

## Chapter 1

# Introduction

### 1.1 The Microarray: Key to Functional Genomics and Systems Biology

The new field of bioinformatics employs computer databases and data mining algorithms to analyze proteins, genes, and complete collections of deoxyribonucleic acid (DNA) on a genome-wide level [222]. As defined by National Institutes of Health (NIH), bioinformatics consists of “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data.”

The rapid development of biological technologies over the last decade has resulted in the complete sequencing of many important model organisms. These include *Haemophilus influenzae* (the first sequenced genome of a free-living organism, 1,830,137 base pairs [100]), *Saccharomyces cerevisiae* (the first eukaryotic genome sequence with 12,068 kilo bases [112]), *Caenorhabditis elegans* (the first complete sequence of a multicellular organism [279]), *Drosophila melanogaster* (120 mega bases [1]), *Arabidopsis thaliana* (an important plant model [195]), and more than 30 microbial genomes. By 2001, the first draft version of the sequence of base pairs in human DNA had been released (human chromosome 22 in 1999 [89] and chromosome 21 in 2000 [127]). We are now moving from the *pre-genomic era* characterized by the effort to sequence the human genome, to a *post-genomic era* that concentrates on harvesting the fruits hidden in the genomic text.

An overarching challenge in this post-genomic era is the management and analysis of enormous quantities of sequence data. The aim of any organizational scheme would be to provide biologists with an inventory of all

genes used to assemble a living creature, analogous to the Periodic Table of chemistry [172]. Understanding the biological systems with tens or hundreds of thousands of genes will similarly require organizing the constituents by properties and will reflect similarities at diverse levels such as:

- Time and place of RNA expression during development;
- Subcellular localization and intermolecular interaction of protein products;
- Physiological response and disease;
- Primary DNA sequence in coding and regulatory regions; and
- Polymorphic variation within a species or subgroup [173].

Recently, the advent of microarray technology has made it possible to monitor the expression levels of thousands of genes in parallel. Arrays offer the first promising tool for addressing the challenges of the post-genomic era, by providing a systematic way to survey variation in DNA and RNA. There have been widespread applications of this technology during the last several years, and it seems likely to become a standard tool both of research in molecular biology and of clinical diagnostics.

## 1.2 Applications of Microarray

Microarrays have already been extensively used in biological research to address a wide variety of questions. As noted by Collins [61], when applied to expression analysis, this approach facilitates the measurement of RNA levels for the complete set of transcripts of an organism. When applied to genotyping, microarrays usher in the possibility of determining alleles at hundreds of thousands of loci from hundreds of DNA samples, allowing the contemplation of whole genome-association studies to determine the genetic contribution to complex polygenic disorders. Moreover, the application of microarrays to mutation screening of disease genes with pronounced allelic heterogeneity is likely to move the possibility of genetic testing for disease susceptibility of individuals, or even entire populations, into the realm of practical reality. To motivate our subsequent discussion, we shall begin by presenting a few examples of some related research. We emphasize that this discussion is by no means exhaustive and in fact represents only a fraction of the universe of potential applications.

### 1.2.1 *Gene Expression Profiles in Different Tissues*

Cells from different tissues perform different functions. Although they can be easily distinguished by their phenotypes, a detailed understanding of the mechanisms of these different behaviors remains elusive [10]. Since cell function is determined by individual proteins and protein synthesis is dependent on which genes are expressed by the cell, the expression pattern of a gene provides indirect information about cell function. For example, a gene expressed only in the kidney is unlikely to be directly involved in the pathology of schizophrenia [67]. Microarray experiments can be used to identify those genes which are preferentially expressed in various tissues. This would enable scientists to gain valuable insights into the mechanisms that govern the functioning of genes and cells [10]. Moreover, the highly selective tissue expression of a drug target is attractive as a means to reduce the potential for unwanted side effects [10].

### 1.2.2 *Developmental Genetics*

Amaratunga and Cabrera [10] described the application of microarrays to developmental genetics. The genes in an organism's genome express differentially at different stages of its developmental process. Interestingly, it has been found that there is a subset of genes involved in early development that is used and reused at different stages in the development of the organism, generally in different order in different tissues, with each tissue having its own combination. Crucial to these processes are growth factors, which can also, later in an organism's development, be involved in causing or promoting cancer; these genes are known as *proto-oncogenes*. Microarrays can, in principle, be used to track the changes in the organism's gene expression profile, tissue by tissue, over the series of stages of the developmental process, beginning with the embryo and up to the adult. Supplementary applications of this line of research include deducing evolutionary relationships among species and assessing the impact of environmental changes on the developmental process of an organism.

### 1.2.3 *Gene Expression Patterns in Model Systems*

Detailed profiling of gene expression in model systems (such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*) will yield valuable insights into the functions of genes and the mechanisms of important cellular processes. This type of analy-

sis has already been described using the yeast *Saccharomyces cerevisiae*. For example, Spellman and his colleagues [263] used DNA microarrays to create a comprehensive catalog of yeast genes whose transcript levels vary periodically within the cell cycle. They reported 800 genes that meet an objective minimum criterion for cell cycle regulation.

In another example of such analysis, Gasch et al. [107] explored genomic expression patterns in the yeast *Saccharomyces cerevisiae* in response to diverse environmental transitions. DNA microarrays were used to measure changes in transcript levels over time for almost every yeast gene as cells responded to a variety of environmental impacts. These included temperature shocks, hydrogen peroxide, the superoxide-generating drug menadione, the sulfhydryl-oxidizing agent diamide, the disulfide-reducing agent dithiothreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into a stationary phase. They found a large set of genes ( $\sim 900$ ) which showed a similar drastic response to almost all of these environmental changes, while additional features of the genomic responses were specialized for specific conditions.

In another study, Gasch et al. [106] used DNA microarrays to observe genomic expression in yeast *Saccharomyces cerevisiae* responding to two different DNA-damaging agents. The genome-wide expression patterns of wild-type cells and defective mutants were compared in Mec1 signaling under normal growth conditions and in response to the methylating agent methylmethane sulfonate (MMS) and ionizing radiation. The comparative study identified specific features of these gene expression responses that are dependent on the Mec1 pathway. Among the hundreds of genes whose expression was affected by Mec1p, one set of genes appeared to represent an MEC1-dependent expression signature of DNA damage, cell cycle, mutations, and stimulus.

#### 1.2.4 Differential Gene Expression Patterns in Diseases

One of the most attractive applications of microarray technology is the study of differential gene expression in disease. There are many *genetic diseases* that are the result of mutations in a gene or a set of genes. The mutations may cause genes to express inappropriately or to fail to express. For example, cancer could occur when certain regulatory genes, such as the p53 tumor suppressor gene, become always transcribed regardless of any regulatory factors [10].

Microarray experiments can be used to identify which genes are differ-

entially expressed in diseased cells versus normal cells. The up- or down-regulation of gene activity can either be the cause of the pathophysiology or the result of the disease. Although targeting disease-causing genes is desirable to achieve disease modification, interfering with genes that are expressed as a consequence of disease can lead to the alleviation of symptoms. The opportunity to compare the expression of thousands of genes between “diseased” and “normal” cells will allow the identification of multiple potential targets [67]. This would enable the development of drugs aimed directly at the difference between diseased and normal cells. Such drugs can be designed to specifically target a particular gene, protein, or signaling cascade, and they are therefore less likely to cause undesirable side effects [10].

There are abundant examples of such microarray applications in the literature. Rheumatoid tissue was analyzed using a microarray of about 100 genes known to have a role in inflammation [130]. Among others, genes encoding interleukin-6 and several matrix metalloproteinases, including matrix metallo-elastase (HME), were markedly up-regulated. The latter result was unexpected, as the distribution of HME was previously thought to be limited to alveolar macrophages and placental cells [67]. Golub et al. [114] studied the expression profiles of patients with two subtypes of leukemia, ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). The clinical diagnosis of these two subtypes of cancer is extremely difficult due to their clinical similarity. Microarray experiments can be used to identify which genes are differentially expressed in the two different types of cancer patients, thereby creating specific disease profiles on the basis of their gene expression patterns [10].

### 1.2.5 *Gene Expression Patterns in Pathogens*

As noted by Debouck [67], activity in the sequencing of bacterial genomes is intense, with a new bacterial genome seemingly sequenced in its entirety every month. The small size of these genomes will allow the easy construction of individual microarrays in which every gene from a given microbe is represented. For microbiologists, confined for years to studying bacteria one gene at a time in a test tube under artificial growth conditions, the horizons appear unlimited. Microarray technology will identify genes that are turned on in vitro but not at the site of infection in vivo, and vice versa, and those genes that are only turned on during infection in vivo. Such genes encode virulence determinants that are regulated by environmental

signals such as the transition from ambient temperature to body temperature [196]. Since traditional genetic techniques used to identify virulence genes are time consuming, they will be quickly supplanted by microarray methods. A similar approach will be used to study viral gene expression during the time course of acute infection or during latency. Microarrays can also be used to study the response of the host to challenges from the pathogen.

### ***1.2.6 Gene Expression in Response to Drug Treatments***

Microarrays are potentially powerful tools for investigating the mechanism of drug action. For example, Interferon- $\beta$  (IFN- $\beta$ ) is the most widely prescribed immunomodulatory therapy for multiple sclerosis (an autoimmune disease of the brain and spinal cord). The therapy is known to exert all its biological effects via gene transcription but there are no validated markers for its long-term efficacy in multiple sclerosis. Although double blind, randomized, placebo-controlled clinical trials have established that IFN- $\beta$  treatment reduces the progression of disability in multiple sclerosis, only 30-40% of patients respond well to the therapy. To define the mechanism of IFN- $\beta$  and investigate the partial responsiveness of various patients, the expression levels of large numbers of genes were monitored for thirteen multiple sclerosis patients during a ten-point time-series [300].

In [67], Debouck gave two other applications of high-density microarrays to examine the effects of drugs on gene expression in yeast as a model system. In one, the effect of potent kinase inhibitors was analyzed on a yeast-genome-wide scale by measuring changes in mRNA levels before and after treatment with inhibitors [115]. The second study reported a gene expression pattern (or “signature”) characteristic of the immunosuppressive drug FK506. This same signature was also observed in yeast cells carrying a null mutation in the FK506 target, establishing that genetic and pharmacological ablation of a gene function results in similar changes in gene expression. Treatment of the null mutants with FK506 also revealed additional pathways distinct from the drug’s primary target [191]. It is possible that yeast will provide a high-throughput platform for studying cellular responses to drugs. However, a similar method applied to human cells and tissues would have even more direct utility in the identification and validation of novel therapeutics.

### 1.2.7 Genotypic Analysis

Variation in DNA sequence underlies most of the differences we observe within and between species. Locating, identifying and cataloging these genotypic differences represent the first steps in relating genetic variation to phenotypic variation in both normal and diseased states [184]. Lipshutz et al. [184] described a specific type of array that can be designed for this purpose. Given a reference sequence for a region of DNA, four probes are designed to interrogate a single position. One is designed to be perfectly complementary to a short stretch of the reference sequence; the other three are identical to the first, except at the interrogation position, where one of the other three bases is substituted. Upon incubation with the reference sequence, the probe complementary to the reference sequence will obtain the highest fluorescence intensity. In the presence of a sample with a different base at the interrogation position (a substitution variant), the probe containing the complementary variant base will obtain the highest fluorescence intensity.

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome; this, and the ease with which they can be identified, recommend them for this type of analysis [184]. For example, the study by David Wang and colleagues [295] identified 3,241 candidate SNPs contained in STSs collected at the Whitehead Institute and Sanger Center and mapped 2,227 of them. Using conventional gel-based sequencing and high-density oligonucleotide arrays, a total of 2.3 million bases of sequence were screened for variations among eight individuals to identify candidate SNPs and create a third-generation genetic map for the human genome. More than two thousands of these SNPs were selected, and sets of appropriate probes were synthesized on a high-density array. The “SNP chips” are intended for commercial distribution and can be used for linkage, linkage disequilibrium and loss of heterozygosity studies. Arrays can also be used to scan the genome for new SNPs. Shrinking the area occupied by each of the interrogating oligonucleotide probes (or “synthesis features”) to approximately  $15 \times 20 \mu\text{m}$ , a total of 50,000 nucleotides (on both strands) can be screened for the presence of polymorphisms on a single array [184].

### 1.2.8 Mutation Screening of Disease Genes

In [10], Amaratunga and Cabrera described using microarrays to study *complex diseases*. Complex diseases are not caused by a few errors in ge-

netic information but by a combination of small genetic variations (polymorphisms) which predisposes an individual to a serious problem. The risk of such an individual contracting a complex disease tends to be amplified by non-genetic factors, such as environmental influences, diet and lifestyle. Coronary artery disease, multiple sclerosis, diabetes, and schizophrenia are complex diseases in which the genetic makeup of the individual plays a major role in predisposing the individual to the disease. The genetic component of these diseases is responsible for their increased prevalence within certain groups such as families, ethnic groups, geographic regions, and genders. Microarray experiments can be used to identify the genetic markers, usually a combination of SNPs, that may predispose an individual to a complex disease.

### 1.3 Framework of Microarray Data Analysis

The enormous quantity of gene expression microarray data and its significant applications in biomedicine require effective approaches to its analysis. Figure 1.1 presents a flowchart illustrating the typical components of microarray data processing and analysis. The framework consists of three major steps, determination of the biological problem and sample preparation, array generation, and data analysis. In the first step, the RNA sources are collected from the tissues of model systems or diseased/normal patients or from a cultivated homogeneous cell population as appropriate to the particular problem being investigated. RNAs are then extracted from these cells.

In the second step, a microarray experiment is carried out. Although there are different types of microarrays, all follow these common basic procedures [278]:

- *Chip manufacture:* A microarray is a small chip (made of chemically-coated glass, nylon membrane or silicon) onto which tens of thousands of DNA molecules (*probes*) are attached in fixed grids. Each grid cell relates to a DNA sequence.
- *mRNA preparation, labeling and hybridization:* Typically, two mRNA samples (a test sample and a control sample) are reverse-transcribed into cDNAs (*targets*), labeled using either fluorescent dyes or radioactive isotopes, and then hybridized with the cloned sequences on the surface of the chip.

- *Chip scanning*: Chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets.

Once the raw microarray data are obtained, several pre-processing steps may need to be performed prior to any further data analysis. These pre-processing steps include data transformation, estimation of missing values, and data normalization. After data pre-processing, the microarray data can usually be represented by a two-dimensional matrix  $\{x_{ij}\}$ , where each row  $r_i$  in the data matrix corresponds to one gene, each column  $c_j$  corresponds to each experimental condition, and each cell  $x_{ij}$  is a real value recording the expression level of gene  $i$  under condition  $j$ . Finally data analysis and visualization algorithms can be applied to the pre-processed data sets.

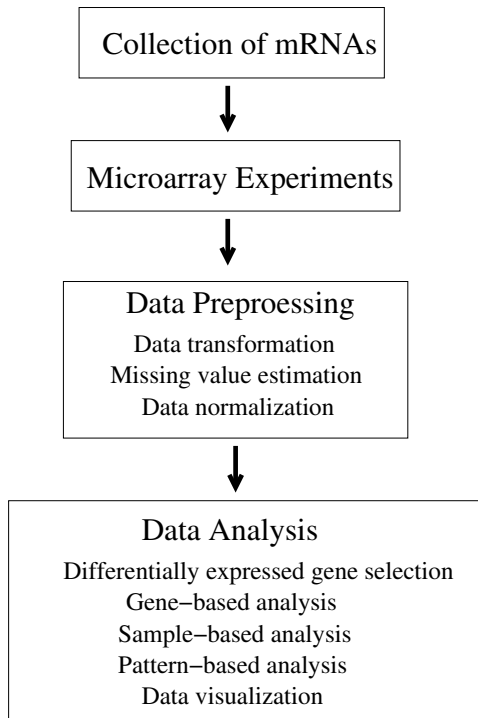


Fig. 1.1 Framework of microarray data analysis.

In this book, we will not discuss in detail the manufacture of microarray chips or the procedures involved in the microarray experimentation. Neither will we elaborate on data pre-processing or statistical approaches to

differential analysis. Instead, we refer readers to the work of Schena [241] and Speed [262] for details on these topics. This book will primarily focus on the development and application of advanced data mining, machine learning, and visualization techniques for the identification of interesting, significant, and novel patterns in microarray data.

The remainder of this book is organized as follows:

- Chapter 2 briefly introduces some concepts central to molecular biology, providing readers with an understanding of the basic mechanisms of microarray technology and the related analysis.
- Chapter 3 discusses various approaches to microarray manufacture, the basic procedures of microarray experimentation, and the pre-processing of the raw data.
- Chapter 4 explores several statistical methods for the identification of differentially expressed genes.
- Chapter 5 focuses on gene-based analysis, introducing various clustering algorithms to identify co-expressed genes and coherent patterns. In particular, an interactive approach which integrates users' domain knowledge into the clustering process is presented. Different cluster validation approaches are also reviewed.
- Chapter 6 presents approaches to sample-based analysis. Since a microarray experiment typically involves a much larger number of genes than that of samples, meaningful sample-based analysis must effectively reduce the extremely high dimensionality of the data set. We will discuss both supervised and unsupervised methods for the reduction of the dimensionality. A series of approaches to disease classification and discovery will then be surveyed.
- Chapter 7 explores methods for ascertaining the relationship between (subsets of) genes and (subsets of) samples. Several classical approaches to mining frequent itemsets, as well as a post-mining method, are introduced to identify interesting association rules among genes and samples. We will also discuss pattern-based clustering algorithms to find the coherent patterns embedded in the sub-attribute spaces. A novel approach will be presented to uncover the inherent correlation among genes, samples, and time-series.
- Chapter 8 looks at various methods for data visualization. The visualization process is intended to transform the data set from high-dimensional space into a more easily understood 2- or 3-

dimensional space.

- Chapter 9 discusses some new trends in mining gene expression microarray data. We focus on combining various data sources and integrating domain knowledge into the data mining process.
- Chapter 10 concludes the book.

## 1.4 Summary

As discussed in this chapter, effective approaches are demanded to analyze gene expression microarray data. Recently, a variety of data-mining techniques have gained acceptance for analyzing gene expression microarray data. This book is intended to provide researchers with a working knowledge of many of the advanced approaches currently available for this purpose. These approaches can be classified into five distinct categories: gene-based analysis, sample-based analysis, pattern-based analysis, visualization, and integration of domain knowledge. Chapters 5 through 9 will treat each of these approaches individually, each starting with an overview of current methods and moving to a more detailed discussion of a selected technique of particular interest. Applications of these approaches to specific gene expression microarray data sets will be discussed and experimental results will be presented. It is hoped that the discussion of new trends offered in Chapter 9 may stimulate further explorations in the field on the part of interested readers.