TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 642 • Publication 642

Matti Nykter

Signal Processing Methods and Information Approach to Systems Biology

Tampere 2006

Tampereen teknillinen yliopisto. Julkaisu 642
Tampere University of Technology. Publication 642

Matti Nykter

Signal Processing Methods and Information Approach to Systems Biology

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 9th of December 2006, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2006

ISBN 952-15-1690-9 (printed)
ISBN 952-15-1750-6 (PDF)
ISSN 1459-2045

Abstract

Recent technological advances have made it possible to observe the behavior of biological systems at the genetic level in a high-throughput manner. The ability to do measurements at the system level has made it possible to move from a traditional reductionistic approach to a more global system level approach. Thus, instead of looking at the behavior of the individual components, the goal of this new approach, the systems biology, is to understand the structural and dynamical properties of the system as a whole.

Living systems differ from non-living systems, for example, by their ability to process information from their environment and to propagate information over time through the mechanism of evolution. As information processing is a fundamental property of all living systems, we can gain insight into the system level properties by studying the information processing and flow. For example, how information is propagated through the evolution or how the system responses to a perturbation.

The content of this thesis is two-fold. In the first part we introduce new signal processing methods for the computational analysis of the biological data. The purpose of the proposed methods is to improve the reliability and the quality of the microarray data. We introduce an unsupervised approach that can be used to verify the clinically determined class labels for the samples. Next we discuss the identification and quantification of the microarray noise sources. We introduce a simulation model that can be used to simulate microarray data with realistic biological and statistical characteristics by utilizing the noise properties of real data. Finally, we discuss how supplemental measurement data can be used to improve the quality of microarray data. As a case study, we show how the cell population distribution can be estimated using fluorescent activated cell sorter data.

The second part of the thesis introduces an information-based approach for studying the complex systems. By using the Kolmogorov complexity based information measure we show how the information processing and flow in biological systems can be used to characterize their structure and behavior at the system level. We show that through the information flow, we can discover evolutionary relationships between organisms. In addition we study the information processing of an innate immunity cell macrophage and show that the dynamics of its information processing exhibit criticality.

Preface

I want to thank my supervisor Professor Olli Yli-Harja for all the support and guidance to my research work. Olli has created an excellent work environment and his continuous encouragement has motivated me along the way. I would also like to thank my friends and colleagues in our Computational Systems Biology group.

I am highly grateful to Professor Ilya Shmulevich for his guidance and deep interest to my research work. Ilya has without a doubt had a major influence to my work and his role in this dissertation can not be overemphasized. I am also indebted to Professor Wei Zhang for his guidance on cancer research. I also which to thank all my co-authors for truly pleasant and insightful collaboration.

This work has been carried out in the Institute of Signal Processing at Tampere University of Technology and partly in the Cancer Genomics Laboratory at The University of Texas M. D. Anderson Cancer Center and at the Institute for Systems Biology. Special thanks goes to all friends and colleagues in these institutes. The financial support of the Tampere University of Technology Graduate School, the Tampere Graduate School in Information Science and Engineering (TISE), the Jenny and Antti Wihuri Foundation, the Etelä-Savo Cultural Foundation, and Tekniikan edistämissäätiö are gratefully acknowledged.

I would like to express my gratitude to my parents, Anja and Markku, and to my sister Hanna for their constant support. And finally, my warm thanks goes to Ulla.

Contents

List of Publications

This thesis is based on the following publications. In the text, these publications are referred to as Publication I, Publication II, and so on.

I Nykter, M., Hunt, K. K., Pollock, R. E., El-Naggar, A. K., Taylor, E., Shmulevich, I., Yli-Harja, O., and Zhang, W. (2006) Unsupervised analysis uncovers changes in histopathologic diagnosis in supervised genomic studies. *Technology in Cancer Research and Treatment*, 5(2), 177–182.

II Nykter, M., Aho, T., Ahdesmäki, M., Ruusuvuori, P., Lehmussola, A., and Yli-Harja, O. (2006) Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, 7(349).

III Niemistö, A.,† Nykter, M.,† Aho, T., Jalovaara, H., Marjanen, K., Ahdesmäki, M., Ruusuvuori, P., Tiainen, M., Linne, M.-L., and Yli-Harja, O. (2004) Distribution estimation of synchronized budding yeast population. In *Winter International Symposium on Information and Communication Technologies*, Cancun, Mexico.

IV Harjunpää, A., Taskinen, M., Nykter, M., Karjalainen-Lindsberg, M.-L., Nyman, H., Monni, O., Hemmer, S., Yli-Harja, O., Hautaniemi, S., Meri, S., and Leppä, S. (2006) Differential gene expression in non-malignant tumour microenvironment is associated with outcome in follicular lymphoma patients treated with rituximab and CHOP. *British Journal of Haematology*, 135(1), 33–42.

V Nykter, M., Yli-Harja, O., and Shmulevich, I. (2005) Normalized compression distance for gene expression analysis. In *IEEE International workshop on Genomic Signal Processing and Statistics*, Newport, Rhode Island, USA.

VI Nykter, M., Yli-Harja, O., Shmulevich, I. (2005) The similarity metric in network analysis. In *3th International Workshop on Computational Systems Biology*, Tampere, Finland.

VII Nykter, M., Kesseli, J., Shmulevich, I., and Yli-Harja, O. (2006) Analyzing Boolean network dynamics using attractor basin structure. In *4th International Workshop on Computational Systems Biology*, Tampere, Finland.

The author of this thesis contributed to the publications as follows. As the first author of Publications I, II, V, VI, VII and joint first author of Publication III, the author designed and implemented computational methods, performed data analysis, and drafted the manuscripts as discussed in the following.

In Publication I the author performed all data analysis and wrote the manuscript, including the integration of biological information.

In Publication II the author designed the microarray simulation model and implemented the model jointly with M. Ahdesmäki. The author performed all the simulations, except the simulation of the artificial genetic regulatory network and the image processing examples. The author wrote the manuscript.

Publication III was jointly written by the author and A. Niemistö. The author designed the algorithms and performed the data analysis, excluding the image analysis part.

In Publication IV the author performed the microarray data analysis and assisted in writing the manuscript.

In Publications V, VI, VII the author designed computational methods and performed all data analysis. The author also wrote all manuscripts.

Chapter 1

Introduction

Advances in biological research have often followed technological breakthroughs. More advanced technologies have made it possible to study biological organisms in more detail and to obtain more knowledge about their properties. For example, the development of new imaging technologies played an important role in the discovery of the double helix structure of DNA (Watson and Crick, 1953). In the last ten years we have seen major advances in both the technology and knowledge. Microarray technologies have made it possible to measure the behavior of biological systems at the genetic level in a highly parallel manner by measuring the expression of all the genes simultaneously (Schena *et al.*, 1995). At the same time the genome sequencing projects have provided large amounts of information about the DNA sequences (The Genome International Sequencing Consortium, 2001; Venter *et al.*, 2001). Increase in knowledge and the availability of the high-throughput measurement technologies have made it possible to move from a reductionistic approach, where only a few components of the system are studied, to a more global system level approach (Kitano, 2002; Ideker *et al.*, 2001). In the system level approach, instead of looking at individual genes or proteins, the goal is to understand the structure and dynamics of a system as a whole.

There are two main types of information that are embedded in biological systems (Hood and Galas, 2003). All the genetic information about the building blocks of life, that is the genes, is stored in the genome. Thanks to the genome sequencing projects, this library of building blocks is available for several organisms. While the genes contain information about the molecules that appear within an organism, they do not tell us much about the system. Equivalently, a catalog of the individual components used in an airplane tells us very little about how to assemble a plane or how an airplane behaves in the air under different weather conditions (Kitano, 2002; Csete and Doyle, 2002). Thus, the other type of information, the regulatory network is in a

key role in understanding the behavior of the system. In a biological system a regulatory network determines which molecules interact with each other and how the system responds to different stimuli or perturbations.

Using microarray technology we can study how a biological system behaves at the genetic level. The system under study can be stimulated using specific drugs that make the system divert from its original behavior (Ideker *et al.*, 2001). The response of the system can then be measured by performing the microarray experiments at several time instants after the stimulus. This type of analysis can give insight into which genes are responsible for generating the response and thus makes it possible to uncover parts of the genetic regulatory network. However, this approach is not sufficient for the full characterization of the behavior of the system. Many important cellular functions happen through, for example, protein interactions. Thus, there are several regulatory networks that operate at the different levels. To gain more information, measurements should be made simultaneously at several different cellular levels (Kitano, 2002).

In addition to multiple cellular levels, networks can operate in different time scales. Some operations, like the response to a change in the environmental conditions happens quickly, typically in a few minutes, while some operations like a cell or a life cycle can take from hours to years. In addition, the genetic information also operates through the evolution in a time scale of tens to millions of years. Because of the different levels of operation and the different time scales, the problem of uncovering the regulatory structure of biological systems is extremely difficult.

Instead of trying to uncover the genetic regulatory network in detail we can look at the emergent properties of the complex networks (Kauffman, 1993, 2004). It is assumed that the properties that are observed in many large networks, such as robustness and adaptability are key elements in sustaining life (Csete and Doyle, 2002). Understanding this kind of system level properties helps us to gain insight into the living systems.

Living systems differ from non-living systems by their ability to process information from their environment and to propagate information over time through the mechanism of evolution. As the information processing is a fundamental property of all living systems, it makes it extremely attractive to analyze the systems by studying their information flow (Yockey, 2005). That is, how the information is propagated through the evolution or how a system responds to a perturbation.

While microarray and other high-throughput measurement technologies have already helped us to gain significant insight into biological systems, there are still several fundamental problems for the applicability of the measurement technologies. A major problem is the lack of the biological ground truth information. Thus, it is not possible to objectively evaluate the ob-

tained results or the algorithms that have been developed for the analysis. In addition, as microarray measurements can usually be done only by using a population of cells, the response of the individual cells can not be measured.

We will introduce the basic concepts of molecular biology in Chapter 2 and the fundamentals of microarray technology and the analysis of biological data in Chapter 3. In Chapter 4 we will discuss how signal processing methods can be used to improve the reliability of the microarray data and to verify the performance of the analysis algorithms. We will first discuss how the reliability of clinically assigned class labels can be evaluated, and we will show that there is a fair chance that the original labeling of samples done by several pathologists may not be reliable. Next, we will discuss how microarray data with realistic biological and statistical characteristics can be simulated and how the obtained data can be used for example in the experimental design or in the verification of data analysis methods. Finally, we will show how the complementary data, measured along the microarray experiment can be used to improve the quality of the microarray data.

The last part of the thesis in Chapter 5 focuses on analyzing complex systems at the system level. We will introduce an information-based approach that can be used to study the information processing and flow in the biological systems. We will show how the structure and dynamics of a system can be characterized through the information processing. We will also show that our information based approach can directly be applied to real measurement data. As an example, by using the data from a microarray experiment we will study and quantify the dynamics of an innate immunity cell macrophage.

Chapter 2

Biological Background

It is the current understanding that the thousands of genes and their products are the building blocks of living systems. Information about the organization and the function of molecular components are embedded into our genome. However, it has become clear that the dynamical behavior of living systems can not be determined solely based on the genetic information. Functional forms of the molecules, like proteins that are constructed based on the genetic information, play an important role in the dynamical behavior of systems. Thus, to understand life, the biological systems need to be studied at different cellular levels.

2.1 The Genome

Life is specified by genomes. The genome includes all the genetic information about the organism. In practice, the genetic information is stored in a deoxyribonucleic acid (DNA) sequence. DNA has a double helix structure which is constructed of individual nucleotides each containing a different base adenine (A), guanine (G), cytosine (C), or thymine (T) (Lodish *et al.*, 2001).

The size of the genome varies significantly between different organisms, ranging from only a few thousand base pairs in the single cell organisms up to billions of base pairs in the eukaryotes. The size of the human genome is about 3×10^9 base pairs (Venter *et al.*, 2001).

The genome sequence can be divided into genes and non-coding regions. A gene is considered to be a region of the DNA that codes for a protein or some other molecules, together with the promoter region that controls when the gene product is produced (Figure 2.1). Genes contains regions known as introns and exons. Exons are the part of the gene sequence that is used to

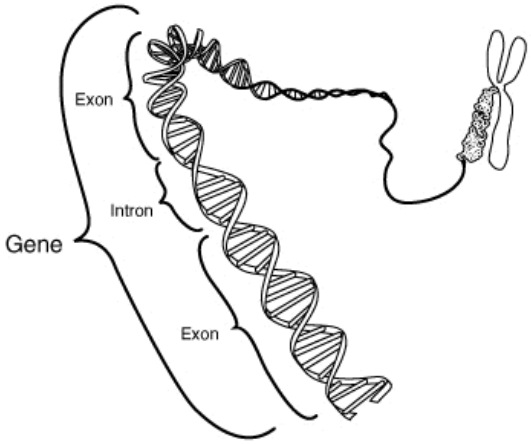

Figure 2.1: Illustration of a gene and the DNA double helix structure in a chromosome (Access Excellence at the National Health Museum, 2006).

code a gene product. Intron regions are spliced out from the sequence before gene product is coded (Lodish *et al.*, 2001). Thus, introns are a type of non-coding region that is located within genes. Only about 1% of the human genome is covered by the genes, remaining 99% being the non-coding region. According to current knowledge, the human genome includes about 25 000 genes (The Genome International Sequencing Consortium, 2004).

2.2 The Cell

Cells are the basic structural and functional units of life. Each cell contains information and the structures that are needed to sustain life (Lodish *et al.*, 2001). Genetic information, that is the DNA sequence, is stored in the chromosomes (Figure 2.1). In the eukaryotic organisms the chromosomes are located inside the nucleus of the cell[1] (Figure 2.2).

DNA sequence contains instructions of how different molecules, involved in cellular functions, are built. Before the sequence can be read the chromosome structure needs to unfold. Reading a gene from the DNA sequence is initiated by ribonucleic acid (RNA) polymerase. Once the RNA polymerase has bound to the promoter region of the gene, DNA unwinds and becomes single-stranded. A copy of one strand of the DNA is synthesized of nucleotides, containing a base adenine (A), guanine (G), cytosine (C), or uracil (U). As a result a messenger RNA (mRNA) molecule is obtained. This process is known as transcription (Lodish *et al.*, 2001).

[1] In addition, some genetic information is stored in the mitochondrial genome.

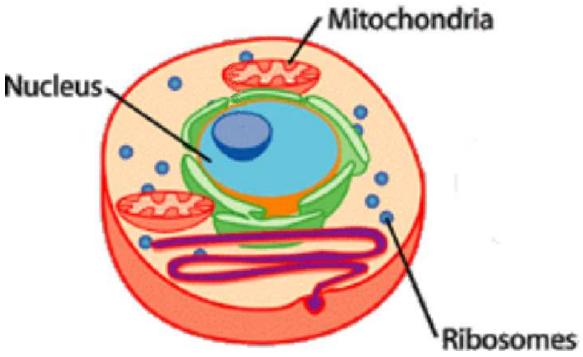

Figure 2.2: Structure of a eukaryotic cell. Different cellular components are shown, including the nucleus, ribosomes and mitochondria (Science Primer, 2006).

The mRNA then moves to another cellular organelle, the ribosome, for protein synthesis. At the ribosomes mRNA is decoded to make proteins. A specialized RNA molecule known as transfer RNA (tRNA) binds to the mRNA. Binding tRNA molecules deliver amino acids that bind together forming a long polypeptide chain. This process is known as translation (Lodish *et al.*, 2001).

Once the entire mRNA sequence has been decoded at the ribosome, the mRNA has been translated into a protein. Right after and already during the translation the protein starts to acquire its functional form by folding into its secondary and tertiary structures. In the process, post-translational modifications, including attachment of functional groups or structural changes, may occur (Figure 2.3) (Lodish *et al.*, 2001).

This process where a gene is first transcribed to form a mRNA molecule which in turn is then translated to form a protein is called the central dogma of molecular biology (Crick, 1970). While it would be convenient to assume that there is direct connection that one gene makes one protein, this is not the case. Many genes are known to have several protein products (Lodish *et al.*, 2001).

In addition to the protein synthesis, there are several other tasks that cells need to perform. For example, cells need to produce descendents by division. Cell division is performed through the cell cycle. The cell cycle can be divided in distinctive phases (Figure 2.4). Before entering the cell cycle, cells can operate in a quiescent state, usually denoted as G_0. Once the cell cycle is initiated the cell enters the first growth phase G_1. Next, the cell enters the synthesis phase S where the DNA is replicated, resulting in a

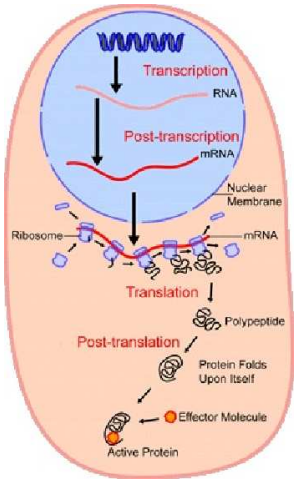

Figure 2.3: An overview of protein synthesis. The DNA sequence of a gene is first transcribed into a mRNA which is then translated into a protein (Science Primer, 2006).

copy of all genetic information. After replication the cell enters the second growth phase G_2, where the cell further grows as it prepares for division. The final phase in the cell cycle is mitosis M, the actual division of the cell.

During the cell division several control mechanisms monitor the cell for DNA damage (Lodish *et al.*, 2001). There are several "checkpoints" along the cell division process that will prevent the division if certain conditions are not met. If something goes wrong, programmed cell death called apoptosis is initiated. This is a vital control mechanism to prevent the damaged cell from dividing further. A failure to enter apoptosis has been found to be related to the emergence of several types of cancers (Vaux *et al.*, 1988; Lockshin and Zakeri, 2001).

While all cells share the same genetic material, DNA, there are several different morphological and functional forms of cells, known as cell types. Different cell types emerge from stem cells through differentiation. There are more than 200 different types of cells present in the human body, each capable to perform very different tasks (Lodish *et al.*, 2001).

2.3 Cellular Networks

As the cell is able to control which proteins are produced and to perform operations like differentiation, cell division or apoptosis, there needs to be a control mechanism for these processes (Sonenberg *et al.*, 2000). This kind of

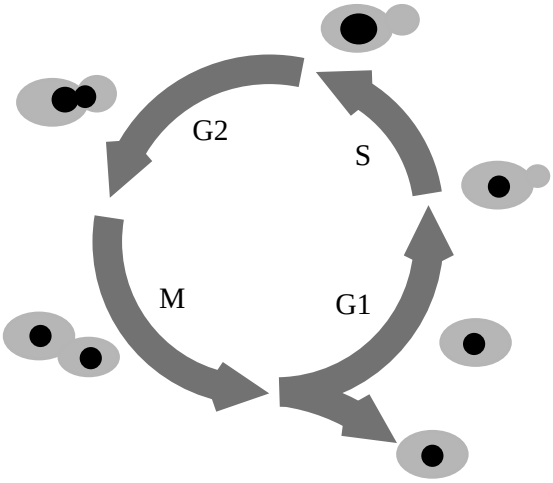

Figure 2.4: Illustration of the cell cycle of an eukaryotic cell. During the G_1 phase the cell grows. During the S phase the genetic material is replicated. Then the cell enters the second growth phase G_2 and finally the cell divides at the M phase.

control occurs through interactions between different molecules. The complexity of these interactions is a key issue when determining the complexity of an organism. A current estimate of the number of genes in the human genome is about 25 000 (The Genome International Sequencing Consortium, 2004). Several other organisms, like plants have about the same number or even more genes. Thus, the number of genes does not correlate with the observed complexity of an organism. Therefore, the complexity has to be due to the underlying regulatory mechanism that controls the amounts of gene products and their interactions (Levine and Tjian, 2003). This kind of control appears at different levels. For example, a mRNA is produced when a gene product is needed. The control signal, an indication that the mRNA needs to be produced can originate from the protein level through a protein that is coded by a different gene (Sonenberg *et al.*, 2000).

As a cell is a highly complex system there are numerous control mechanisms that operate independently of each other in a highly parallel manner. Several different subnetworks that control, for example, metabolism and transcription have been identified (Sonenberg *et al.*, 2000). Still, these independent subnetworks are dependent and closely connected to each other through the gene products that are needed in the reactions.

Regulatory mechanisms can be studied as a genetic regulatory network. Genes can be considered to be the nodes of the regulatory network and their states, or the expression levels, are regulated by the other genes or proteins (Kauffman, 1969). Control mechanisms and interactions can be studied at

different cellular levels and at different levels of detail (Bolouri and Davidson, 2002). For example, we can study the behavior of the cell at gene, protein or metabolite level, yielding genetic regulatory, protein-protein interaction or metabolic networks, respectively. Inputs of the network can be taken from different cellular levels, or the network can be modeled, for example, at the genetic level using only gene to gene interactions.

Different types of models can be used to model the behavior in different levels of detail, ranging from a simple binary model (Kauffman, 1993) to continuous differential equation models (de Jong, 2002; Bolouri and Davidson, 2002; Chen *et al.*, 2004). The selection of the model class depends on the level of detail required and the goal of the modeling. For example, if we are interested in understanding the general properties that emerge in a large complex network, Boolean networks, where each gene has the state on or off, can be sufficient models (Kauffman, 1993; Aldana *et al.*, 2003). If we want to predict how different interactions between the gene products occur, then obviously more detailed models are needed (Chen *et al.*, 2004; Lee *et al.*, 2002).

Chapter 3

Microarray Technology

Traditionally, research in molecular biology has focused on studying individual genes or proteins and their behavior under different conditions. A lot of information has been obtained using this approach. However, this traditional approach is not sufficient for understanding the system as a whole. The introduction of microarray technology has allowed us to study biological systems at the system level by looking at the expression of thousands of genes simultaneously.

A microarray is a microscopic slide containing numerous probes, that is binding sites for genetic material, in predetermined locations (Schena *et al.*, 1995). Several different types of gene expression microarrays, based on different manufacturing technologies, have been introduced (Schena *et al.*, 1995; Chee *et al.*, 1996; Hughes *et al.*, 2001). The most common type of microarray is the cDNA microarray, where complementary DNA is hybridized to the slide. Microarrays where instead of genetic material other types of samples are hybridized have also been introduced. These include, for example, tissue, lysate, and protein microarrays (Kononen *et al.*, 1998; Paweletz *et al.*, 2001; Haab *et al.*, 2001; Knezevic *et al.*, 2001). In addition, several different types of microarray chips that have been designed to find a specific type of biological information from the samples, are available (Buck and Lieb, 2004; Snijders *et al.*, 2001; Ohnishi *et al.*, 2001). For example, the ChIP-chip microarrays can be used to find transcription factors that bind to a specific gene (Ren *et al.*, 2000; Lee *et al.*, 2002). In this chapter, we focus solely on the cDNA microarray technology.

Performing a microarray experiment is a highly complex process that involves several consecutive steps (Figure 3.1). These steps include extraction and isolation of the biological sample, reverse transcription and labeling of the RNA, preparation of the microarray slide by printing the probes, hybridization of the labeled sample to the slide, reading the slide by a laser scanner, extraction of the information from the scanned image using image

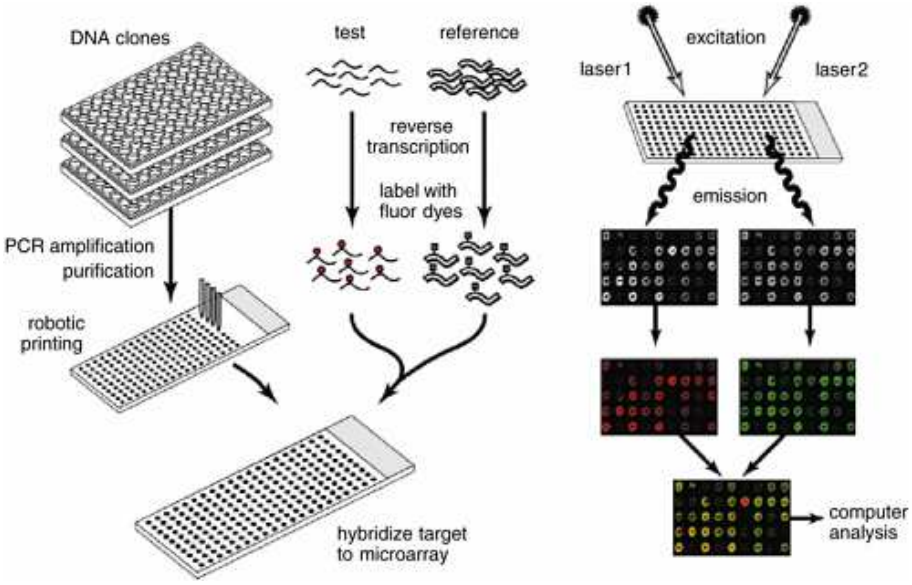

Figure 3.1: Outline of a microarray experiment (Duggan *et al.*, 1999).

processing, and finally performing computational analysis using the obtained data (Duggan *et al.*, 1999; Zhang *et al.*, 2004). Many of these steps are sensitive to different types of errors. Thus, there are several potential sources of noise and systemic bias in a microarray experiment. (Dror *et al.* 2003; Cho and Lee 2004; Publication II).

3.1 Sample Preparation

Before we can perform a microarray experiment, a biological sample needs to be obtained. Typical samples may include, for example, cancer tumors of different types (Zhang *et al.*, 2004). In addition, normal tissues can be used as reference samples. A requirement for the biological sample is that it contains enough genetic material so that a sufficient amount of RNA can be isolated (Brownstein and Khodursky, 2003). There are several well established protocols available that can be used for the RNA extraction (Coombs *et al.*, 1999).

To obtain good quality data, a very small homogeneous population of cells, or even a single cell, needs to be obtained. Recently, sophisticated technologies, such as the laser capture micro-dissection, have been developed. These technologies can be used to obtain pure samples (Emmert-Buck *et al.*, 1996). The isolated RNA sample is reverse transcribed to complementary DNA (cDNA). In many cases the amount of available cDNA from the

sample is not enough to perform a microarray experiment. In these cases, a polymerase chain reaction (PCR) amplification can be used to amplify the genetic material and obtain more cDNA (Duggan *et al.*, 1999). In the next step, the cDNA is labeled using a fluorescent dye, typically Cy3 (green) or Cy5 (red). If the purpose is to hybridize two samples to the same slide, then different samples are labelled with different dyes (Churchill, 2002).

3.2 Array Fabrication

Microarrays can be fabricated using different technologies and materials (Brownstein and Khodursky, 2003). They can be printed, for example, on a glass, plastic, or silicon slide using printing with fine pointed pins, photo lithography, or ink-jet printing (Schena *et al.*, 1995; Lipshutz *et al.*, 1999; Hughes *et al.*, 2001; Heller, 2002).

Gene expression microarrays that are based on the hybridization of the expression product can be divided into two groups depending on the type of the hybridization setup. The first type of microarrays are the *spotted* or two channel microarrays (Figure 3.2). In these arrays two samples, labeled with different dye colors, are hybridized to one slide. To prepare the slide, small fragments of the PCR products or DNA clones are printed to the slide using a robotic printing machine with fine pointed print pins (Duggan *et al.*, 1999). Recently, two channel microarrays have been prepared using ink-jet printing of the oligonucleotides (Hughes *et al.*, 2001). With this approach, a sequence corresponding to the mRNA of a specific gene is printed to the spot. With two channel microarrays we assume that a specific gene product will only bind to one spot.

The second type of arrays are the oligonucleotide or single channel microarrays (Figure 3.2). These are usually constructed by printing short oligonucleotide sequences to the slide using photo lithography (Lipshutz *et al.*, 1999). In photo lithography, specific areas on the chip are lighted. This causes chemical coupling to occur at the illuminated sites allowing the nucleotides to bind. This process is repeated several times to build up piles of nucleotides, as illustrated in Figure 3.3 (Lipshutz *et al.*, 1999).

Numerous microarrays with different types of probes are commercially available. These include, for example, the whole genome chips for several different organisms including the human. The whole genome chips contain corresponding probes for all the known genes present in a genome. In addition to readily available arrays, custom microarrays can be prepared in well equipped laboratories or ordered from commercial manufacturing companies. Thus, special array layouts can be designed for specific research problems.

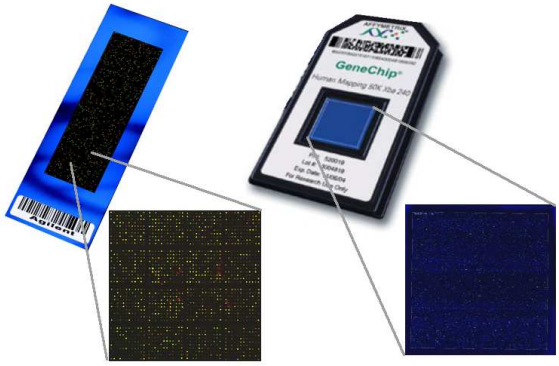

Figure 3.2: Two types of microarrays. On the left is a typical two channel microarray printed on a glass slide (Agilent Technologies, 2006). On the right, a single channel array is shown (Affymetrix, 2006). In two channel microarrays the spots are typically arranged in several subarrays whereas in single channel arrays they are in one subarray.

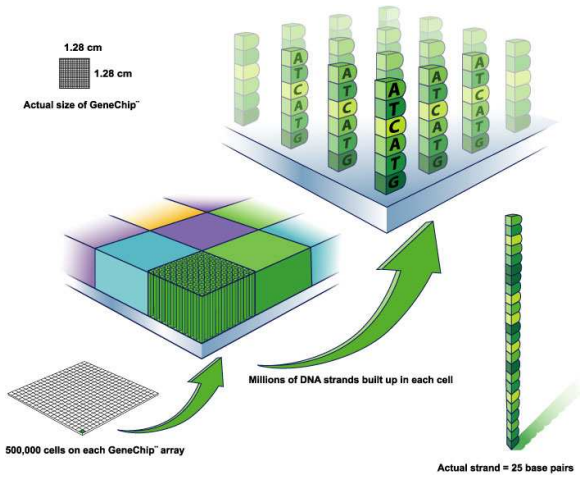

Figure 3.3: Illustration of a single channel microarray. Piles of nucleotides can be built using photo lithography. Each pile is a probe with a specific nucleotide sequence. Typically, one probe includes 25 nucleotides. The expression products of several genes can bind to the same probe sequence. The intensity level for the individual genes is obtained by computationally combining the intensity levels of a set of probes (Affymetrix, 2006).

3.3 Microarray Experiment

Once the cDNA sample has been labeled with a fluorescent dye and the microarray chips have been acquired, the sample can be hybridized to the microarray slide. In the hybridization, the microarray slide is covered with the sample and the cDNA sequences bind to the corresponding probes. As the hybridization is a highly sensitive process, specific hybridization chambers are usually used. These chambers are used to keep the temperature constant and to prevent the hybridization solution from evaporating. Microarray slides are kept in the hybridization chamber typically overnight. Once the hybridization process is completed, the microarray slide is washed to remove any unbound material from the slide. Once the slide has dried, it is ready to be analyzed.

Fluorescent dyes that have bound to the cDNA can be read from the slide using a laser excited microscope, known as microarray scanner. As the dyes have bound to each fragment of the cDNA in the sample, the amount of emitted light is proportional to the amount of cDNA present in the sample. In the case of the two channel microarray, each slide is scanned twice using lasers with different wavelengths. Thus, the intensity information for the Cy3 and Cy5 dyes can be read independently, resulting in intensity measurements of both samples.

3.4 Data Preprocessing

Even though there are significant differences between different microarray technologies, for example how the slides are fabricated and how the experiment is conducted, all technologies produce a similar outcome. As a result a digital image is obtained. To be able to do computational analysis with the microarray data, intensity information needs to be extracted from the image (Speed, 2003; Zhang *et al.*, 2004). Subsequently, information from different probes can be combined and different kinds of quality control methods can be applied to make sure the obtained data is of good quality.

While different array technologies require very different algorithms for the data preprocessing, the workflow is still the same (Leung and Cavalieri, 2003). In the following we discuss the required analysis steps in the context of two channel microarrays.

3.4.1 Image Processing

To be able to extract intensity information from the scanned microarray slides, spot areas need to be identified from the slide image. This can be

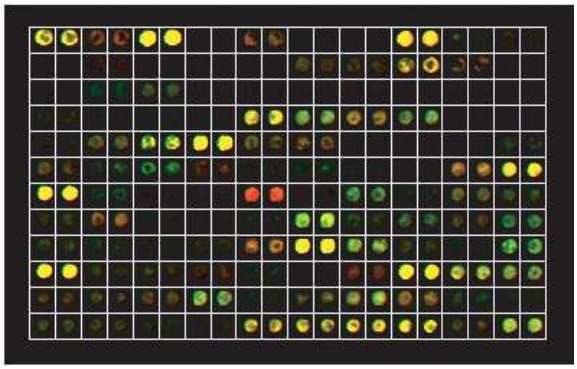

Figure 3.4: An example of a grid alignment over a microarray slide. The grid alignment for one subarray from a cDNA microarray image is shown (Tuimala and Laine, 2005).

done with a two step process. First, a rectangular grid is aligned over the slide such that each spot is assigned into a block of its own (Figure 3.4). Next, a segmentation algorithm is applied to each block. As a result a boundary between the spot area and the background is obtained (Figure 3.5). There are several different algorithms available that can be used for these tasks (Yang *et al.*, 2002*a*; Speed, 2003).

Once the spots are segmented using an automatic or interactive segmentation algorithm, different statistics can be computed from the spot. Usually, at least the spot area and spot surrounding, that is the background, intensity values are computed (Speed, 2003). In addition, several parameters that characterize the quality of the spot can be computed. For example, the number of pixels in the spot area or spot roundness can be quantified. If the spot area is small compared to other spots, it may indicate that information from the spot is not reliable. This kind of quality parameters are important when the reliability of the individual data points are evaluated.

Most microarray scanners come bundled with an analysis software. Thus, the extraction of information from the microarray slide can be done in a very straightforward manner. However, the algorithms that are used to extract information have an impact on the obtained data (Lehmussola *et al.*, 2006). This issue is briefly discussed in Publication II.

3.4.2 Quality Control

After the image processing, spot intensity values are available. The first step in the analysis is to compensate for technological limitations and to transform the data in a form suitable for the analysis (Speed, 2003; Quack-

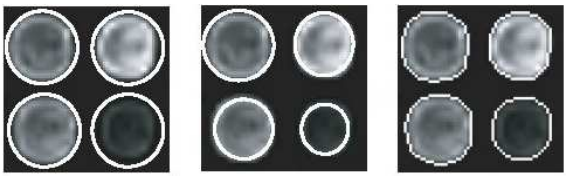

Figure 3.5: Examples of spot segmentation using different algorithms. There is a clear difference in the segmentation accuracy (Tuimala and Laine, 2005).

enbush, 2002). During the hybridization the genetic material and thus the fluorescent dyes can attach to areas other than the spots. When the microarray slide is scanned background area will yield non-zero intensity reading. Thus, it is assumed that there is an additive hybridization bias that can be observed from the background area. Therefore, before any further analysis the background intensity is subtracted from the spot intensity

$$\hat{I}_i = I_i^s - I_i^b. \tag{3.1}$$

Here I_i^s is the spot and I_i^b is the background intensity. Index $i \in \{3, 5\}$ corresponds to the dye color red or green. The background intensity can be estimated globally for the entire slide or independently for each spot. While the background subtraction is a common practice in the microarray analysis, there are counter arguments why it should not be done (Gottardo *et al.*, 2003). For example, the binding properties of the spot and the background area are different. Thus, the assumption about an additive noise model may not hold.

Binding efficiencies of all the probes are not similar. Some probes bind extremely well and will be observed at high intensity levels while others have significantly lower binding efficiency (Hein *et al.*, 2005; Dror *et al.*, 2003; Weng *et al.*, 2006). In addition, the variation within large intensity values is much larger than within small intensity values. To compensate this a log transform can be applied

$$I = \log_2(\hat{I}), \tag{3.2}$$

where $\hat{I}$ is the background subtracted intensity value. This transformation has the property that it makes the variation of intensities more independent of the absolute magnitude of the intensity values (Durbin *et al.*, 2002).

When a reference sample is available, instead of using log intensity values directly, log ratios can be used. With a two channel microarray, reference sample is usually readily available in the other dye channel. With a single channel microarray, reference sample can be from other microarray chip. A

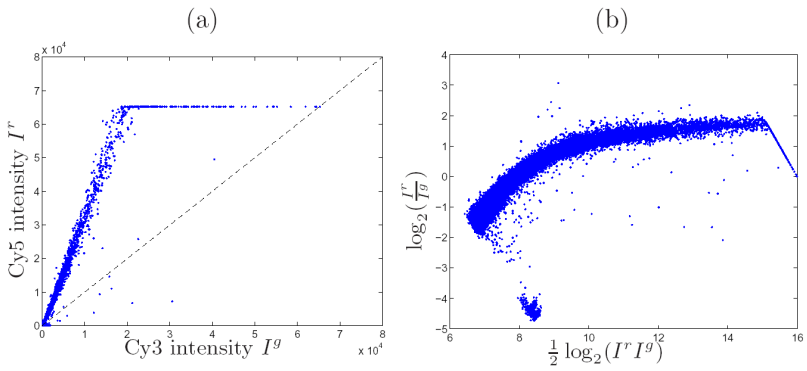

Figure 3.6: An intensity scatter plot and a MA-plot are shown. Intensity plot (a) shows that there is a clear linear relationship between different dyes. However, there is an observable dye bias on both scatter plots. Scatter plots also show the saturation of intensity values. In addition, a group of erroneous measurements can be identified from the MA-plot. A cluster of points at the lower left corner includes bad quality data.

log ratio is obtained from the intensity values as

$$R = \log_2(\frac{I^t}{I^r}) = \log_2(I^t) - \log_2(I^r), \tag{3.3}$$

where I^t is the test and I^r reference sample intensity and R is the obtained log ratio. Conveniently, log transformed cDNA intensity ratio data is usually approximately Gaussian distributed (Zhang *et al.*, 2004).

Once the data is presented in the desired form, the next step in the analysis is to remove all bad quality data (Speed, 2003). This data can be identified using spot quality statistics that were extracted by the image processing algorithms. Typically, spots that have a very low intensity value or spots that could not be segmented at all are candidates for the removal. In addition, saturated spots can be removed. Saturation may indicate that several genes have bind to the same spot or that there is some other kind of hybridization error.

The quality of the data can be assessed using different types of scatter plots (Figure 3.6). The most commonly used scatter plot is the MA-plot (Chambers *et al.*, 1983; Bolstad *et al.*, 2003), where the product of intensities from two channels is plotter against the ratio of intensities. These kinds of plots can be used to identify the bad quality data points, as illustrated in Figure 3.6.

If replicated microarray experiments are available, it is not necessary to remove bad quality data directly. Instead, replicated values can be used to

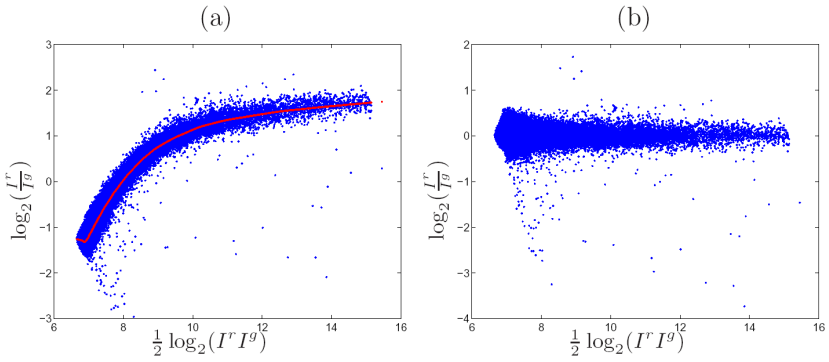

Figure 3.7: Original MA-plot after quality filtering and a lowess fit to the data is shown on the left (a). On the right (b) MA-plot of the lowess normalized data is shown. Clearly, the systematic bias has been removed.

compute a combined expression statistic for the gene expression (Quackenbush, 2002; Tseng *et al.*, 2001). Most straightforward way to combine the replicates is to compute a mean or median over all the replicates. Also more advanced replicate combining methods have been proposed (Ideker *et al.*, 2000). There are several different ways to perform the replication of microarray data (Churchill, 2002). One slide can include more than one probe for an individual gene or several similar arrays can be hybridized using the same sample yielding technical replicates. Alternatively, the whole biological experiment can be repeated resulting in biological replicates.

3.4.3 Normalization

After the data extraction and quality filtering, intensity values are not yet directly comparable. Due to various sources of systemic bias, for example, different dyes have different incorporation efficiencies, the data needs to be compensated for biases before further analysis are made. This process is called normalization (Quackenbush, 2002). There are two different normalization steps, within slide and between slide normalization (Yang *et al.*, 2002*b*). The goal of the normalization process is to remove variation from the data that is not from a biological origin.

Within slide normalization is particularly important with the two channel microarray data, because of the above mentioned fluorescent dye bias. This kind of bias can be addressed using a robust local regression. The most commonly used method is to fit a curve to the log ratio data using locally weighted scatter plot smooth (lowess) algorithm (Figure 3.7) (Cleveland, 1979).

This normalization approach is based on the assumption that the genes that are not differentially expressed should have the same expression value with both dye colors (Speed, 2003). Examples of this kind of genes are the control spots and house keeping genes whose expression should be the same under all conditions. However, it has proven to be more robust to do the normalization using all the genes. If the number of under-expressed and over-expressed genes is approximately the same, then on the average there should not be any significant bias due to the differential expression of genes. This assumption has proven to hold under most conditions and thus, lowess normalization can be done using all the genes. Another type of error that can be addressed by within slide normalization is uneven hybridization. Different areas of the slide might have different hybridization efficiencies. This problem can be solved by doing the above described lowess normalization separately for each subarray of spots (Speed, 2003).

After the within slide normalization the intensity values within the slide are comparable. If a microarray experiment includes several slides, further between slide normalization might be needed to compensate the systematic differences between the slides. If the lowess normalization has already been applied, the data distribution should have a zero mean. Thus, if the variation of data is the same at different slides, further between slide normalization is not needed. This is the case with some high quality microarray technologies (Yang *et al.*, 2002*b*).

If the data from different arrays have different statistical characteristics, such as the mean or variance, between slide normalization is needed. Several different approaches have been proposed including a median and quantile normalization (Shmulevich and Zhang, 2002; Huang and Pan, 2002; Bolstad *et al.*, 2003). With this kind of approach each slide is normalized to have the same median or quantile values.

Normalization schemes that are based on global statistics of the dataset are sufficient for most large scale analyzes. However, a much more detailed model based normalization algorithms have been proposed. These models try to identity and compensate errors from different sources (Ideker *et al.*, 2000; Hartemink *et al.*, 2001). The model based approach makes it possible to asses the reliability of individual intensity values through the p-values yielding a more reliable analysis of the data. The model based approach and different sources of error will be discussed in more detail in Chapter 4.

3.5 Data Analysis

After the normalization, microarray data is ready to be used in the subsequent data analysis (Quackenbush, 2001). The goal of the data analysis

is to extract biologically interesting information from a large dataset using computational methods. Data analysis can be done in unsupervised fashion without using any additional knowledge or alternatively in supervised fashion, where additional information, for example, clinically determined class labels for the samples is used. A special characteristic of the microarray data analysis is that the number of observations (samples) m is typically much smaller than the number of variables (genes) n. Thus, a dataset is $\mathbb{R}^{n \times m}$ dimensional, where $n \gg m$. Here, we will briefly discuss some of the most common analysis tasks and introduce standard computational methods for the analysis. These methods have been applied to real biologically motivated data analysis tasks in Publication I and Publication IV.

3.5.1 Unsupervised Analysis

In the unsupervised analysis microarray data is analyzed without utilizing any additional information. A goal of this kind of analysis is to extract information about the underlying structure of the dataset. Thus, common unsupervised tasks include the clustering of data, for example, to find functional modules, or the projection of the data in a lower dimensional space for illustration.

Unsupervised analysis can be applied to a dataset in gene-wise or sample-wise. That is, we may want to study the similarities between the individual genes or samples. When the sample-wise analysis is performed it is of interest to exclude uninformative genes from the analysis. Thus, the genes that do not change their expression between the samples can be removed as they do not contribute any information to the analysis. To remove uninformative genes, a fold change can by used to measure information (Cui and Churchill, 2003). For example, a 5 percentile and 95 percentile can be compared. The pth percentile is obtained from m ordered values by computing the rank $k = p(m + 1)/100$, rounding k to the nearest integer, and selecting the kth value. For example, if the fold change between the 5 percentile and 95 percentile is less that two-fold the gene is deemed to be uninformative and can be removed. In addition, the genes can be removed if the maximum intensity over all the samples does not exceed a given threshold. This kind of filtering of uninformative genes can significantly improve the results obtained with the unsupervised analysis as we remove the characteristics that can not be reliably detected from the data.

If the analysis is done gene-wise, additional standardization of the gene expression profiles over all the samples can be applied. This is needed to bring the variation between the genes into the same scale, and thus to make the expression profiles comparable. The most common gene-wise standardization method is to make the expression profile to have zero mean and unit

variance

$$\hat{\boldsymbol{x_i}} = \frac{\boldsymbol{x_i} - \mu_{x_i}}{\sigma_{x_i}}, \tag{3.4}$$

where $\boldsymbol{x_i}$ is the expression profile of gene i and μ_{x_i} and σ_{x_i} are the mean and standard deviation of $\boldsymbol{x_i}$, respectively. An alternative standardization method is to apply a fixed norm standardization

$$\hat{\boldsymbol{x_i}} = \frac{\boldsymbol{x_i} - \mu_{x_i}}{|\boldsymbol{x}_i|}, \tag{3.5}$$

where $|\boldsymbol{x}_i|$ is the norm of $\boldsymbol{x}_i$. This method is more appropriate especially if the variance of the data is expected to change in time during the experiment due to biological reasons.

Clustering algorithms can be used to find co-regulated genes or to measure the similarity between samples. Hierarchical clustering is an often used tool for gene expression analysis (Eisen *et al.*, 1998; Quackenbush, 2001). First, the distances between the genes or the samples are computed using a distance metric. Possible choices of the metric include the Euclidean distance

$$d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}, \tag{3.6}$$

or correlation

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x\sigma_y}, \tag{3.7}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two expression profiles of length n, μ_x and μ_y are the means and σ_x and σ_y are the standard deviations of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

Based on the selected distance metric, the expression profiles are linked using a linkage method. Again, there are several options for the linkage method including, for example, the single linkage

$$l(r, s) = \min(d(x_{ri}, x_{sj})), i \in 1, \ldots, n_r, j \in 1, \ldots, n_s, \tag{3.8}$$

complete linkage

$$l(r, s) = \max(d(x_{ri}, x_{sj})), i \in 1, \ldots, n_r, j \in 1, \ldots, n_s, \tag{3.9}$$

and average linkage

$$l(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} d(x_{ri}, x_{sj}), \tag{3.10}$$

where n_r is the number of objects in the cluster r and n_s is the number of objects in the cluster s. Based on the linkage, a tree that represents the distances between different samples is obtained.

In addition to hierarchical clustering, several other more advanced clustering methods are available. These include, for example, the k-means clustering (Hastie *et al.*, 2001), fuzzy c-means clustering (Dembélé and Kastner, 1999), and self-organizing maps (Tamayo *et al.*, 1999; Kohonen, 2001). These methods find the cluster structure using more advanced measures for cluster similarity than the hierarchical clustering. Typically, this means that a fixed number of clusters is specified and then the algorithm finds a cluster structure that minimizes the given cost function. For example, the k-means algorithm clusters n data points $x_1, x_2, \ldots, x_n$ into k disjoint clusters $U_1, U_2, \ldots, U_k$. The clustering is carried out in such a way that the sum

$$\sum_{m=1}^{k} \sum_{x_n \in U_m} (x_n - \mu_m)^2, \tag{3.11}$$

where μ_m is the centroid of the cluster U_m, is minimized. The minimum of the expression in Equation 3.11 can be found by iterating two steps: (i) Assign each data point to the cluster that has the closest centroid, (ii) Recalculate the positions of the centroids. The algorithm minimizes the sum of point-to-centroid distances for all clusters. The method is guaranteed to converge to a local optimum (Hastie *et al.*, 2001).

Recent research has also focused on so called biclustering methods, where both the samples and the genes are clustered simultaneously (Tanay *et al.*, 2004; Prelić *et al.*, 2006). Obtained biclusters are allowed to overlap with each other. Thus, the difference to the traditional clustering methods is that the same genes and samples are allowed to appear in several clusters.

Dimensionality reduction techniques can be used to visualize or to find similarities in the data. Multidimensional scaling (MDS) is a commonly used method for this purpose (Borg and Groenen, 2005). The basic form of the MDS is known as a classical or metric multidimensional scaling. MDS takes a general distance matrix $D^{n \times n}$ as an input and the goal of the algorithm is to find a configuration of points $\boldsymbol{x}_k = (x_k^1, \ldots, x_k^p)^T$, $k \in 1, \ldots, n$ that produce the given distance structure in the p-dimensional Euclidean space. That is, $d^*_{i,j}$ the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the configuration space approximates $d_{i,j}$ the distance between i and j in D for all pairs of i, j. The only requirements for the input matrix D are the symmetry $d_{ij} = d_{ji}$, non-degeneracy $d_{ii} = 0$, and triangular inequality $d_{ij} + d_{jk} \geq d_{ik}, \forall i, j, k$. Thus, D can include distances computed by any metric that fulfills these criteria (Borg and Groenen, 2005).

When dissimilarities between $d_{i,j}$ and $d^*_{i,j}$ are treated as Euclidean distances, the solution is obtained by minimizing the cost function

$$E = \sum_{i,j} (d_{i,j} - d^*_{i,j})^2. \tag{3.12}$$

The result obtained with this approach is analogous to the principal component analysis (PCA) (Johnson and Wichern, 1998).

In modern multidimensional scaling more complex dissimilarity measures can be used to measure the difference between $d_{i,j}$ and $d^*_{i,j}$ (Borg and Groenen, 2005). Instead of trying to approximate the dissimilarities themselves, as done in classical MDS, non-metric scaling can be used to approximate a nonlinear, but monotonic, transformation of dissimilarities (Kruskal, 1964*a*). This will make the MDS more general by allowing only the rank of the distances to be preserved. Non-metric scaling can be done by minimizing the stress function

$$S = \frac{\sum_{i,j}(\delta_{i,j} - d^*_{i,j})^2}{\sum_{i,j}(d^*_{i,j})^2}, \tag{3.13}$$

where $\delta_{i,j} = f(d_{i,j})$ is a monotonic transformation of $d_{i,j}$. Several different methods have been proposed to minimize the stress function. Typically, this includes two iterative steps, searching the optimal coordinates and the optimal monotonic transformation (Borg and Groenen, 2005; Kruskal, 1964*b*).

3.5.2 Supervised Analysis

In the supervised analysis, in addition to the expression data, some other biological knowledge is utilized. This kind of biological knowledge may include the clinical class labels for the samples or the functional categories of genes (Eisen *et al.*, 1998; Golub *et al.*, 1999; Alon *et al.*, 1999; Ashburner *et al.*, 2000). Several unsupervised clustering methods can easily be extended to take advantage of, for example, the functional categories of genes.

A typical supervised analysis task is to find the differentially expressed genes between two types of samples, for example, between two different types of cancers. This can be done by using a statistical test to test the similarity of samples in two groups (Cui and Churchill, 2003). Commonly used tests include two-sample t-test and Mann-Whitney test.

Parametric two sample t-test assumes that the testable data is normally distributed. Then the null hypothesis "means μ_1 and μ_2 of two populations are equal" is tested against the alternative "means are not equal" hypothesis with a predetermined significance level α. The test statistic is given as

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{3.14}$$

where μ_1 and μ_2 are the means, σ_1 and σ_2 are the standard deviations of the two corresponding populations, and n_1 and n_2 denote the number of samples in the populations. By using the value of the test statistic a p-value for the significance can be obtained. If the p-value is smaller than the

significance level α, the null hypothesis should be rejected. Non-parametric version of the t-test can also be defined. In this case, the p-value is obtained by randomly permuting the labels of the samples and computing the p-value from the obtained empirical distribution.

Unlike t-test, the Mann-Whitney test is non-parametric and does not make the normal distribution assumption. It does, however, assume that both samples are from the same distribution. The test statistic is given by

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \tag{3.15}$$

where R_1 is the sum of ranks in the first population. Terms n_1 and n_2 denote the number of samples in the populations.

In addition to these basic statistical tests, numerous methods specially designed to detect differentially expressed genes from microarray data have been proposed (Golub *et al.*, 1999; Tusher *et al.*, 2001; Subramanian *et al.*, 2005). For example, the weighted voting algorithm is based on correlating the expression profiles with ideal class labels (Golub *et al.*, 1999). Higher the degree of correlation, more significant the gene is deemed to be. With this kind of methods a p-value can be obtained from an empirical distribution estimated by permuting the class labels.

A problem in finding the differentially expressed genes by testing individual genes is that the number of tests is equal to the number of genes. Thus, if we are testing 10 000 genes with significance level $\alpha = 0.05$ we will find 500 false positives, that is differentially expressed genes, by chance. This is known as the multiple testing problem. The experiment wide significance level α_e is dependent on the number of tests n

$$\alpha_e = 1 - (1-\alpha)^n. \tag{3.16}$$

To compensate the increase in α_e, several computational approaches have been proposed. The most straightforward approach is to use the Bonferroni correction (Johnson and Wichern, 1998). This method controls the family-wise error rate (FWER) which is the number of false positives. This is done by adjusting the significance level α so that the experiment wide significance level will be, for example, $\alpha_e = 0.05$. This can be done by dividing the desired significance level by the number of tests $\alpha = \alpha_e/n$.

An alternative approach is to use the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). In this approach, a predetermined rate of false positives is allowed. The false discovery rate is defined as

$$Q = \frac{V}{V+S}, \tag{3.17}$$

where V is the number of false positives and S is the number of true positives. One wants to keep the value of Q under the threshold q. As V and S are

random variables the value of Q can not be computed directly. However, there are algorithms to ensure that the expected value of Q is less than q. Let $p_1, \ldots, p_n$ be the p-values from n independent tests. We can order these in the increasing order of magnitude, denoted by $p_{(1)}, \ldots, p_{(n)}$. Then, given q we can find the largest k, the number of significant tests, by

$$\forall i \leq k : p_{(i)} \leq \frac{i}{n} q. \tag{3.18}$$

Instead of just finding the differentially expressed genes, the goal of the supervised analysis can be the classification or class discovery of the samples. That is, given a set of samples with the known class labels assign a new sample to the class whose members are the most similar.

Because the number of genes n is a significantly larger than the number of samples m, we need to take care not to overfit the classifier. Thus, a key issue in the microarray data classification is the selection of the features that are used in the classification. A large number of papers have been published proposing how the features should be selected. These include statistical methods based on t-test (Jaeger *et al.*, 2003), different kinds of clustering methods (Getz *et al.*, 2000), methods based on information theoretic criteria (Tabus *et al.*, 2003; Ding and Peng, 2003), and many more. A problem is that methods for feature subset selection make assumptions about the properties of data. Thus, different algorithms will select different sets of features and the classification accuracy will depend on the feature selection algorithm. Ideally, the classification should be done without any feature selection, utilizing all available features. However, this is usually not possible without significantly overfitting the classifier. In addition, most of the genes do not contain any information about the class separation and thus, to reduce noise, they should not be included as features.

Feature subset selection can be done before the classification is considered. This kind of a filter approach typically selects a list of genes that give the best separation between the classes of samples in terms of the feature selection criteria (Jaeger *et al.*, 2003). Alternatively, feature selection can be done using a wrapper approach where the features are selected along the classifier design (Krishnapuram *et al.*, 2004).

In addition to feature subset selection, selection of the classification algorithm needs to be considered. Numerous different classification algorithms have been used with microarray data, including the nearest neighborhood, support vector machine, Bayesian, and linear discriminant classifiers (Hastie *et al.*, 2001). The nearest neighborhood classifier is often used as it is intuitive and simple. It is based on a distance metric, like the correlation or Euclidean distance, between samples. Sample can be classifier using the k nearest neighborhood classifier as follows. Find k nearest samples in terms

of the distance metric. Perform a majority vote and assign the sample the the class that has the majority of k nearest neighbors (Hastie *et al.*, 2001).

Chapter 4

Microarray Data Quality

As discussed in the previous chapter, a microarray experiment includes several steps that may contribute stochastic variation or systemic bias to the data. To obtain a good quality data, these sources of error should be controlled during the experiment and systemic biases should be removed from the data by the preprocessing and normalization.

In addition to the microarray experiment process, error sources may lead back to the sample preparation. Identification of a right type of high quality samples is not always straightforward. In the case of cancer research the samples of interest are usually different types of tumors (Golub *et al.*, 1999; Alon *et al.*, 1999; Alizadeh *et al.*, 2000). It is a challenging task to dissect a tumor in such a way that only cancer cells are present in the sample. This may lead to sample heterogeneity (Lähdesmäki *et al.*, 2005). Furthermore, identification of the cancer tumor type is not always unambiguous and the classification of cancer types is constantly changing to incorporate new knowledge (Harris *et al.*, 1994). This may lead to revision and introduction of new cancer types.

If the sample is, for example, from a cell culture, sample heterogeneity or unambiguous of the sample origin are not issues. However, even with a cell culture there are issues that need to be taken into account in the experiment design. As numerous cells live in the culture they are at the different phase of their live span. Thus, they may have a different response to a given stimulus. This will lead to a problem since with microarray technology we are only able to observe the average behavior of the entire population[1]. To observe the behavior in more detail, the cell population should be synchronized or individual cells should be observed (Lähdesmäki *et al.*, 2003).

In this chapter, signal processing methods that can be used to improve the quality and reliability of microarray data are discussed. The presented

[1] Unless a single cell experiment is conducted.

methods can be used to address the problems that have been raised above.

4.1 Class Label Verification

Classification algorithms can be used to find the class labels for new samples. Once we have samples with a known ground truth, a classifier can be trained. By using the obtained classifier, a class label can be assigned for new samples according to the classification outcome. This class discovery problem has been studied in several publications (Golub *et al.*, 1999; Alon *et al.*, 1999). However, to be able to address this problem one need to have the training samples with the known ground truth.

It is not always obvious how to obtain a biological ground truth of samples unambiguously. In the field of cancer research it is common to look for available tumor samples from an institutional database. Based on the information at the database the samples that are relevant for the experiment at hand can be obtained. This kind of approach possesses several dangers. Normally pathologists use several different types of information including topological properties and histopathological diagnosis of the tumor to determine the type of the tumor. With these features the tumor is assigned to a class which is then used as a basis for the treatment. It is not uncommon that different pathologists make a different diagnosis as all the characteristics of the cancer tumors are not always clear (Trotter and Bruecks, 2003; Xavier *et al.*, 2005).

Another source of error may be the information stored in the institutional database. Even though the initial diagnosis of a pathologist would be correct, the classification criteria may have chanced over time (Harris *et al.*, 1994). This has happened, for example, with the leiomyosarcomas (LMS). Recently, a new class of sarcoma tumors known as the gastrointestinal stromal tumors (GIST) has emerged (Antonescu *et al.*, 2004; de Schipper *et al.*, 2004). Thus, if one looks for LMS tumors from the institutional database, one will obtain a number of tumors that in the light of current knowledge are GIST (Publication I).

Computational methods can be used to help in the verification of class labels. In the verification we can not rely on the available class labels, and thus supervised analysis techniques are not directly applicable. We could study the prediction error of a classifier and identify the samples that are not classified correctly. By repeating this analysis using several different classifiers, we could conclude that those samples that are constantly classified wrong have potentially wrong class label. However, it is more straightforward to evaluate the reliability of class labels using an unsupervised approach.

Multidimensional scaling (MDS) can be used to illustrate a microarray

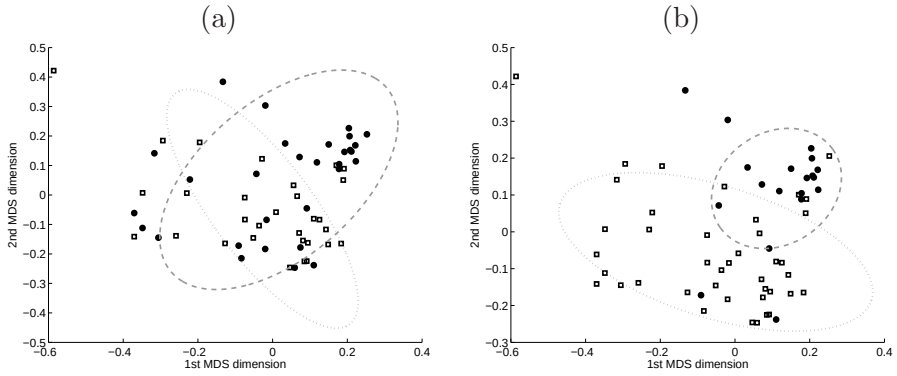

Figure 4.1: Two dimensional MDS representation with (a) labels from the institutional clinical database and (b) after pathological re-review and validation. Samples from patients with GIST are denoted by squares and those from patients with LMS by circles. The spread of the populations is demonstrated by ellipses whose size corresponds to the variance of the class spread. Dashed and dotted ellipses correspond to LMS and GIST, respectively. In (a) the ellipses are overlapping, thus there is no visible separation between LMS and GIST. In (b) with the corrected class labels GIST and LMS appear as a distinct clusters.

dataset in a lower dimensional space. In Figure 4.1(a) two dimensional MDS presentation of the microarray data from 60 samples is shown. From 60 samples 30 are from GIST and 30 from LMS according to the institutional database. As these are distinct types of tumors, we should expect to see a clear separation between them in the MDS space (Antonescu *et al.*, 2004). However, all the samples appeared mixed together in a one big cluster of points. This observation raises a doubt about the reliability of class labels. In subsequent pathological re-analysis of the tumors it prove out that 11 of the samples that were originally labeled as LMS were in fact GIST. With the corrected class labels there were more evident separation of classes in the MDS presentation of the data (Figure 4.1(b)).

If the unsupervised analysis of microarray data is successful, we should see as many clusters as there are different classes of data. If the assigned class labels are consistent with the obtained clusters, that is, each cluster includes samples of different type, we can trust that the class labels are correct. Then, the supervised analysis, for example, the identification of differentially expressed genes can be done with confidence. If the result is not what is expected, then in addition to erroneous class labels we should considered whether there are some other type of biological variation between the samples that could explain the observed behavior. For example, sample heterogeneity or different experiment conditions may cause variation that

can be more prominent than the differences between cancer types.

It should be noted that different types of samples are not always separable by unsupervised analysis. Thus, unsupervised analysis can only be used to estimate the quality of biological ground truth if distinct clusters can be observed in the first place.

4.2 Errors and Noise in Microarray Data

As discussed earlier, numerous steps in a microarray experiment contribute to the quality of data (Zhang *et al.*, 2004). Thus, there are several sources that may cause errors in the form of stochastic fluctuation or systemic bias. Some of the sources are due to human interaction in the process and some are due to the properties of the materials involved in the experiment. Errors that originate from the human interaction can be controlled with a good experimental design and detailed laboratory work (Zhang *et al.*, 2004). Still, there remains a large number of error sources that can not be controlled directly. As microarray technology has matured, better protocols and array manufacturing techniques have been introduced (Hughes *et al.*, 2001). This has improved the quality of data significantly. It should be noted that this is the reason why the data from different technologies have different statistical characteristics. The technology in use determines what kind of sources of error are present in the obtained data.

One source of error, present in all measurements, is stochastic fluctuations within the cell. Several studies have tried to characterize the structure of this type of intrinsic noise (Blake *et al.*, 2003; Fraser *et al.*, 2004). However, as a microarray experiment typically is a measurement from a cell population, this kind of variation is averaged out and thus, is not a significant source of error in the microarray studies.

More important error sources are those related to the sample preparation and experimental setup, for example, the sample heterogeneity. Furthermore, the amplification of cDNA with a PCR may cause errors in replicated cDNA fragments. A microarray experiment itself also contains several potential error sources. There may be quality problems within the slide. Some probes may have a sequence other than expected due to printing errors or there might be a difference in binding efficiency of the surface within or between different slides. Further problems may arise from uneven hybridization or from a change in conditions during the hybridization (Balagurunathan *et al.*, 2002).

There have been several studies focusing on characterizing the properties of microarray noise (Nykter *et al.*, 2003; Hartemink *et al.*, 2001; Ideker *et al.*, 2000; Hein *et al.*, 2005; Tu *et al.*, 2002; Dror *et al.*, 2003; Rocke and

Durbin, 2001; Cho and Lee, 2004; Weng *et al.*, 2006). These include both global approaches where all the sources of error are modeled together and more detailed approaches where error sources have been identified and modeled separately. An example of a detailed model is the hierarchical error model that has been proposed to model the noise in cDNA microarray data (Cho and Lee, 2004). This model includes different terms for different error sources. The error model is defined at the log scale in two stages as

$$y = X + \epsilon \tag{4.1}$$

$$X = \mu + g_i + c_j + r_{ij} + b_{ijk}, \tag{4.2}$$

where μ is the ground truth intensity, g_i is the noise specific to the gene i, c_j is the noise specific to the chip j, r_{ij} is the noise specific to the chip j and gene i, b_{ijk} is the noise specific to the gene i, chip j, and biological sample k, and ϵ is the random noise independent of the gene i, chip j, and biological sample k. It should be noted that this kind of separation of variances is a common technique in analysis of variance (ANOVA). Thus, the model parameters could be estimated using the proposed Bayesian or more traditional ANOVA approach (Cho and Lee, 2004).

This model is one example of the characterization of microarray noise. To estimate the model parameters, assumptions about different noise terms need to be made. In the case of hierarchical error model all sources of error are assumed to be from a zero mean Gaussian distribution with the standard deviation dependent of the noise type (Cho and Lee, 2004). As the model is formulated at the log scale, this means that the noise is multiplicative and, thus, nonlinear in nature.

When compared to traditional methods, the model based analysis has proven to improve the results when applied to normalization and analysis of microarray data (Hartemink *et al.*, 2001; Dror *et al.*, 2003; Ideker *et al.*, 2000). However, a problem is that each microarray technology requires a model of its own. As the noise characteristics and the underlying technology are very different, the formulation and assumptions about the noise need to be different as well (Dror *et al.*, 2003; Hein *et al.*, 2005; Weng *et al.*, 2006). In addition, it is difficult to estimate whether the proposed noise model really works well as there is no ground truth about the data available. Thus, the validation has been done based on the statistical properties of the data or by comparing the scatter plots. Here less scatter would indicate a better performance. To be able to utilize the model based analysis effectively, an additional control data that can be used to tune the model parameters needs to be generated (Dror *et al.*, 2003). This kind of data can be obtained by performing the experiment with replicates. Then, the statistical properties of the data can be reliably estimated (Ideker *et al.*, 2000).

Even though there are practical problems that limit the applicability of the model based approach to data analysis, the characterization of the

statistical properties of microarray data has other uses as well. Error models can be used to simulate microarray data with realistic characteristic.

The simulation of microarray data is of interest as it provides a way to obtain a realistic ground truth data. A major problem in developing algorithms for biological applications is that the ground truth of the data is not known. This makes it difficult to estimate the performance of the algorithms. Having a realistic simulated data available, the performance of the algorithms can be evaluated and compared objectively (Quackenbush, 2001). It should be pointed out that a problem with this approach is that a model based evaluation of algorithms always favors the one that makes the same assumptions about the data as the model does. Thus, to get an objective estimate of the algorithms performance, the algorithms should be tested using data generated with several different models.

In addition to data analysis algorithm evaluation, there are also other applications where simulated microarray data can be used. By changing the noise parameters, the effect of different error sources can be studied. This information can be used to refine the microarray experiment protocols. If we have an accurate enough computational model about the system we are studying, it would be possible to model the entire microarray experiment. This would help to find potential problems in the experimental design, and make it possible to redesign the experiment such that a hypothesis can reliable be tested when real data is generated.

Simulation of realistic microarray data is a challenging task as there are several steps that effect the outcome as discussed earlier. First, simulated ground truth data from biological system under study needs to be obtained. Depending on the usage of simulated data, there are several options how this can be done. In the ideal case a kinetic model for the system could be used. However, in practice this kind of models are rarely available. Instead, a random network model with realistic reaction kinetics can be used to simulate the behavior of a mock system (Mendes, 1993; Pettinen *et al.*, 2005). This kind of data, simulated using a network with random connections, should still have the essential characteristics of real system (Mendes *et al.*, 2003). If we are only interested in simulating realistic data, not to model the system, the ground truth data can be generated directly from a distribution that corresponds to the properties of real data.

As a next step, there is a need to model biological and stochastic variation due to different error sources. We have implemented several noise models that have been proposed earlier (Nykter *et al.*, 2003; Dror *et al.*, 2003; Rocke and Durbin, 2001; Cho and Lee, 2004; Hein *et al.*, 2005). Along with the models, methods for estimating the model parameters from a real measurement data have been proposed (Dror *et al.*, 2003; Cho and Lee, 2004). These methods can be used to estimate realistic parameters for the

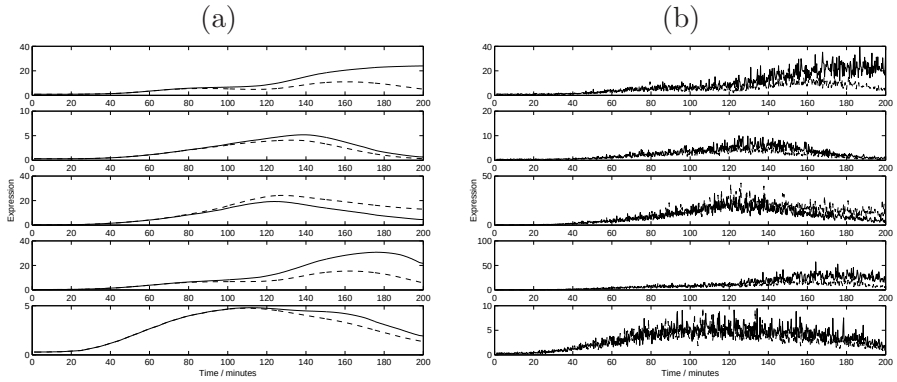

Figure 4.2: Gene expression profiles of the selected genes, simulated as explained in Publication II. Selected noise free expression profiles (a) and the same expression profiles after the hierarchical error model has been applied (b) are shown.

simulation. Selected gene expression profiles, simulated using a random network model as explained in Publication II, are shown in Figure 4.2(a). The same expression profiles after the hierarchical error model has been applied are shown for comparison in Figure 4.2(b).

Finally, a microarray experiment including the slide manufacturing process needs to be simulated. We have developed a simulation model that includes all the error sources that are commonly observed in microarray images. These error sources can be used to test the robustness of image processing algorithms and also to test, for example, how different normalization methods perform when spatial errors are introduced into microarray slides. Examples of simulated slides are shown in Publication II.

We have proposed a modular framework that can be used to model realistic microarray data. Our approach uses noise models that have been developed earlier. A ground truth data generation can be done using available network simulation programs (Mendes, 1993; Pettinen *et al.*, 2005). For the microarray slide manufacturing and hybridization we have developed a model that takes into account several possible sources of error. Our simulation approach is discussed in detail in Publication II.

4.3 Supplemental Measurement Data

Along a microarray experiment, other measurement technologies can be used to obtain a supplemental data. For example, we can measure the concentrations of cell populations, the amount of cell mass, or the phase of cell

cycle (Spellman *et al.*, 1998; Pitkänen *et al.*, 2004; Bar-Joseph *et al.*, 2004). We can use this kind of data to verify that the experiment is performed as planned. In addition, we can also use it to improve the quality of the obtained microarray data.

An example of an application where a microarray experiment can benefit from a supplemental measurement data is the cell cycle studies. To be able to study the cell cycle behavior with microarrays, we need to obtain a synchronized cell population[2]. There are several methods that can be used to obtain an approximately synchronous population of cells (Spellman *et al.*, 1998; Shedden and Cooper, 2002*a*,*b*). Once a synchronized population is let to grow freely, it will start to lose the synchrony. This significantly limits the time frame when we can study the behavior of a synchronized population. As a solution, computational methods that can be used to invert the effect of the loss of synchrony have been proposed (Lähdesmäki *et al.*, 2003; Bar-Joseph *et al.*, 2004). These methods are based on modeling the loss of synchrony by convolution. Thus, by deconvolving time series expression data, we can obtain data that corresponds to a measurement from an ideal synchronized population. For this purpose, we need to obtain an estimate of the cell population distribution.

A fluorescence activated cell sorter (FACS) is a device that can be used to measure the amount of DNA within a cell (Lodish *et al.*, 2001). This process is based on fluorescent dyes that bind to the genetic material within a cell. As the cells are run through a measurement point and the fluorescent is illuminated with a laser, the amount of emitted light can be measured. The amount of light is relative to the amount of genetic material. This process can be done for tens of thousands of cells in a high-throughput manner (Figure 4.3). As the amount of DNA within the cell is dependent on the phase of the cell cycle, this kind of data can be used to estimate the distribution of a cell population.

When we are conducting a time series experiment to measure the cell cycle behavior we can perform a FACS analysis along microarray measurements. As a FACS experiment is significantly simpler and cheaper to perform than a microarray experiment, the data can easily be generated in a more dense intervals. Thus, we can assume that we have obtained K FACS measurements from the time instants $T_1, \ldots, T_K$.

Let us assume that the wild type asynchronous cell population is distributed as $p(t) = 2^{(1-t)}$, $t \in [0, 1]$ (Cooper, 2004). Here, t denotes the cell

[2]By a synchronous cell population we mean that all the cells in the population are at the same phase of the cell cycle and thus, by definition have the same amount of DNA. More strict definition of synchrony would require that in addition to the same amount of DNA, all the cells should be at the same state of their life cycle, that is to be of the same age.

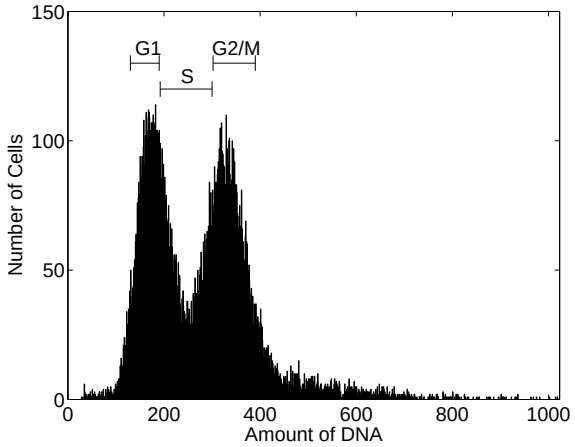

Figure 4.3: Example of the histogram that is obtained with FACS. The amount of DNA, corresponding to different phases of the cell cycle, is marked to the figure. There is about the same number of cells in all the phases of the cell cycle.

cycle phase normalized to the interval $t \in [0,1]$[3]. As we know N, the total number of cells used in the FACS measurement, we can compute the number of cells at each small time interval $[t_0, t_1]$ as $c(t) = N(2^{(1-t_0)} - 2^{(1-t_1)})$, where $t_1 > t_0$. Furthermore, the cumulative number of cells at time t is $C(t) = \sum_{i=0}^{t} c(i) = N(2 - 2^{(1-t)})$. That is, for a given t, $C(t)$ is the total number of cells at the earlier phases of the cell cycle.

By using a measured asynchronous FACS histogram h_a and the cumulative number of cells $C(t)$, we can estimate a mapping that we call the DNA replication function $f(t)$. This function maps the *number of cells - cell cycle phase* -space to *number of cells - amount of DNA* -space as

$$f(t) = \arg\min_{K} \left(\left| \sum_{i=0}^{K} h_a(i) - C(t) \right| \right), \tag{4.3}$$

where $h_a(i)$ is the value of FACS histogram of the asynchronous population at the point i, and $K \in \mathbb{N}$ making $f(t)$ to be a discrete approximation of the DNA replication function.

The function $f(t)$ presents the amount of DNA that is present at each time instant of the cell cycle. Having this information, we can use a FACS histogram of a synchronous population to evaluate the number of cells that this amount of DNA corresponds to. Thus, the distribution of the cell

[3] As the observed FACS data is discrete, for convenience the cell cycle phase variable t needs to be a discrete variable as well.

population is obtained by

$$x(t) = \sum_{i=0}^{f(t)} h_s(i) - \sum_{i=0}^{t-1} x(i), \tag{4.4}$$

where $f(t)$ is the value of the DNA replication function and $h_s(i)$ is the value from a FACS histogram of a synchronous population at the point i. The obtained distribution is obviously discrete. By using the obtained estimate of $f(t)$, this population estimation process can be repeated for all FACS measurements from the time instants $T_1, \ldots, T_K$. Thus, for each time instant T_k, we can obtain a separate estimate of the population distribution over the cell cycle phase t (Publication III).

Traditionally, a population estimate is obtained by estimating the number of cells in each phase of the cell cycle by hand, as demonstrated in Figure 4.3 (Bar-Joseph *et al.*, 2004). These cell counts are then used to draw the cell population distribution. As the proposed estimation method is automatic and non-parametric, it provides a more objective estimate of the population distribution.

Chapter 5

Information in Biology

A reductionistic approach to molecular biology, where the effects of a single gene or a group of connected genes have been studied, has helped us to understand how different parts of the organism interact and what the underlying control mechanisms are. This reductionistic approach, however, has its limitations. Focusing on studying individual genes in isolation from the rest of the system offers only a limited view to the behavior of the system. Thus, a system level approach to biology has recently become a major field of research (Ideker *et al.*, 2001).

In the system level approach computational models have an essential role. We can model the behavior of the system and make predictions on the effects of different stimuli. Building these kinds of models requires a lot of biological knowledge and extensive measurements of the system. While there have been several successful attempts to model the behavior of biological systems, these have focused only on a fraction of the system (Chen *et al.*, 2004; Gilchrist *et al.*, 2006). There is still not enough knowledge and data that would make it possible to model real biological systems in detail at the system level.

Instead of trying to model a system in detail we can use qualitative modeling (Bornholdt, 2005). In this approach the focus is on understanding general emergent properties of large networks. With this approach we can address several fundamental questions of biology. For example, why is an organism able to robustly process information from a variable environment while maintaining adaptability. This kind of a behavior is observed with several biological systems, for example with an innate immunity cell macrophage (Kitano and Oda, 2006).

To understand the behavior of real organisms at the system level, we can look at how information is propagated within and between organisms. It can be argued that studying information processing is a key factor in understanding life (Hood and Galas, 2003; Yockey, 2005). Living systems differ

from non-living systems, for example, in their ability to process information from their environment and to propagate information over time through the mechanism of evolution. This difference between living and non-living systems motivates to study biology as an information science (Yockey, 2005).

Information theory has successfully been applied in biological research to quantify the information in genetic sequences or proteins (Yockey, 2005; Krasnogor and Pelta, 2004; Kocsor *et al.*, 2005). Using genome sequences, evolutionary relationships between different organisms have been shown (Li *et al.*, 2004). Here we will show how the information-based approach can be applied to study the information processing of biological systems. We will show that the information-based approach can be used to uncover information flow at different levels. We can study how information is propagated through evolution in the structure of a regulatory network or how a system responds to a perturbation. Thus, the information-based approach can be used to study both the structural and dynamical properties of the system. We will demonstrate the applicability of our approach by using a simple computational model class, Boolean networks. In addition, we will show how the proposed approach can be applied to real data by studying the evolutionary relationships through metabolic networks of different organisms and characterizing the dynamical behavior of macrophage using time series microarray data.

5.1 Information Theory

Here some fundamental results of information theory and interesting new developments are discussed. The presented results will form basic tools that allow us to study biological systems through the concepts of information content and processing.

There are two commonly used definitions for information, Shannon information (Shannon, 1948) and Kolmogorov complexity (Kolmogorov, 1965; Solomonoff, 1964; Chaitin, 1969). Both theories provide a measure of information using the same unit: a bit. A natural interpretation of information is the length of the description of an object in bits. Here we discuss the fundamental differences between these two theories and give definitions for information. In addition, we discuss how information can be used to measure the similarity of two objects.

5.1.1 Shannon Information

In Shannon information theory the amount of information is measured by entropy. For a discrete random event x with k possible outcomes, the entropy

H is given as

$$H = \sum_{i=1}^{k} p_i I_i = -\sum_{i=1}^{k} p_i \log p_i, \tag{5.1}$$

where p_i is the probability of an event x_i to occur (Cover and Thomas, 1991). Quantity $I_i = -\log p_i$ is the information content of an event x_i. Natural interpretation for entropy is that it is the expected number of bits that are needed to encode the outcomes of a random event x. It can be observed that entropy is maximized when the probabilities of all events are equal, that is $p_i = \hat{p}, \forall i \in 1, \ldots, k$ (Cover and Thomas, 1991).

As indicated earlier, Shannon information measures information of a distribution. Thus, it is based on the underlying distribution of the observed random variable realizations. The distribution can be obtained based on assumptions about the data generation process or it can be estimated from the data. This distribution based definition has some obvious drawbacks. For example, consider the bit strings

$$11111111111111110000000000000000$$

and

$$01000101010011010110010011101101.$$

If we assume that both strings are from a random variable X with alphabet $\{0, 1\}$ they have exactly the same information content, empirical entropy $H = 1$, even though the first one obviously shows a simpler bit pattern.

5.1.2 Kolmogorov Complexity

Unlike Shannon information, Kolmogorov complexity[1] or algorithmic information is not based on statistical properties, but on the information content of the object itself (Li and Vitanyi, 1997). Thus, Kolmogorov complexity does not consider the origin of an object. The Kolmogorov complexity $K(x)$ of a finite object x is defined as the length of the shortest binary program that with no input outputs x on a universal computer. Thus, it is the minimum amount of information that is needed to generate x. Unfortunately, in practice this quantity is not computable (Li and Vitanyi, 1997).

While the computation of Kolmogorov complexity is not possible, an upper bound can be estimated using lossless compression (Li and Vitanyi, 1997). Several real-life compression algorithms, like the Huffman (Huffman, 1952), Lempel-Ziv (Ziv and Lempel, 1977), and arithmetic coding (Rissanen and Langdon, 1979) have proven to give useful approximations of Kolmogorov complexity in practical applications (Li and Vitanyi, 1997).

[1] Algorithmic information theory was independently introduced by R.J. Solomonoff (Solomonoff, 1964), A.N. Kolmogorov (Kolmogorov, 1965) and G. Chaitin (Chaitin, 1969). However, this theory is commonly known as Kolmogorov complexity.

5.1.3 Information Distance

As information is an absolute measure, related to a single object or a distribution, it is not directly suitable for comparing the similarities of two objects. Small or large information alone does not tell much about the similarity of objects. Thus, measures to jointly compare the information content of two objects have been proposed.

With Shannon information a joint entropy between two discrete random variables X and Y is defined as

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y), \tag{5.2}$$

where $p(x,y)$ is the probability of observing a pair of events x and y and the sum is computed over all the pairs of x and y (Cover and Thomas, 1991). In a similar manner we can define conditional entropy, that is the entropy of X given Y

$$H(X|Y) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} = H(X,Y) - H(Y). \tag{5.3}$$

Mutual information is one of the best known information-based measures of similarity (Cover and Thomas, 1991). It is a measure of how much information can be obtained about random variable X by observing Y. The mutual information of X relative to Y is defined as

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5.4}$$

and by using the notations of joint and conditional entropy it can be written as

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y). \tag{5.5}$$

Thus, mutual information is simply the sum of entropies of X and Y minus the joint entropy (Cover and Thomas, 1991).

Information-based similarity measures can also be defined based on Kolmogorov complexity. This topic has been studied in recent years with the goal of finding an information measure than can be approximated computationally (Bennett *et al.*, 1998; Li *et al.*, 2004).

We denote as $K(x,y)$ the length of the shortest binary program that outputs x and y, and a description how to tell them apart. Analogously to Shannon information, we can define a conditional Kolmogorov complexity $K(x|y)$ as the length of the shortest binary program that with a given input

y outputs x (Li and Vitanyi, 1997). Thus, information about y, contained in x can be defined as (Li *et al.*, 2004)

$$I(x;y) = K(y) - K(y|x). \tag{5.6}$$

It can be shown that the relation

$$K(x,y) = K(x) + K(y|x) = K(y) + K(x|y) \tag{5.7}$$

holds up to an additive precision[2] (Li and Vitanyi, 1997). Therefore, there exists a symmetry property $I(y;x) = I(x;y)$, up to an additive precision.

Kolmogorov complexity based similarity measure, or information distance, between two objects is the shortest binary program that computes x from y, or vice versa. Thus, information distance can be defined as (Bennett *et al.*, 1998)

$$d_{ID}(x,y) = \max(K(y|x), K(x|y)). \tag{5.8}$$

This is a measure of absolute information distance between two objects. As the size of an object has a direct impact to the Kolmogorov complexity of the object, we should define a normalized version of the information distance that takes the size of an object into account. A normalized information distance can be defined as (Li *et al.*, 2004)

$$d_{NID}(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}. \tag{5.9}$$

While normalized information distance can be motivated solely from the information theory point of view, it has some general properties that make it interesting in other ways. The normalized information distance has been shown to incorporate all effective computable distance metrics including, for example, the Euclidean and Hamming distances. Thus, the normalized information distance can be argued to be a universal measure of similarity.

5.1.4 Normalized Compression Distance

While normalized information distance, like Kolmogorov complexity itself, is not computable, it has been shown that this metric can be approximated by any real-life compression algorithm that fulfills several natural criteria of a *normal compressor* (Cilibrasi and Vitanyi, 2005).

Let us denote compressed length of a string x by $C(x)$. Similarly, compressed length of the concatenation of strings x and y is denoted by

[2]There is a constant $c > 0$, independent of x and y such that equalities in Equation 5.7 holds up to c additive precision.

$C(xy)$. A compressor C is considered to be *normal* if it asymptotically fulfills the following criteria (Cilibrasi and Vitanyi, 2005): 1) *Monotonicity* $C(xy) \geq C(x)$; 2) *Idempotency* $C(xx) = C(x)$, and $C(\lambda) = 0$ where λ is the empty string; 3) *Symmetry* $C(xy) = C(yx)$; and 4) *Distributivity* $C(xy)+C(z) \leq C(xz)+C(yz)$. Details about these properties can be found in (Cilibrasi and Vitanyi, 2005). Earlier work has demonstrated that several real-life compression algorithms approximate a *normal compressor* in sufficient detail (Cilibrasi and Vitanyi, 2005).

By using a compressor C instead of the Kolmogorov complexity K, we can write Equation 5.9 in a computable form. After we apply Equation 5.7, to the the numerator of Equation 5.9, the numerator can be written as $\max\{K(x,y) - K(y), K(x,y) - K(x)\}$ (Li *et al.*, 2004). For compression convenience we can approximate $K(x,y)$ by the concatenation of these strings $K(x,y) = K(xy) = K(yx)$[3]. Using these properties the normalized compression distance (NCD) can be defined as

$$d_{NCD}(x,y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}. \tag{5.10}$$

It can be shown that this approximation has the same metric properties as the normalized information distance, up to an additive constant (Li *et al.*, 2004; Cilibrasi and Vitanyi, 2005). However, it is important to understand the limitations that are faced when a *normal compressor* is approximated with a real-life compression algorithm in real-life applications.

When we estimate NCD we are inherently limited to the metrics that are covered by the compression algorithm. Thus, even though NCD is shown to be quasi-universal (Cilibrasi and Vitanyi, 2005), this does not hold in applications where real compression algorithms are used. However, if the compression algorithm is able to uncover the similarities that are of interest in the underlying analysis task, NCD will be an effective analysis tool as has been shown in several applications (Cilibrasi and Vitanyi, 2005; Krasnogor and Pelta, 2004). Furthermore, it has been observed that the performance of the NCD is not dependent of the compression algorithm, but several very different algorithms will yield consistent results (Cilibrasi and Vitanyi, 2005).

As NCD is an asymptotic approximation of the normalized information distance it only holds only up to an additive precision. In addition, as Kolmogorov complexity is not computable, we can not directly determine how good our approximation is. These problems are observed in practice in the dynamic range of NCD. In theory, NCD should cover the range $[0, 1]$ of distances. In practice, the observed distances does not cover this full

[3]This holds up to an additive precision. In addition to objects x and y we need to encode the separator between these objects in the term $K(x,y)$.

range and in some cases the distances may even exceed one. The range of observable distances is related to how accurately the normalized information distance is approximated (Cilibrasi and Vitanyi, 2005). Thus, the range is dependent of the amount of data and the compression algorithm. Even though the range of distances is more limited when a small amount of data is used, NCD is still able to uncover the differences between different objects remarkably well.

Additional problems are caused by the limitations in the implementations of the compression algorithms. For example, the popular gzip compression program is implemented using a block size of 32 kilobits. This means that if the length of a bit string is, for example, 100 kilobits similarities between x and y in the estimation of the term $C(xy)$ in Equation 5.10 are not found. This is because the codebook that is used in compression is cleared always after 32 kilobit block of data. Thus, if the amount of data is less than 32 kilobits the assumption about stream-basedness of the gzip compressor does not hold. Other limitation comes from the fact that Lempel-Ziv algorithm only looks for repetitions in a bit string. Thus, the estimated NCD between bit strings x and $\bar{x}$, where $\bar{x}$ is x with all bits flipped, is approximately one even though these strings are obviously similar. Even with these limitations, gzip is a powerful compression tool and it has been shown to perform well in real applications (Cilibrasi and Vitanyi, 2005).

5.2 Discrete Networks

Discrete network models are commonly used to model genetic regulatory networks at the system level (Aldana *et al.*, 2003; Bolouri and Davidson, 2002; Shmulevich *et al.*, 2003). Even though these models introduce numerous assumptions and simplifications, significant insights about the behavior and structure of biological systems have been obtained (Bornholdt, 2005; Barabási and Albert, 1999).

When discrete computational models are used to model biological systems one needs to understand the limitations of the model and thus, use the model only to address the questions that can be answered reliably at the selected level of abstraction. For example, the model class may operate in synchronous fashion. This is clearly not true in real biological systems and the significance of this assumption depends on whether it is an important characteristic in our modeling task. Similarly, the model can be simplified by using abstract regulators so that genes and proteins can not be distinguished. In the model, the regulators are only considered as nodes in the network. Furthermore, we may choose not to consider the quantities of reaction products. It may be enough to know whether the product is present or absent. When we are working with discrete systems, the data we are

processing is inevitably discrete. This can be a useful property as the quantization reduces noise in the data, or it may be a problem as the dynamic range of the data can be very limited.

Here we will focus on the Boolean network model (Kauffman, 1993; Erdös and Rényi, 1959). This is a simple dynamical system model where each node can have only two possible states, on or off. Despite the apparent simplicity, this model class is able to produce highly complex behavior, for example, in the form of a phase transition. Furthermore, as this model has been studied extensively, several of its properties are well understood (Kauffman, 1993; Aldana *et al.*, 2003; Kesseli *et al.*, 2005, 2006). Thus, as there are several earlier results which our approach can be compared, this model is an excellent choice for the illustration of our information-based analysis approach.

Boolean network model class can be defined as follows. Let $s_i(t) \in \{0,1\}$, $i = 1, \ldots, N$, where N is the number of nodes in the network, be the state of i:th node in a Boolean network at time t. The state of this node at time $t+1$ is determined by the states of nodes $j_1, j_2, \ldots, j_{k_i}$ at time t as

$$s_i(t+1) = f_i(s_{j_1}(t), s_{j_2}(t), \ldots, s_{j_{k_i}}(t)), \tag{5.11}$$

where $f_i : \{0,1\}^{k_i} \to \{0,1\}$ is a Boolean function of k_i variables. A binary vector $\boldsymbol{s}(t) = (s_1(t), \ldots, s_N(t))$ is the state of the network at time t. In the classical model, all nodes are updated synchronously as the system transitions from state $\boldsymbol{s}(t)$ to state $\boldsymbol{s}(t+1)$ (Kauffman, 1993). It should be noted that this model can directly be generalized to a larger alphabet by defining $s_i(t) \in \{0, \ldots, L-1\}$ and $f_i : \{0, \ldots, L-1\}^{k_i} \to \{0, \ldots, L-1\}$, where L is the size of the alphabet.

To construct a Boolean network, the inputs $j_1, j_2, \ldots, j_{k_i}$ for each node i needs to be determined. This can be done by selecting the inputs randomly among all N nodes or by selecting the inputs using some systematic pattern. The number of inputs k_i can be endowed with a probability distribution, such as the power-law (Barabási and Albert, 1999; Aldana and Cluzel, 2003) or Poisson distribution, with a mean $K = E[k_i]$. The mean K is known as the average connectivity of the network.

Once the connections have been set, we can choose a Boolean function f_i for each node. Functions can be parameterized by the bias $b = E[f_i]$, the probability that the function outputs one on an arbitrary input vector. If $b = 0.5$, then the function is said to be unbiased. The functions can be selected randomly among all $2^{2^{k_i}}$ Boolean functions or they can be selected from some class of functions (Stauffer, 1987; Shmulevich *et al.*, 2003; Kauffman, 2000; Harris *et al.*, 2002). If both the functions and connections are selected randomly, then the obtained network is called random Boolean network (RBN) (Kauffman, 1993).

As a Boolean network is a discrete system, it has a finite state space. A boolean network with N nodes has 2^N different states. Thus, the state space is $S = \{0, 1\}^N$. We can define a transition from state $\boldsymbol{s}_i$ as $F(\boldsymbol{s}_i)$ where $F = (f_1, \ldots, f_N)$ and f_i is the Boolean function of node i with predetermined connections from the nodes $j_1, j_2, \ldots, j_{k_i}$. As the state space is finite, at some point any trajectory, that is a path from any initial state, will return to one of the previously visited states. This kind of a state cycle where the same states are repeated infinitely is known as an attractor cycle and the states within the cycle are called attractor states. A set of states that leads to the same attractor is called the basin of attraction (Wuensche, 1999).

5.2.1 Quantization of Microarray Data

Measurement data that can be obtained with microarray technology is continuous in nature[4]. To utilize measurement data in the context of discrete models, the data needs to be quantized into elements of a discrete alphabet.

In quantization some information contained in the data is lost. However, quantization can be seen as a noise removal process. Reducing the precision of data representation removes noise from the data and makes it possible to emphasize meaningful trends in the data. Thus, the quantization algorithm needs to make a trade-off between data presentation accuracy and noise reduction. If the modeling approach is qualitative, then it may be of interest to quantize to a small number of levels, even to a binary domain. Even in the binary domain the most important characteristic of gene expression, whether the gene is regulated or not, is captured.

Standard approach for data quantization is to cluster the samples in k classes using the k-means clustering algorithm (Shmulevich *et al.*, 2005). To take the noise into account in the quantization, a noise floor can be applied before the k-means clustering. Purpose of the noise floor is to remove all variation, that can be assumed to be due to the noise and are not of biological origin. This can be done by setting all the intensity values below a threshold to a constant value.

After the quantization, the information content of microarray data can be estimated in a more straightforward manner. Estimation of Kolmogorov complexity from the quantized data can be considered as a lossy compression of the original data or alternatively the compression of essential features of the data. As the unquantized microarray data is extremely noisy, information can not be estimated reliably by using a general purpose compression

[4]Microarray data is actually discrete with the dynamic range depending on the scanner. However, after prepossessing the obtained intensity values are floating point numbers and thus effectively continuous.

algorithm. Thus, the compression of microarray data would require a specially designed compression algorithm. So far this has not been seen as an important research problem and only very little work has been done in order to compress microarray expression data effectively (Jörnsten, 2001).

If we estimate information, or Kolmogorov complexity, from the quantized microarray data we can compare the samples in terms of their information content. By using normalized information distance instead of a more specific distance measure like correlation, we can potentially uncover more detailed similarities between the samples. In addition, by estimating the information content we avoid the problem of feature subset selection. In practice, to reliably separate different classes of samples using a traditional distance measure, a subset of informative features needs to be identified. Thus, the class separation accuracy is dependent on the selected features. With information distance, we can obtain a reliable separation even with all the features (Publication V).

5.3 Structure

Properties of a network are related to its structure. Traditionally, networks have been analyzed assuming that the connections between different nodes are selected randomly (Erdös and Rényi, 1959; Kauffman, 1969, 1993). Recent discoveries have shown that this assumption does not hold for most real world networks (Albert and Barabási, 2002; Barabási, 2002; Babu *et al.*, 2004; Guelzim *et al.*, 2002). Instead, several networks, including gene regulatory networks show a scale free structure (Barabási, 2002). Characteristic property of a scale free network is the existence of hubs, that is, the nodes that have a very high number of connections. In a random topology all the nodes have approximately the same number of connections.

Network structure determines the robustness of the network to structural perturbations. If the nodes are knocked out randomly from the network, then a network with a scale free structure is highly robust to structural perturbations (Albert and Barabási, 2002). Knockouts have a significant effect to a scale free network only if a knockout hits a hub. However, if the number of nodes is high compared with the number of hubs, probability for a hub knockout is very small.

Structure of the network can be characterized using summary statistics, that can be computed for any given network (Watts and Strogatz, 1998; Aldana *et al.*, 2003). Input and output degrees are defined as a distribution of the number of inputs and outputs for each node. Input and output degree distributions characterize the connectivity of a network. For example, if a network has a scale free input degree distribution, then there must exist highly connected nodes, that is, the hubs.

Clustering coefficient is a measure for the connectivity of a network (Watts and Strogatz, 1998). It is defined for a given node as the number of neighboring nodes that are connected to each other. That is, for a set of nodes $N = n_1, \ldots, n_k$ we have a set of connections (edges) $E = \{e_{ij}\}$, where $i, j \in 1, \ldots, k$. Thus e_{ij} is an edge between the nodes n_i and n_j. We can define a neighborhood B for the node n_i as its immediately connected neighbors $B_i = \{n_j\} : e_{ij} \in E$. The connectivity k_i of the node n_i is the size of the neighborhood $|B_i|$. The clustering coefficient C_i for the node n_i is the proportion of links between the neighborhood nodes divided by the number of links that could possibly exist. For each neighborhood the maximum number of links is $k_i(k_i - 1)$. Thus, the clustering coefficient is given as

$$C_i = \frac{|\{e_{lm}\}|}{k_i(k_i - 1)} : n_l, n_m \in B_i, e_{lm} \in E. \tag{5.12}$$

The clustering coefficient for the whole network is the average of the clustering coefficients of all the nodes

$$\hat{C}_i = \frac{1}{n} \sum_{i=1}^{k} C_i. \tag{5.13}$$

Another measure of network topology is characteristic path length (Watts and Strogatz, 1998). First, the path length L_{ij}, that is the minimum number of edges that are needed to get from the node n_i to the node n_j, is computed. The characteristic path length L is then L_{ij} averaged over all pairs of nodes.

Topological statistics like clustering coefficient and characteristic path length can be used to determine the type of network and to compare different network topologies (Watts and Strogatz, 1998). For a network with regular wiring the clustering coefficient and characteristic path length are both high, whereas in a random network both statistics have a small value. Other network topologies can have a high clustering coefficient and still a low average path length. Networks with this kind of topology are known as small world networks and they have the property that $L > L_r$ but $C \gg C_r$, where L_r and C_r denotes the characteristic path length and clustering coefficient of a random network, respectively. Usually this kind of a network also has a scale free topology (Watts and Strogatz, 1998).

While these topological statistics can successfully be used to compare and classify different types of networks, it is not obvious what measures are able the uncover all the interesting characteristics of a network. Furthermore, measures like the characteristic path length and clustering coefficient are most useful in the comparison of different topologies. They are not that informative when, for example, two scale free networks are compared. This is a problem if we want to compare networks that have the same topological properties.

Instead of computing individual statistics from networks, we should compare the entire networks directly. While there are several aspects that make this comparison difficult, for example a difference in the number of nodes, we can do the comparison using the information-based approach. Thus, we can compare the networks by their information content (Publication VI).

To demonstrate the benefits of the information-based network analysis, we downloaded metabolic network structures for 107 different organisms from the KEGG database (Kanehisa and Goto, 2000; Ma and Zeng, 2003). The information content of these networks has been formed through millions of years of evolution. Thus, it is expected that the information distance between the species will be a function of their evolutionary history. The information distance is a powerful tool for reconstructing phylogenies, as has been shown by applying the NCD to mitochondrial genomes (Li *et al.*, 2004).

It has been shown that the choice of distance metric such as the Jaccard index, the Simpson index, and the Korbel distance, all of which are defined in terms of the number of enzymes within the organisms and shared between the organisms, produce different phylogenies from the metabolic networks (Zhu and Qin, 2005; Podani *et al.*, 2001). By using the information-based approach this issue can be avoided. A phylogenetic tree, computed using the NCD to measure information between the metabolic networks and using the complete linkage method to construct a tree from the obtained distances is shown in Figure 5.1. It should be noted that to compress networks effectively we need to present the network structure in a form from which the compression algorithm can find similarities effectively. Details about the network presentation are discussed in Publication VI.

This result shows that the organisms are clearly grouped into the three domains of life. The bacteria form three distinct clades, with parasitic bacteria, encoding more limited metabolic networks, separating from the others as has been observed previously (Podani *et al.*, 2001). The fact that the phylogenetic tree reproduces the known evolutionary relationships between species suggests that closely related organisms are also close in terms of the information content of their networks.

5.4 Dynamical Behavior

While the structural analysis of a system can uncover important insight into the robustness and connections between different components, it does not consider the behavior of the system. To understand the behavior we need to look at the dynamics of the system, for example by studying its response to perturbations. Most dynamical systems can operate in the ordered or

Figure 5.1: A phylogenetic tree generated using NCD applied to all pairs of metabolic network structures from 107 organisms in the KEGG. Bacteria are shown in red, archaea in blue, and eukaryotes in green. Subclasses are labeled on the right.

chaotic regime or at the phase transition boundary between the two regimes (Aldana *et al.*, 2003). This phase transition area can also be referred to as the edge of chaos (Kauffman, 1993).

When a network is operating in the ordered regime, it is intrinsically robust while its dynamical behavior is simple. The robustness can be observed through both the structural and transient perturbations. Perturbations of any size have a small effect to the behavior of the network. Networks in the chaotic regime, on the other hand, are extremely sensitive to perturbations. Even a small perturbation will quickly propagate through the entire network. Thus, networks in the chaotic regime are not robust and will fail under perturbations. A phase transition between the ordered and chaotic regimes represents a tradeoff between the need for stability and the need to have a wide range of dynamical behavior to respond to variable perturbations (Kauffman, 1993).

By varying the parameters K and b in the random Boolean network model, dynamical phase transition can take place. The parameter

$$\sigma = 2b(1-b)K \tag{5.14}$$

determines the dynamical regime. If $\sigma > 1$ then the system is chaotic and for $\sigma < 1$ the system is ordered. (Derrida and Pommeau, 1986; Luque and Sole, 1997, 2000; Shmulevich and Kauffman, 2004). It is easy to see that for unbiased random Boolean networks the critical connectivity is $K_c = 2$.

Dynamical behavior of a system can be characterized using an order parameter. For random Boolean networks one order parameter is the slope of the Derrida curve (Derrida and Pommeau, 1986). This order parameter is based on the annealed approximation of a Boolean network. That is, the state of a node in a network is determined based on a distribution of all possible states of the node (Derrida and Pommeau, 1986).

The Derrida curve is defined as follows. Let $\boldsymbol{s}^{(1)}(t)$ and $\boldsymbol{s}^{(2)}(t)$ be two states of the system at time t. A normalized Hamming distance between the states is $d(t) = \frac{1}{n}\sum_{i=1}^{N}(s_i^{(1)}(t) \oplus s_i^{(2)}(t))$, where $\oplus$ is XOR operator and N is the number of nodes. The Derrida curve can be drawn by plotting the expected distance $d(t+\Delta t)$ versus the distance $d(t)$. The expectation here is relative to the distribution over the state space of a particular system or over some ensemble of systems, or both (Shmulevich and Kauffman, 2004). In practice, the state space of a dynamical system can be sampled for constructing an empirical Derrida curve (Figure 5.2). If the slope of the Derrida curve at the origin is greater than 1, then the system can be said to be chaotic; if less than 1, ordered; and if equal to 1, critical.

In addition to the Derrida curve several other order parameters have been proposed for Boolean networks. In the case of random Boolean networks these are all equivalent in terms of the phase transition (Shmulevich

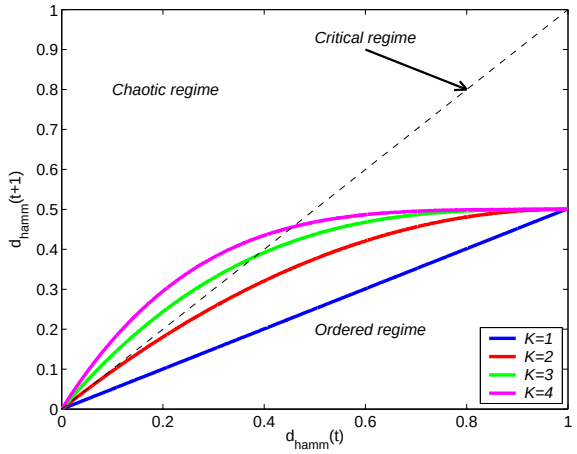

Figure 5.2: Derrida curve for Boolean networks from ensembles $K = 1, 2, 3, 4$ with $b = 0.5$. The slope of the Derrida curve at the origin determines the dynamical regime.

and Kauffman, 2004; Flyvbjerg, 1988; Luque and Sole, 2000). Order parameters are usually defined in the context of a specific model class. Thus, the definition is dependent on the selected distance metric, in the case of Derrida curve, the Hamming distance. Making the definition of an order parameter dependent of a model class poses limitations for measuring the behavior. For example, the order parameter can only be used with one type of a model class and thus, the properties of different model classes can not be compared. Furthermore, the purpose of an order parameter is to study the propagation of information through the system. Thus, instead of looking at the propagation of individual bits, it is more justified to study the propagation of information.

We propose a new information-based order parameter for measuring the information propagation through a system. This measure is based on the normalized information distance and thus it can directly be applied to any model class as it makes no assumptions about the model or the alphabet the model is using. We have defined our order parameter analogously to the Derrida curve. Instead of using the Hamming distance as the measure of similarity, we are using the normalized information distance. In computational applications normalized compression distance can be used as an approximation. Thus, the information-based Derrida curve is obtained by computing the distances between the states $\boldsymbol{s}^{(1)}(t)$ and $\boldsymbol{s}^{(2)}(t)$ using $d(t) = d_{NCD}(\boldsymbol{s}^{(1)}(t), \boldsymbol{s}^{(2)}(t))$ (Figure 5.3).

When compared with the traditional Derrida curve for random Boolean networks, our information based version has an interesting property. For

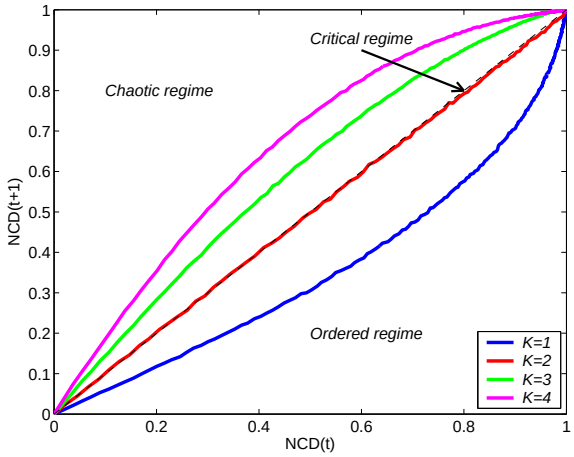

Figure 5.3: Information-based Derrida curve for Boolean networks from ensembles $K = 1, 2, 3, 4$ with $b = 0.5$. The dynamical regime can be observed throughout the curve.

a critical network the curve stays at the diagonal for all the distances, not just close to the origin. With the traditional approach that is based on the Hamming distance the dynamical regime can be characterized only by using very small perturbations, as the order parameter is defined by the slope at the origin. Our information-based version allows us to use perturbations of any size as the same dynamical behavior is observed throughout the curve. For example, when a stimulus is given to a biological system, it is usually not known what the exact response is. Thus, our measure allows the usage of biological data even though the size of the response, or the perturbation, is not known.

We have applied our information-based order parameter to real microarray data, measured from a mouse macrophage, with the aim to characterize the dynamical behavior of a living system. The macrophage is an innate immune cell responsible for initiating the host defense against an infection. The macrophage is able to recognize a broad variety of pathogens and rapidly mount appropriate responses to each (Aderem, 2001). Thus, to perform these functions the macrophage needs to be both robust and adaptable (Kitano and Oda, 2006).

To characterize the dynamical behavior we need measurement data that shows responses to perturbations. For the macrophage several ligands that are known to cause a different response are available (Aderem, 2001). Thus, as we measure the response for each stimulus at times t and $t + 1$ we can construct the information-based Derrida curve that shows how the perturbations propagate in the system.

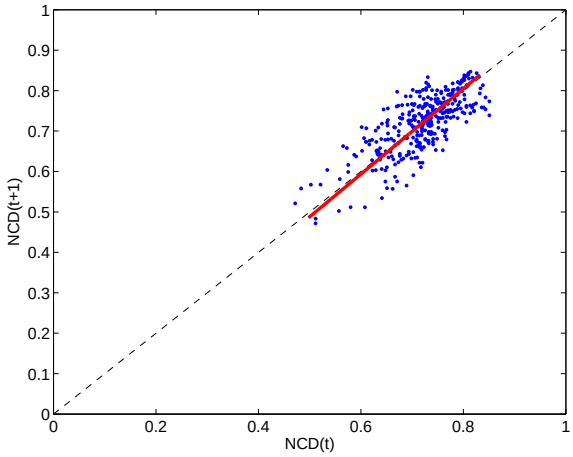

Figure 5.4: Information-based Derrida curve computed using the time-course microarray data from the murine bone marrow-derived macrophages treated with various toll-like receptor stimuli. The red line shows the least-squares fit of the data points with residual distances being orthogonal to the main diagonal.

We have generated a microarray dataset that includes time series measurements from six different stimuli (Gilchrist *et al.*, 2006). To reduce the amount of noise in the data we choose to binarize the measurement data using the k-means approach. After computing the distances between all the pairs of stimuli at times t and $t+1$, an information-based Derrida curve can be constructed. The result is shown in Figure 5.4 (Nykter *et al.*, 2006). Based on our order parameter, dynamics of the macrophage seems to operate at the phase transition boundary between order and chaos. Thus, this observation supports the hypothesis that living systems operate at the edge of chaos (Kauffman, 1993).

5.5 Correlation of Structure and Dynamics

Information-based analysis allows us to use the same approach for comparing both the dynamical and structural similarities between networks. In addition to the order parameters, the dynamical behavior between different networks can be compared by measuring the similarity of time series data. Similar networks should produce dynamics of similar complexity. The structural comparison of networks can be done as discussed in Section 5.3. As the same approach can be used for both structural and dynamical comparisons, this allows us to study whether there is a correlation between dynamical and structural complexity.

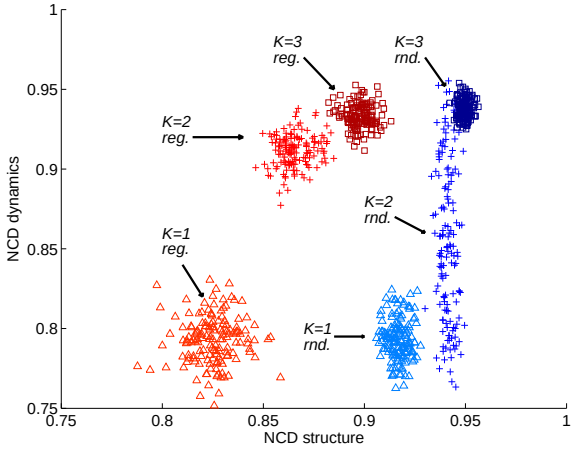

Figure 5.5: The normalized compression distance (NCD) applied to network structure and dynamics. Six ensembles of random Boolean networks ($K = 1, 2, 3$ each with random or regular topology; $N = 1000$) were used to generate 150 networks from each ensemble.

We draw several networks from different network ensembles. As an example we used Boolean networks with the connectivity $K = 1, 2, 3$ and with regular and random wiring. After computing the distances between different networks within the ensembles, the results can be illustrated as shown in Figure 5.5.

This result shows that there is a clear correlation between the dynamical and structural complexity. A complex structure will yield more complex dynamics. It is interesting to observe that the networks at the critical regime show the most variation in dynamical behavior, ranging in dynamics from ordered to chaotic. This observation further supports the hypothesis that networks at the critical regime are the most evolvable. Even though we have demonstrated the applicability with Boolean networks, this approach is directly applicable to any model class.

5.6 Basin Structure

So far we have applied the information-based approach to compare the network structure and dynamics and to define an order parameter for the dynamical behavior. As the analysis of dynamics is based on sampling the state space of a network, it can be argued that analyzing the entire state space as a whole would give a more global view to the properties of the system. As the size of the state space grows exponentially with the number

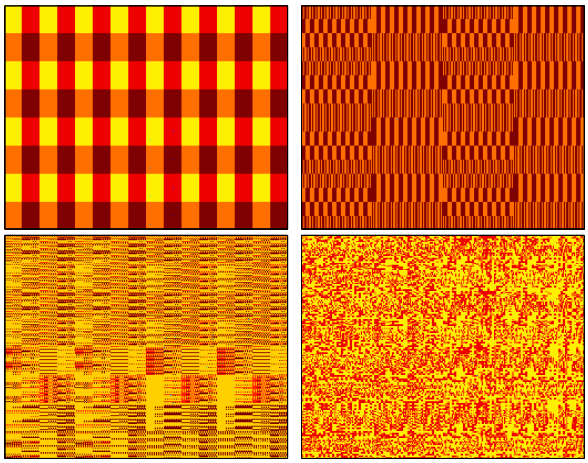

Figure 5.6: Basin of attraction illustration for $K = 1$ (top left), $K = 2$ (top right), $K = 3$ (bottom left), and $K = 4$ (bottom right) random Boolean networks with $N = 16$ nodes.

of nodes in the network, it is evident that this kind of an analysis can not be done for very large networks (Wuensche, 1999).

By extracting features from the state space, networks from different ensembles can be distinguished. Useful state space properties that can successfully be used as features include the total number of attractor states, number of Garden of Eden states and transient lengths (Publication VII).

In addition, networks from different ensembles can be compared in terms of their basin structure. A basin includes all the states of the state space that lead to the same attractor cycle. Basins can be illustrated as a tree, where each node is a state and each edge is a state transition (Wuensche, 1999). Alternatively, the basin structure can be illustrated as a two dimensional grid, where each point is a state (Publication VII). This grid can be obtained by using $2^{\lfloor N/2 \rfloor}$ least significant bits as indices at the vertical axis and $2^{\lceil N/2 \rceil}$ most significant bits as indices at the horizontal axis; analogously to the construction of the Karnaugh map (Karnaugh, 1953). By assigning all the states in a given basin the same color, we have a color image presentation of the basin structure. This kind of an illustration shows a clear separation between the networks from different dynamical regimes (Figure 5.6).

Instead of visually comparing the images, information-based approach can be used to compare the information content of the basins. Using the same coding for the basin structure, that is, a different symbol for the states in each basin, we can compute the information distance between the basins. An example of the basin comparison is shown in Figure 5.7. It can be seen that while there are a number of outliers, different dynamical regimes can

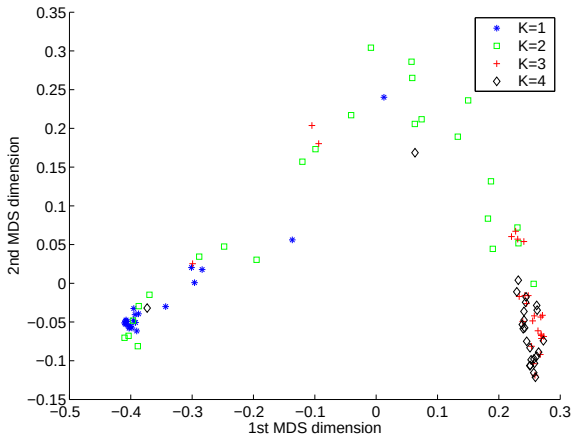

Figure 5.7: MDS presentation of the NCD matrix computed from the state space presentations. Ensembles from different dynamical regimes are clearly separable.

still be observed. As we are working with small networks, $N = 16$ in this example, it is not surprising to see outliers. Again, we observe that the networks at the critical regime show the most variation, while the networks in the ordered and chaotic regimes are more compactly clustered. This is consistent with the observation that was made when structure and dynamics were compared.

5.7 General Laws of Biology

Since systems biology is a relatively new field of research, general laws for the system behavior have not yet been formulated. However, there are several properties that are observed over and over again with different organisms and in different contexts.

The most well known general property of biological systems is the scale free structure of the regulatory networks (Albert and Barabási, 2002; Babu *et al.*, 2004; Guelzim *et al.*, 2002). Scale free networks are common in the real world appearing, for example, in the world wide web, social networks, power grids, phone lines, and in biological systems (Barabási, 2002). Most of the identified biological networks show a scale free structure (Babu *et al.*, 2004; Guelzim *et al.*, 2002). For example, the gene regulatory network of *E. coli* which has a scale free output and Poisson input degree connectivity (Salgado *et al.*, 2006). In addition, metabolic networks that have been identified for hundreds of organisms show a scale free structure (Jeong *et al.*, 2000; Zhu and Qin, 2005), although this property is inherently dependent on how the network structure is presented (Arita, 2004). The observation that scale free

topology is a fundamental property of many biological systems has helped to form hypotheses and to understand the properties of biological systems.

Another hypothesized general property of biological systems is that the dynamics of the system operate at the critical regime, at the edge of chaos (Kauffman, 1995). While the evidence for this hypothesis is not yet as convincing as the evidence for the scale free property, several independent pieces of evidence that support critical or slightly ordered dynamics do exist (Shmulevich *et al.*, 2005; Serra *et al.*, 2004; Rämö *et al.*, 2006; Nykter *et al.*, 2006);

We believe that in the future the information-based approach will prove to be an important tool in uncovering general laws for biology. Information processing is a key property of all living systems. Thus, information is a powerful tool that can be used to understand how systems behave, evolve and interact. In this chapter we have presented some basic approaches that can be used to address these questions. However, there is still a need to further develop the methodology before the full power of the information-based approach can be utilized.

Chapter 6

Conclusions

We have introduced signal processing methods that can be used to estimate and improve the quality and reliability of microarray data and data analysis algorithms. First, we discussed the basic methodology of microarray data analysis and applied the methods to real biologically motivated data analysis tasks. Next, we discussed the reliability of clinically defined class labels for cancer tumors. We showed that as the classification of different cancer types evolves over time, clinical databases do not provide a reliable source for the ground truth information of class labels. Thus, to obtain reliable class labels, all the samples should be re-classified by a single pathologist. We demonstrated how the unsupervised learning approach can be used to gain confidence about the class labels. Our case study showed that multidimensional scaling can be used to verify the reliability of class labels. This is an important sanity check, since if the supervised analysis would be done using incorrect class labels, the conclusions would be erroneous.

As the second computational approach we discussed the identification and quantification of microarray experiment error sources. As we do not have the ground truth information about microarray data available, we proposed using knowledge about microarray noise characteristics to simulate data with realistic biological and statistical characteristics. Simulated microarray data can then be used to validate data analysis algorithms or to improve the experimental setup. As we have a detailed simulation model available, we can study the effects of each error source and focus on improving those steps that have the largest effect to the quality of obtained data.

The third computational method involves using supplemental measurement data in addition to microarray measurements. As a case study, we introduced a computational method that can be used to estimate the distribution of a synchronous cell population. Our approach is based on using a fluorescent activated cell sorter to measure the number of cells at each

phase of the cell cycle. The obtained cell count histograms are then used to estimate the distribution of the cell population. The obtained distribution estimates can then be used to improve the quality of microarray data, for example, by inverting the effects of cell population asynchrony by deconvolution.

Next we showed how an information-based approach can be used to analyze biological systems at the system level. We used a Kolmogorov complexity based measure of similarity to compare different network structures and to quantify the dynamical behavior. This analysis showed that by analyzing information processing and flow in a system, we can uncover important insight into the properties of the system. By studying the structure of metabolic networks we showed that a phylogenetic tree can be built solely on the basis of how information has been propagated from one organism to another in evolution. In addition, we used microarray data to quantify the dynamical behavior of an innate immunity cell macrophage. We showed that the robustness and adaptability that have experimentally been observed in the macrophage can be explained by the fact that information propagation of the macrophage has the characteristics of the critical regime.

Bibliography

Access Excellence at the National Health Museum (2006). http://www.accessexcellence.org/. Retrieved Oct 26.

Aderem, A. (2001) Role of toll-like receptors in inflammatory response in macrophages. *Critical Care Medicine,* 29(Suppl 7), S16–S18.

Affymetrix (2006). http://www.affymetrix.com/. Retrieved Oct 26.

Agilent Technologies (2006). http://www.agilent.com. Retrieved Oct 26.

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics,* 74(1), 47–97.

Aldana, M. and Cluzel, P. (2003) A natural class of robust networks. *Proceedings of the National Academy of Sciences USA,* 100(15), 8710–8714.

Aldana, M., Coppersmith, S. and Kadanoff, L. P. (2003) Boolean dynamics with random couplings. In *Perspectives and Problems in Nonlinear Science. A Celebratory Volume in Honor of Lawrence Sirovich*, (Kaplan, E., Marsden, J. E. and Sreenivasan, K. R., eds), Springer Applied Mathematical Sciences Series. Springer-Verlag, New York, 23–89.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A. and Boldrick, J. C. (2000) Distinct types of diffuse large B-cell lymphoma identied by gene expression proling. *Nature,* 403, 503–511.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences USA,* 96(12), 6745–6750.

Antonescu, C. R., Viale, A., Sarran, L., Tschernyavsky, S. J., Gonen, M., Segal, N. H., Maki, R. G., Socci, N. D., DeMatteo, R. P. and Besmer, P. (2004) Gene expression in gastrointestinal stromal tumors is distinguished by KIT genotype and anatomic site. *Clinical Cancer Research,* 10(10), 3282–3290.

Arita, M. (2004) The metabolic world of *Escherichia coli* is not small. *Proceedings of the National Academy of Sciences USA,* 101(6), 1543–1547.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *Nature Genetics,* 25(1), 25–29.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology,* 14(3), 283–291.

Balagurunathan, Y., Dougherty, E. R., Chen, Y., Bittner, M. L. and Trent, J. M. (2002) Simulation of cDNA microarrays via a parameterized random signal model. *Journal of Biomedical Optics,* 7(3), 507–523.

Bar-Joseph, Z., Farkash, S., Gifford, D. K., Simon, I. and Rosenfeld, R. (2004) Deconvolving cell cycle expression data with complementary information. *Bioinformatics,* 20(Suppl 1), i23–i30.

Barabási, A.-L. (2002) *Linked: The New Science of Networks.* Perseus Books Group, Cambridge, Massachusetts.

Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science,* 286(5439), 509–512.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,* 57(1), 289–300.

Bennett, C. H., Gacs, P., Li, M., Vitanyi, P. M. B. and Zurek, W. (1998) Information distance. *IEEE Transactions on Information Theory,* 44(4), 1407–1423.

Blake, W. J., Kærn, M., Cantor, C. R. and Collins, J. J. (2003) Noise in eukaryotic gene expression. *Nature,* 422(6932), 633–637.

Bolouri, H. and Davidson, E. H. (2002) Modeling transcriptional regulatory networks. *Bioessays,* 24(12), 1118–1129.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* 19(2), 185–193.

Borg, I. and Groenen, P. J. F. (2005) *Modern Multidimensional Scaling.* Springer Series in Statistics, 2nd edition, Springer, New York.

Bornholdt, S. (2005) Systems biology: less is more in modeling large genetic networks. *Science,* 310(5747), 449–451.

Brownstein, M. J. and Khodursky, A. B. (2003) *Functional Genomics Methods and Protocols.* Methods in Molecular Biology, Humana Press, Totowa, New Jersey.

Buck, M. J. and Lieb, J. D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics,* 83(3), 349–360.

Chaitin, G. J. (1969) On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the Association of Computer Machinery,* 16(1), 145–159.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods of Data Analysis.* The Wadsworth statistics/probability series, Duxbury Press, Boston, Massachusetts.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. and Fodor, S. P. A. (1996) Accessing genetic information with high-density DNA arrays. *Science,* 274(5287), 610–614.

Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novak, B. and Tyson, J. J. (2004) Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell,* 15(8), 3841–3862.

Cho, H. and Lee, J. K. (2004) Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics,* 20(13), 2016–2025.

Churchill, G. A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nature Genetics,* 32(Suppl), 490–495.

Cilibrasi, R. and Vitanyi, P. (2005) Clustering by compression. *IEEE Transactions on Information Theory,* 51(4), 1523–1545.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* 74(368), 829–836.

Coombs, N. J., Gough, A. C. and Primrose, J. N. (1999) Optimisation of DNA and RNA extraction from archival formalin-fixed tissue. *Nucleic Acids Research,* 27(16), e12.

Cooper, S. (2004) Bacterial growth and division. In *Encyclopedia of Molecular Cell Biology and Molecular Medicine Volume 1*, (Meyers, R. A., ed.). 2nd edition, Wiley, Hoboken, New Jersey.

Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory.* Wiley-Interscience, Hoboken, New Jersey.

Crick, F. (1970) Central dogma of molecular biology. *Nature,* 227(5258), 561–563.

Csete, M. E. and Doyle, J. C. (2002) Reverse engineering of biological complexity. *Science,* 295(5560), 1664–1669.

Cui, X. and Churchill, G. A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology,* 4(210).

de Jong, H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology,* 9(1), 67–103.

de Schipper, J. P., Liem, R. S., van den Ingh, H. F. and van der Harst, E. (2004) Revision of gastrointestinal mesenchymal tumours with CD117. *European Journal of Surgical Oncology,* 30(9), 959–962.

Dembélé, D. and Kastner, P. (1999) Fuzzy c-means method for clustering microarray data. *Nucleic Acids Research,* 27(16), e12.

Derrida, B. and Pommeau, Y. (1986) Random networks of automata: a simple annealed approximation. *Europhysics Letters,* 1, 45–49.

Ding, C. and Peng, H. (2003) Minimum redundancy feature selection from microarray gene expression data. In *Proc. Computational Systems Bioinformatics*, 523–528, Stanford, California.

Dror, R. O., Murnick, J. G., Rinaldi, N. J., Marinescu, V. D., Rifkin, R. M. and Young, R. A. (2003) Bayesian estimation of transcript levels using a general model of array measurement noise. *Journal of Computational Biology,* 10(3–4), 433–452.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. M. (1999) Expression profiling using cDNA microarrays. *Nature Genetics,* 21(Suppl), 10–14.

Durbin, B. P., Hardin, J. S., Hawkins, D. M. and Rocke, D. M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics,* 18(Suppl 1), S105–S110.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA,* 95(25), 14863–14868.

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., Weiss, R. A. and Liotta, L. A. (1996) Laser capture microdissection. *Science,* 274(5289), 998–1001.

Erdös, P. and Rényi, A. (1959) On random graphs. *Publicationes Mathematicae,* 6, 290–297.

Flyvbjerg, H. (1988) An order parameter for networks of automata. *Journal of Physics A: Mathematical and General,* 21(19), L955–L960.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J. and Eisen, M. B. (2004) Noise minimization in eukaryotic gene expression. *PloS Biology,* 2(6), e137.

Getz, G., Levine, E. and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences USA,* 97(22), 12079–12084.

Gilchrist, M., Thorsson, V., Li, B., Rust, A. G., Korb, M., Kennedy, K., Hai, T., Bolouri, H. and Aderem, A. (2006) Systems biology approaches identify ATF3 as a negative regulator of innate immunity. *Nature,* 441(7090), 173–178.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* 286(5439), 531–537.

Gottardo, R., Raftery, A. E., Yeung, K. Y. and Bumgarner, R. E. (2003). Robust estimation of cDNA microarray intensities with replicates. Technical Report 438 Department of Statistics, University of Washington.

Guelzim, N., Bottani, S., Bourgine, P. and Kepés, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics,* 31(60), 60–63.

Haab, B. B., Dunham, M. J. and Brown, P. O. (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology,* 2(2).

Harris, N. L., Jaffe, E. S., Stein, H., Banks, P. M., Chan, J. K., Cleary, M. L., Delsol, G., Wolf-Peeters, C. D., Falini, B., Gatter, K. C., Grogan, T. M., Isaacson, P. G., Knowles, D. M., Mason, D. Y., Muller-Hermelink, H.-K., Pileri, S. A., Piris, M. A., Ralfkiaer, E. and Warnke, R. A. (1994) A revised European-American classification of lymphoid neoplasms: a proposal from the international lymphoma study group. *Blood,* 84(5), 1361–1392.

Harris, S. E., Sawhill, B. K., Wuensche, A. and Kauffman, S. (2002) A model of transcriptional regulatory networks based on biases in the observed regulation rules. *Complexity,* 7(4), 23–40.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2001) Maximum-likelihood estimation of optimal scaling factors for expression array normalization. In *Proc. SPIE Microarrays: Optical Technologies and Informatics*, (Bittner, M. L., Chen, Y., Dorsel, A. N. and Dougherty, E. R., eds), vol. 4266, 132–140, San Jose, California.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K. and Green, P. J. (2005) BGX: a fully Bayesian integrated approach to the analysis of affymetrix genechip data. *Bioinformatics,* 6(3), 349–373.

Heller, M. J. (2002) DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering,* 4, 129–153.

Hood, L. and Galas, D. (2003) The digital code of DNA. *Nature,* 421(6921), 444–448.

Huang, X. and Pan, W. (2002) Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Functional & Integrative Genomics,* 2(3), 126–133.

Huffman, D. A. (1952) A method for the construction of minimum-redundancy codes. *Proceedings of the Institute of Radio Engineers,* 40, 1098–1102.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H. and Linsley, P. S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology,* 19(4), 342–347.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics,* 2, 343–372.

Ideker, T., Thorsson, V., Siegel, A. F. and Hood, L. E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology,* 7(6), 805–817.

Jaeger, J., Sengupta, R. and Ruzzo, W. L. (2003) Improved gene selection for classification of microarrays. In *Proc. Pacific Symposium on Biocomputing*, 53–64, Kauai, Hawaii.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A.-L. (2000) The large-scale organization of metabolic networks. *Nature,* 407(6804), 651–654.

Johnson, R. A. and Wichern, D. (1998) *Applied Multivariate Statistical Analysis.* 4th edition, Prentice Hall, Upper Saddle River, New Jersey.

Jörnsten, R. (2001). *Data Compression and Its Statistical Implications with an Application to the Analysis of Microarray Images.* PhD thesis, University of California Berkley.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* 28(1), 27–30.

Karnaugh, M. (1953) The map method for synthesis of combinational logic circuits. *Transactions of American Institute of Electrical Engineers,* 72(9), 593–599.

Kauffman, S. A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology,* 22, 437–467.

Kauffman, S. A. (1993) *The Origins of Order: Self-organization and selection in evolution.* Oxford University Press, New York.

Kauffman, S. A. (1995) *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity.* Oxford University Press, New York.

Kauffman, S. A. (2000) *Investigations.* Oxford University Press, New York.

Kauffman, S. A. (2004) The ensemble approach to understand genetic regulatory networks. *Physica A,* 340(4), 733–740.

Kesseli, J., Rämö, P. and Yli-Harja, O. (2005) Tracking perturbations in Boolean networks with spectral methods. *Physical Review E,* 72(2), 026137.

Kesseli, J., Rämö, P. and Yli-Harja, O. (2006) Iterated maps for annealed Boolean networks. *Physical Review E,* 74(4), 046104.

Kitano, H. (2002) Systems biology: a brief overview. *Science,* 295(5560), 1662–1664.

Kitano, H. and Oda, K. (2006) Robustness trade-offs and hostmicrobial symbiosis in the immune system. *Molecular Systems Biology,* 2(2006.0022).

Knezevic, V., Leethanakul, C., Bichsel, V. E., Worth, J. M., Prabhu, V. V., Gutkind, J. S., Liotta, L. A., Munson, P. J., III, E. F. P. and Krizman, D. B. (2001) Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics,* 1(10), 1271–1278.

Kocsor, A., Kertész-Farkas, A., Kaján, L. and Pongor, S. (2005) Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics,* 22(4), 407–412.

Kohonen, T. (2001) *Self-Organizing Maps.* 3rd edition, Springer, New York.

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information. *Problems in Information Transmission,* 1(1), 1–7.

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., ans M. J. Mihatsch, J. T., Sauter, G. and Kallioniemi, O. (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine,* 4(7), 844–847.

Krasnogor, N. and Pelta, D. (2004) Measuring the similarity of protein structures by means of the universal similarity metric. *Bioinformatics,* 20(7), 1015–1021.

Krishnapuram, B., Hartemink, A. J., Carin, L. and Figueiredo, M. A. (2004) A bayesian approach to joint feature selection and classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26(9), 1105–1111.

Kruskal, J. B. (1964*a*) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* 29, 1–27.

Kruskal, J. B. (1964*b*) Nonmetric multidimensional scaling: a numerical method. *Psychometrika,* 29, 115–130.

Lähdesmäki, H., Aho, T., Huttunen, H., Linne, M.-L., Niemi, J., Kesseli, J., Pearson, R. and Yli-Harja, O. (2003) Estimation and inversion of the effects of cell population asynchrony in gene expression time-series. *Signal Processing,* 83(4), 835–858.

Lähdesmäki, H., Shmulevich, I., Dunmire, V., Yli-Harja, O. and Zhang, W. (2005) *in silico* microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics,* 6(54).

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young, R. A. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science,* 298(5594), 799–804.

Lehmussola, A., Ruusuvuori, P. and Yli-Harja, O. (2006) Evaluating the performance of microarray segmentation algorithms. *Bioinformatics,* to appear.

Leung, Y. F. and Cavalieri, D. (2003) Fundamentals of cDNA microarray data analysis. *Trends in Genetics,* 19(11), 649–659.

Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature,* 424(6945), 147–151.

Li, M., Chen, X., Li, X., Ma, B. and Vitanyi, P. (2004) The similarity metric. *IEEE Transactions on Information Theory,* 50(12), 3250–3264.

Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications*. 2nd edition, Springer-Verlag, New York.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nature Genetics,* 21(Suppl 1), 20–24.

Lockshin, R. A. and Zakeri, Z. (2001) Programmed cell death and apoptosis: origins of the theory. *Nature Reviews Molecular Cell Biology,* 2(7), 545–550.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D. and Darnell, J. E. (2001) *Molecular Cell Biology*. Freeman, New York.

Luque, B. and Sole, R. V. (1997) Phase transitions in random networks: simple analytic determination of critical points. *Physical Review E,* 55(1), 257–260.

Luque, B. and Sole, R. V. (2000) Lyapunov exponents in random Boolean networks. *Physica A,* 284(1-4), 33–45.

Ma, H.-W. and Zeng, A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics,* 19(2), 270–277.

Mendes, P. (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Computer Applications in the Biosciences,* 9(5), 563–571.

Mendes, P., Sha, W. and Ye, K. (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics,* 19(Suppl 2), ii122–ii129.

Nykter, M., Aho, T., Kesseli, J. and Yli-Harja, O. (2003) On estimation of statistical characteristics of microarray data. In *Proc. Finnish Signal Processing symposium*, Tampere, Finland.

Nykter, M., Price, N. D., Larjo, A., Aho, T., Aldana, M., Ramsey, S., Kauffman, S. A., Hood, L., Yli-Harja, O. and Shmulevich, I. (2006) Information flow in complex networks and evolution: a universal approach. *Submitted.*

Ohnishi, Y., Tanaka, T., Ozaki, K., Yamada, R., Suzuki, H. and Nakamura, Y. (2001) A high-throughput SNP typing system for genome-wide association studies. *Journal of Human Genetics,* 46(8), 471–477.

Paweletz, C. P., Charboneau, L., Bichsel, V. E., Simone, N. L., Chen, T., Gillespie, J. W., Emmert-Buck, M. R., Roth, M. J., III, E. F. P. and Liotta, L. A. (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene,* 20(26), 1981–1989.

Pettinen, A., Aho, T., Smolander, O.-P., Manninen, T., Saarinen, A., Taattola, K.-L., Yli-Harja, O. and Linne, M.-L. (2005) Simulation tools for biochemical networks: evaluation of performance and usability. *Bioinformatics,* 21(3), 357–363.

Pitkänen, J.-P., Törmä, A., Alff, S., Huopaniemi, L., Mattila, P. and Renkonen, R. (2004) Excess mannose limits the growth of phosphomannose isomerase PMI40 deletion strain of *Saccharomyces cerevisiae*. *Journal of Biological Chemistry,* 279(53), 55737–55743.

Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A.-L. and Szathmáry, E. (2001) Comparable system-level organization of archaea and eukaryotes. *Nature Genetics,* 29(1), 54–56.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics,* 22(9), 1122–1129.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nature Reviews Genetics,* 2(6), 418–427.

Quackenbush, J. (2002) Microarray data normalization and transformation. *Nature Genetics,* 32(Suppl), 496–501.

Rämö, P., Kesseli, J. and Yli-Harja, O. (2006) Perturbation avalanches and criticality in gene regulatory networks. *Journal of Theoretical Biology,* 242(1), 164–170.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000) Genome-wide location and function of DNA binding proteins. *Science,* 290(5500), 2306–2309.

Rissanen, J. and Langdon, G. G. (1979) Arithmetic coding. *IBM Journal of Research and Development,* 23, 149–162.

Rocke, D. M. and Durbin, B. (2001) A model for measurement error for gene expression array. *Journal of Computational Biology,* 8(6), 557–569.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Peñaloza-Spínola, M. I., Martínez-Antonio, A., Karp, P. D. and Collado-Vides, J. (2006) The comprehensive updated regulatory network of *Escherichia coli* K-12. *BMC Bioinformatics,* 7(5).

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* 270(5235), 467–470.

Science Primer (2006). http://www.ncbi.nlm.nih.gov/About/primer/. Retrieved Oct 26.

Serra, R., Villani, M. and Semeria, A. (2004) Genetic network models and statistical properties of gene expression data in knock-out experiments. *Journal of Theoretical Biology,* 227(1), 149–157.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal,* 27, 379–423.

Shedden, K. and Cooper, S. (2002*a*) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proceedings of the National Academy of Sciences USA,* 99(7), 4379–4384.

Shedden, K. and Cooper, S. (2002*b*) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Research,* 30(13), 2920–2929.

Shmulevich, I., Gluhovsky, I., Hashimoto, R. F., Dougherty, E. R., and Zhang, W. (2003) Steady-state analysis of genetic regulatory networks modeled by probabilistic Boolean networks. *Comparative and Functional Genomics,* 4(6), 601–608.

Shmulevich, I. and Kauffman, S. A. (2004) Activities and sensitivities in Boolean network models. *Physical Review Letters,* 93(4), 048701.

Shmulevich, I., Kauffman, S. A. and Aldana, M. (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proceedings of the National Academy of Sciences USA,* 102(38), 13439–13444.

Shmulevich, I., Lähdesmäki, H., Dougherty, E. R., Astola, J. and Zhang, W. (2003) The role of certain Post classes in Boolean network models of genetic networks. *Proceedings of the National Academy of Sciences USA,* 100(19), 10734–10739.

Shmulevich, I. and Zhang, W. (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics,* 18(4), 555–565.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics,* 29(3), 263–264.

Solomonoff, R. (1964) A formal theory of inductive inference. *Information and Control,* 7, 1–22.

Sonenberg, N., Hershey, J. W. B. and Mathews, M., eds (2000) *Translational Control of Gene Expression.* 2nd edition, Cold Spring Harbor Laboratory Press, New York.

Speed, T. (2003) *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC, New York.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell,* 9(12), 3273–3297.

Stauffer, D. (1987) On forcing functions in Kauffman random Boolean networks. *Journal of Statistical Physics,* 46(3–4), 789–794.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences USA,* 102(43), 15545–15550.

Tabus, I., Rissanen, J. and Astola, J. (2003) Classification and feature gene selection using the normalized maximum likelihood model for discrete regression. *Signal Processing, Special issue on Genomic Signal Processing,* 83(4), 713–727.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA,* 96(6), 2907–2912.

Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences USA,* 101(9), 2981–2986.

The Genome International Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature,* 409(6822), 860–921.

The Genome International Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature,* 431(7011), 931–945.

Trotter, M. J. and Bruecks, A. K. (2003) Interpretation of skin biopsies by general pathologists: diagnostic discrepancy rate measured by blinded review. *Archives of Pathology and Laboratory Medicine,* 127(11), 1489–1492.

Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C. and Wong, W. H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research,* 29(12), 2549–2557.

Tu, Y., Stolovitzky, G. and Klein, U. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences USA,* 99(22), 14031–14036.

Tuimala, J. and Laine, M. M., eds (2005) *DNA Microarray Data Analysis.* 2nd edition, CSC – Scientific Computing Ltd, Helsinki.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA,* 98(9), 5116–5121.

Vaux, D. L., Cory, S. and Adams, J. M. (1988) *Bcl-2* gene promotes haemopoietic cell survival and cooperates with *c-myc* to immortalize pre-B cells. *Nature,* 335(6189), 440–442.

Venter, J. C., Adams, M. D., Myers, E. W., W., P., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A. *et al.* (2001) The sequence of the human genome. *Science,* 291(5507), 1304–1351.

Watson, J. D. and Crick, F. H. (1953) Molecular structure of nucleic acids. *Nature,* 171(4356), 737–738.

Watts, D. J. and Strogatz, S. H. (1998) Collective dynamics of 'small-world' networks. *Nature,* 393(6684), 440–442.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B. and Bassett, D. E. (2006) Rosetta error model for gene expression analysis. *Bioinformatics,* 22(9), 1111–1121.

Wuensche, A. (1999) Discrete dynamical networks and their attractor basins. *Complexity International,* 6, 2–23.

Xavier, A. C. G., Siqueira, S. A. C., Costa, L. J. M., Mauad, T. and Saldiva, P. H. N. (2005) Missed diagnosis in hematological patients – an autopsy study. *Virchows Archiv,* 446(3), 225–231.

Yang, Y. H., Buckley, M., Dudoit, S. and Speed, T. (2002*a*) Comparison of methods for image analysis on cDNA microarray data. *Journal of Computional and Graphical Statistics,* 11(1), 108–136.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002*b*) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* 30(4), e15.

Yockey, H. P. (2005) *Information Theory, Evolution, and The Origin of Life.* Cambridge University Press, New York.

Zhang, W., Shmulevich, I. and Astola, J. (2004) *Microarray Quality Control.* John Wiley and Sons, Hoboken, New Jersey.

Zhu, D. and Qin, Z. S. (2005) Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics,* 6(8).

Ziv, J. and Lempel, A. (1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory,* 23(3), 337–343.

Publications

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O. Box 527
FIN-33101 Tampere, Finland