.97% to 15.53%, and situates JPEG2000 as the overall best image compression standard for microarray images, with bitrates only 0.788 to 1.266 bpp larger than the best results claimed in the literature.

Lossy compression does not guarantee exact image fidelity, but can produce almost arbitrary compressed file sizes. Several lossy methods for microarray images have been proposed, but none of them provides an objective way of measuring the relevancy of the compression loss whose results can be extrapolated to new analysis techniques. Traditional image distortion metrics like PSNR or SSIM are not suitable for this task. Thus, new metrics specific for microarray images must be devised. The microarray distortion metric (MDM) has been proposed. Its design is based on the image features that are used in any image analysis algorithm. Experimental results suggest that this metric is suitable for the task of measuring the loss of relevant information, and is able to differentiate important and unimportant distortions in microarray images.

In the foreseeable future, it is not likely that great advances can be achieved for the lossless compression of DNA microarray images. Small improvements may be obtained by combining and refining existing approaches like pixel prediction and per-bitplane context-based compression. However, all data gathered in Section 8.3 suggest that lossless compression ratios of 3:1 are not going to be obtained anytime soon. On the other hand, there is a large margin for improvement for the lossy compression of microarray images. New distortion metrics like the MDM defined in Section 8.4 are currently being researched and validated. Once a consensus is reached and some of these metrics are generally accepted by biologists and physicians, the attention of the microarray image coding community will likely be drawn to lossy compression. Therefore, great advances in the microarray image compression performance will probably be observed.

8.6 ACKNOWLEDGMENTS

The MicroZip image set was kindly provided by Neves and Pinho from the University of Aveiro. The Arizona image set was provided by David Galbraith and Megan Sweeney from the University of Arizona. The research leading to this chapter has been partially funded by the European Union, by the Spanish Ministry of Economy and Competitiveness (MINECO), and the Catalan Government under projects FP7-PEOPLE-2009-IIF FP7-250420, FPU AP2010-0172, TIN2009-14426-C02-01, TIN2012-38102-C03-03 (LIFE-VISION), and 2009-SGR-1224.

REFERENCES


5. SIB Computational Genomic Group, “ISREC microarray image set.”


8. David Galbraith Laboratory, “Arizona microarray image set.”


should not be overlooked by data analysts.

Microarray grid alignment is one of the basic processing steps of microarray images. Understanding this step is critical for performing an optimal microarray data analysis and obtaining conclusions from every experiment. There are several sources of error during the grid alignment process that could cause the information stored in the CEL file from the DAT file to be erroneous, and thus affect the data analysis workflow. Therefore, the grid alignment process should not be overlooked in the analysis of Affymetrix arrays.

Every oligonucleotide array contains spatial flaws to some degree [34], and efforts should be directed to evaluate and assess the reproducibility of microarray data (e.g., [39]). The most common flaws appear to be the consequence of trapped bubbles and are manifested as rings or arcs [13], irregular shaped blobs, or occasional “scratches” [2]. These are usually seen towards a side of the array [8].

As noted in [31], several computational tools (e.g., RMA) whose purpose is to assess overall chip quality and improve the accuracy of microarray gene expression fail to incorporate adequate spatial information into their outlier detection methods or normalization routines. The image processing algorithms described in this chapter seek to address these deficiencies and complement existing methods to improve the accuracy of gene expression measurements and the implications biologists draw from them. According to the authors of [35], it is usually not clear how to proceed once spatial artifacts have been detected, and the researcher must decide if the contaminated chip is to be included or excluded from further analysis. However, the decision of excluding information from an experiment could be cost-intensive.

The identification of spatial biases using image processing techniques is a topic often neglected in microarray data analysis. A possible explanation for this is that several scientists regard this problem as part of the manufacturing stage rather than as part of the data analysis stage. In other words, scientists often ignore the problem because they believe that the manufacturers have taken appropriate measures to prevent the existence of blemishes. Since this is not always the case, we suggest both encouraging microarray manufacturers to improve their quality protocols to prevent the incorporation of these biases, and incorporating image processing techniques and spatial normalization algorithms in the microarray data analysis work flow so that these algorithms are applied routinely.

REFERENCES


22. F. Naef and M. O. Magnusco, “Solving the riddle of the bright mismatches: Labelling
array and the spatial nature of most artifacts can be taken into account. This approach was also superior to MBR as well as Harshlight, which in general appeared to be too stringent on probe quality as it also flagged many probes on arrays without apparent defects.

It should be noted here that the results of this evaluation are independent of the particular summarization method used later to calculate probeset values, as methods were applied to the individual probe values before summarization. Furthermore, performance of the approaches was evaluated in terms of the number of spiked probes identified correctly and not the fold-changes of the summarized probeset values. However, probe values were normalized before applying artifact detection using quantile normalization. Thus, we cannot completely exclude the possibility that the relative performance of the methods might be different for a different normalization method. Nevertheless, as quantile normalization is one of the most commonly used normalization methods, this evaluation is likely representative for most standard applications.

In summary, these results illustrate the importance of properly addressing artifacts in microarray analysis. Furthermore, they show that quite simple methods already perform very well for this purpose and can outperform more complicated approaches. The main challenge is now not to develop even more sophisticated methods but to establish artifact detection and correction methods as an integral part of microarray expression analysis. Furthermore, as microarray analysis is more and more replaced by RNA sequencing (RNA-seq) due to rapidly decreasing costs of sequencing, it will be important to establish similar quality control procedures for RNA-seq. Although the artifacts will likely be different in this case, the importance of addressing such artifacts will remain the same.

REFERENCES


REFERENCES


67. D. A. Clunie, “DICOM structured reporting and cancer clinical trials results,” Cancer


Cancer, the Ligue Nationale Contre le Cancer (label DB), and the Institut National de la Santé et de la Recherche Médicale. Support for the computational infrastructure was obtained from a Fondation pour la Recherche Médicale grant. Maxime Garcia is funded by the Institut National de la Santé et de la Recherche Médicale—Région Provence-Alpes Cote d’Azur Fellowship. Support for Raphaèle Millat-Carus was obtained from the Institut National du Cancer Grant. Thanks to Sabrina Carpentier (Ipsogen, Marseille, France) for helpful discussions on the original ITI method, and Wahiba Gherraby for proofreading the manuscript.

REFERENCES


49. E. Segal, M. Shapira, A. Regev, et al., “Module networks: identifying regulatory mod-


between \(g_{j_1}, t_{j_2}\), and \(s_{j_3}, \mathcal{X} = \begin{bmatrix} \mathcal{C}, \mathcal{G}, \mathcal{T}, \mathcal{S} \end{bmatrix}\) can be matricized into

\[
\begin{align*}
X_{(1)} & \approx GC_{(1)}(S \otimes T)^T \\
X_{(2)} & \approx TC_{(2)}(S \otimes G)^T \\
X_{(3)} & \approx SC_{(3)}(T \otimes G)^T
\end{align*}
\] (13.49)

Generally speaking, there are no constraints on the core tensor and mode matrices in Tucker decomposition. However, constraints such as orthogonality, non-negativity, and non-Gaussianity can be enforced in a decomposition algorithm. For instance, HOSVD enforces the orthogonality constraints on the mode matrices and is among the most popular Tucker algorithms. It calculates the left singular matrices for different matrices in different modes as factors. The core tensor is obtained through \(\mathcal{C} = \mathcal{X} \times_1 \mathcal{G}^T \times_2 \mathcal{T}^T \times_3 \mathcal{S}^T\). Interested readers are referred to [51] for more details.

Higher-order orthogonal iterations (HOOI) is an alternating least squares (ALS) algorithm initialized by HOSVD, which gives better decomposition than HOSVD itself (see [52] and [48] for details). HOOI also generates orthogonal mode matrices. HONMF imposes non-negativity constraints on the core tensor and mode matrices. Multiplicative updates rules [53] corresponding to core and mode matrices have been extended in [21] for HONMF. The core tensor and mode matrices are alternatingly updated until the convergence criteria are met. The authors of [53] have observed that good interpretation and learning performance can be benefited by adding non-negativity and sparsity constraints to matrix factorization. Even though it can be imposed and controlled, sparsity is sometimes a by-product of non-negativity constrained matrix (maybe tensor also) factorization without explicit sparsity constraint.

REFERENCES


37. B. Savas, “Analyses and tests of handwritten digit recognition algorithms,” Master’s
TABLE 14.8
Most significant shared GO terms (process, function, component) PDNS for biclusters on Yeast Cell-Cycle dataset.

<table>
<thead>
<tr>
<th>Biclusters</th>
<th>Biological Process</th>
<th>Molecular function</th>
<th>Cellular component</th>
</tr>
</thead>
<tbody>
<tr>
<td>78 genes;</td>
<td>translation</td>
<td>structural constituent of ribosome</td>
<td>cytosolic ribosome</td>
</tr>
<tr>
<td>9 conditions</td>
<td></td>
<td>(58.1%, 8.71e-37)</td>
<td>(53.00%, 5.97e-70)</td>
</tr>
<tr>
<td>83 genes;</td>
<td>nucleic acid metabolic process (34.0%,</td>
<td>phosphatase regulator activity (1.7%,</td>
<td>nucleus</td>
</tr>
<tr>
<td>11 conditions</td>
<td>2.45e-11)</td>
<td>0.00041)</td>
<td></td>
</tr>
</tbody>
</table>

The performance of the presented algorithms are evaluated on three DNA microarray datasets. We have tested the biological significance using a gene annotation tool to show that these works are able to produce biologically relevant biclusters.

Given the diversity of microarray data and evaluation criteria, it seems that no single biclustering approach dominates all the other approaches. The existing biclustering tools have specific advantages and limitations and could be jointly applied. Moreover, given that large and complex microarray data become available, it is useful to devise more powerful bicluster methods. One way to achieve this is to make the search method more intelligent by integrating various specific knowledge into the search mechanisms and operators.

REFERENCES
### TABLE 14.9
Most significant shared GO terms (process, function, component) EvoBic for biclusters on Yeast Cell-Cycle.

<table>
<thead>
<tr>
<th>Biclusters</th>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Cellular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>2527 genes; 6 conditions</td>
<td>cytoplasmic translation (2.57%; 2.10e-4)</td>
<td>structural molecule activity (6.60%; 9.7e-3)</td>
<td>replication fork (1.60%; 4.67e-7)</td>
</tr>
<tr>
<td></td>
<td>response to stimulus (11.40%; 4.90e-6)</td>
<td></td>
<td>nuclear replication fork (1.40%; 9.10e-7)</td>
</tr>
<tr>
<td></td>
<td>cellular response to stimulus (9.66%; 2.22e-5)</td>
<td></td>
<td>non-membrane-bounded organelle (20.60%; 1.20e-6)</td>
</tr>
<tr>
<td>20 genes; 3 conditions</td>
<td>nucleoside transport (20%; 6.30e-4)</td>
<td>nucleobase-containing compound transmembrane transporter activity (30%; 4.44e-6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>purine-containing compound transmembrane transport (20%; 1.51e-3)</td>
<td>nucleobase transmembrane transporter activity (20%; 1.40e-4)</td>
<td></td>
</tr>
</tbody>
</table>


43. J. Liu, Z. Li, F. Liu, and Y. Chen, “Multi-objective particle swarm optimization biclus-


ment and internal validation of the networks produced, it remains an art to choose the best model. Since simpler models are easier to interpret and less prone to overfitting, when two models provide equally good performances for explaining the observations, it is recommended to use the simpler model. “Everything should be made as simple as possible, but not simpler” (Albert Einstein).

So far no single model is error-free and no single inference method has optimal performance for all data sets [72]. It is a natural idea to combine different models. Marbach et al. [72] showed that the integrated results of multiple inference methods have robust performance across different data sets. Vignes et al. [73] proposed a meta-analysis that integrates the inferred results of three different methods. Li et al. [74] proposed a method called differential equation-based local dynamic Bayesian network (DELDBN), which integrates differential equations and the DBN model. They solve the differential equations and infer dependence by identifying the Markov blanket of target variables.

With the rapid accumulation of microarray data and biological knowledge, as well as the development of other high-throughput techniques, integrating data and knowledge from different sources is important to improve inference accuracy. Wang et al. [75] proposed a method based on linear programming and SVD to find the most consistent network structure with respect to multiple microarray data sets and the sparseness of the network. Chen et al. [76] integrated Bayesian prior from the epigenetic data of histone modification, and applied DBN to infer GRN from time series gene expression data. Wang et al. [77] proposed a GRN inference method using genome-wide fitness data from knockout libraries, which is an emerging high-throughput data type.

In conclusion, to generate better and useful hypotheses, scientists need a GRN with finer resolution. So far, no single inference method can achieve this goal and there is always lack of data. Combining the advantages of different inference models and integrating data and knowledge from different sources are promising directions for future studies.

15.6 FURTHER READING

For more information, please refer to survey papers on gene regulatory network reconstruction [3, 20, 33, 78, 79, 80, 81, 82, 83].

REFERENCES


Reconstruction of Regulatory Networks from Microarray Data


other sources of data such as pathways, ontologies, and external and previously published data can augment primary microarray experiment data, making it more complex, but in many ways easier to comprehend. We should not neglect the domain expertise that can play a major role in microarray data analysis since only a domain expert may be able to determine whether identified patterns or lists of genes carry biological significance, and has the capacity to test them in a wet lab setting. In any case, microarrays proved to be an indispensable technology for studies related to the behavior of genes under various conditions. The techniques described above represent the basis for the visual analysis of microarray experiment data and are an important step of its general analysis. At the same time, they serve as a stepping stone for further utilization with advanced and novel visual techniques in the rapidly evolving field of biomedical visualization.

In the future, the field of biomedical visualization will continue to focus on the development of tools as well as integration of tools and data from various sources and modes under the umbrella known as “visual analytics” [74]. The principles of visual analytics will advance the field of biomedical visualization by integrating analytical reasoning, interaction, data transformation and representation into novel, yet to be developed tools [75]. Since the effective analysis of microarray and related data relies on collaboration of analysts and domain experts, these efforts will focus on the development of collaborative tools that will facilitate exploration of the data by entities that are potentially not co-located. This means that the tools of the future will make heavy use of hardware such as multi-core CPUs, cloud storage and computing resources and capable GPUs as well as networking and distributed databases, requiring well-defined mechanisms for integration of disparate tools and sources of data. Overall, the future of biomedical visualization is bright and will get even brighter as new technologies, requiring sophisticated tools, are introduced.

ACKNOWLEDGMENTS

We would like to thank Phillip C.S.R. Kilgore for his assistance and for proofreading this manuscript.

REFERENCES


REFERENCES


ing differential expression in microarray experiments,” *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.