Statistical and Functional Analysis of Genomic and Proteomic Data

Liu, Yingchun

2007

Statistical and Functional Analysis of Genomic and Proteomic Data

Yingchun Liu

Department of Theoretical Physics
Lund University, Sweden

Dissertation for the degree of Doctor of Philosophy

Advisor: Markus Ringnér

Faculty Opponent: Sayan Mukherjee
Duke University, USA

To be presented, with the permission of the Faculty of Natural Sciences of Lund University, for public criticism in Lecture Hall F of the Department of Physics on Friday, the 26th of January 2007, at 10.15 a.m.

DOKUMENTDATABLAD enl SIS 61 41 21

Organization
LUND UNIVERSITY
Department of Theoretical Physics
Sölvegatan 14A
SE-223 62 LUND

Document Name
DOCTORAL DISSERTATION

Date of issue
December 2006

CODEN:

Author(s)
Yingchun Liu

Sponsoring organization

Title and subtitle

Statistical and functional analysis of genomic and proteomic data

Abstract

High-throughput technologies have led to an explosion in the availability of data at the genome scale. Such data provide important information about cellular processes and causes of human diseases, as well as for drug discovery. Deciphering the biologically relevant results from these data requires comprehensive analytical methods. In this dissertation, we present methods for gene and protein expression data analysis. Our major contributions include a method for differential in-gel electrophoresis data analysis capable of removing protein-specific dye bias in the data, a method for finding unknown biological groups using expression data, and a method for identifying active and inactive signaling pathways in a gene expression signature based on the enrichment of downstream target genes of pathways.

Key words
2D-gel, dye bias, expression data, linear mixed model, microarray, regulatory motif, signaling pathway, TGF-beta, unsupervised classification

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title

ISBN
91-628-6997-3

Recipient's notes

Number of pages 80

Price

Security classification

Distribution by (name and address)
Yingchun Liu, Dept. of Theoretical Physics,
Sölvegatan 14 A, SE-223 62 LUND

Signature _____ **Date** 2006-12-05

Information is only as good
as your ability to make use of it

Acknowledgments

I am grateful to a large number of people who have inspired and helped me throughout the years of my studies.

First and foremost I would like to thank my advisor Markus Ringnér for his guidance and support throughout the past four years. I have benefited greatly from his brilliance and excellent combination of knowledges. I have learned so much from him, including how to think analytically, how to do research, and how to write and present research. Except for being a great advisor, Markus also has a fantastic personality. His sense of humor and encouragement for me make it a great pleasure to work with him. Markus, it has been a wonderful and enriching experience to work closely and learn from you.

I would also like to especially thank Morten Krogh for his significant influence on my mathematical and scientific thinking. I am indebted to him for the skills I learned from him and for his insightful comments. Morten is also full of good ideas. I was often inspired by discussing with him and I really enjoyed working with him.

I am very grateful to Carsten Peterson. His concerns and support have been a big encourage to me. Carsten has great insight into problems and discusses ideas openly. I always feel enlightened after talking to him.

All my co-authors have tremendous contributions to the work in this dissertation. In particular, Stefan Karlsson and Göran Karlsson have brought me ideas in biology and have led me to the field of signaling pathways. I am thankful to them for this.

Other wonderful folks in the department include Anders Irbäck, who

always greets me with a warm smile and has been extremely patient in explaining difficult concepts with which I was unfamiliar; Michael Green, who has been a great colleague and friend. Whenever I turn to him for help, he always welcomes me with a big smile. I have learned many computer skills from him; Peter Johansson, who helped me a lot with physics and Swedish; Jari Häkkinen, who assisted in my projects; Liwen You and Carl Troein, who brought motivating discussions; Patrik Edén, who helped me with the pathway project; Henrik Jönsson, who brought helpful ideas; Mattias Ohlsson and Leif Lönnblad, who helped me with computer problems. I feel very grateful to all of you.

I would also like to thank Anders Blomberg for organizing fantastic scientific activities in the research school, and thank Olle Nerman and Ziad Taib for writing a letter of recommendation on my behalf.

Many friends have brought me a lot of joys and made my life colorful over the past years. Especially, I would like to thank Paul for always being my best friend. Aaron, who has inspired me and shown me what it means to be excellent.

Finally, I want to give my deep thanks to my parents, sister, and brother. I am grateful for their love and support for me.

This dissertation is based on the following papers:

I M. Krogh, Y. Liu, S. Bengtsson, B. Valastro and P. James
Analysis of DIGE data with a linear mixed model incorporating protein-specific dye effects
LU TP 06-40 (submitted)

II Y. Liu and M. Ringnér
Multiclass discovery in array data
BMC Bioinformatics **5**:70 (2004)

III G. Karlsson, Y. Liu, J. Larsson, M-J. Goumans, J-S. Lee, S.S. Thorgeirsson, M. Ringnér and S. Karlsson
Gene expression profiling demonstrates that TGF-β1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-β signaling
Physiological Genomics **21**, 396-403 (2005)

IV Y. Liu and M. Ringnér
Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis
LU TP 06-36 (submitted)

Contents

Biological Background

We begin with a brief overview of the concepts of molecular biology that are relevant to this dissertation followed by an introduction to technologies for generating genome-wide data. For further knowledge of molecular biology, the reader can refer to the book [1].

Basic Concepts of Molecular Biology

DNA, gene, mRNA, and protein

Cells are basic units of life. All living organisms are built from cells. At the center of the cell, there is the cell nucleus which contains the genetic code, DNA (Deoxyribonucleic acid), of the cell. The DNA molecule consists of two long sequences of nucleotides, where each nucleotide is composed of one sugar molecule, one phosphate molecule, and one of the four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases can form complementary base pairs in the form of A-T and C-G, joined by hydrogen bonds. The two sequences of the DNA are joined by such base pairs and twisted into a double helix (Figure 1). DNA is organized into separate chromosomes in the cell nucleus, and the whole genetic information encoded in the DNA for an organism is termed genome.

A *gene* is a region on the DNA sequence that codes for proteins. In the human genome, there are over 30,000 genes, and there are even more proteins than genes, because each gene can code for multiple proteins. Genes encode proteins through two main steps. Firstly, the DNA sequence of a gene is transcribed into another molecule called mRNA

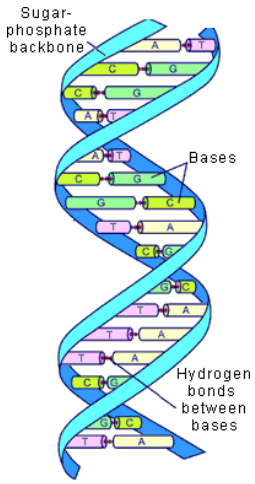

Figure 1: An illustration of the DNA molecule. Image taken from http://web.jjay.cuny.edu/~acarpi/NSC/12-dna.htm

(Ribonucleic acid) by base pairing, thereby the mRNA contains a nucleotide sequence complementary to its template DNA sequence. This process is termed *transcription*. When a gene is transcribed into mRNA, the gene is said to be expressed. The *expression level* of this gene refers to its mRNA abundance. Secondly, the mRNA leaves the cell nucleus and travels to the cellular processing units called ribosomes, where it serves as template for protein synthesis. Every three consecutive nucleotides of the mRNA sequence are converted into one amino acid, and the amino acids are linked together by peptide bonds into a poly-peptide chain. This process is termed *translation*. Finally, the chain folds into a protein with specific three dimensional structure (see Figure 2).

Gene expression regulation

The are often many types of cells in an organism. In the human body, there are brain cells, lung cells, liver cells, skin cells, and so on. Although all the cells contain the same DNA, they appear different and have different functions. This is because different genes are expressed in

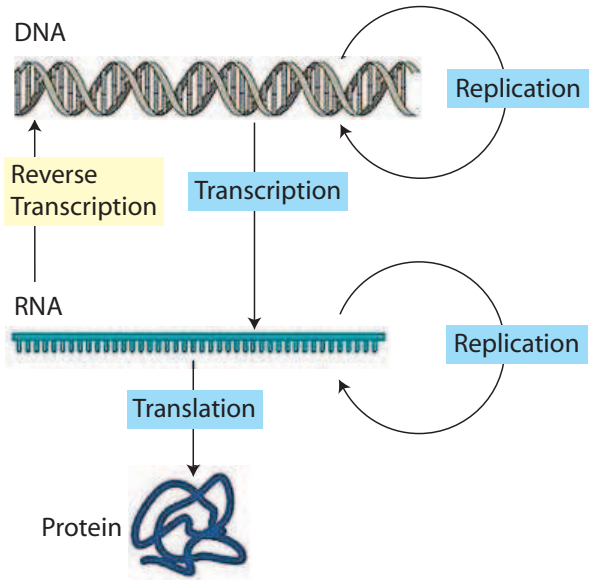

Figure 2: The central dogma of molecular biology. Genetic information encoded in the DNA is passed to RNA through transcription and then to proteins through translation.

different types of cells, which leads to production of different proteins.

Gene expression is largely controlled by regulatory proteins called *transcription factors*. To control gene expression, transcription factors bind to specific short sequences on the DNA in the region upstream of the transcription start site of the gene. This binding recruits or impedes proteins necessary for transcription to enhance (*up-regulate*) or inhibit (*down-regulate*) the transcription of this gene. Such an upstream region of a gene is called a *promoter* region.

In the cell, one gene can be regulated by many transcription factors. One transcription factor can bind to many short sequences in the promoter regions of different genes to regulate their transcription. This forms a complex regulatory network. Similar binding sequences of a transcription factor are represented by a common pattern called the *motif* of the transcription factor. One transcription factor could have multiple motifs.

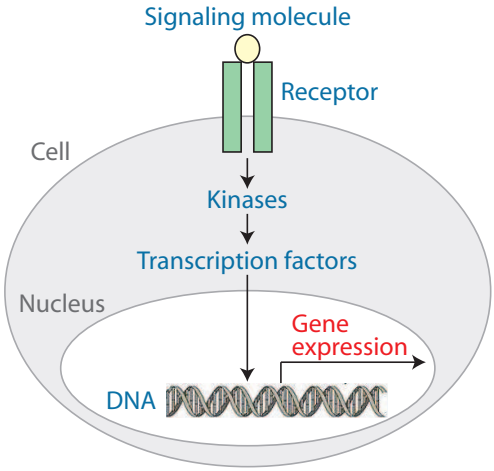

Figure 3: A schematic representation of the signal transduction within a eukaryotic cell. Once the signaling molecule binds to the receptor, the signal is passed through a number of steps into the cell nucleus, where gene expression is affected.

Signaling and metabolic pathways

Cells have the ability to communicate with their internal and external environment, and adjust their functions in response to environmental changes. This ability is achieved through a number of signaling pathways that receive and process signals.

A signaling pathway consists of a set of molecules, such as ligands, receptors, enzymes, and transcription factors. Ligands are signaling molecules from the external or internal environment, and receptors are proteins located on the cell membrane or within the cell that bind to signaling molecules. Once ligands bind to receptors, the signal is propagated within the cell through a cascade of biochemical interactions between receptors and transcription factors (Figure 3). As a result, various transcription factors are activated or inactivated, which in turn alters the expression levels of many genes and eventually alters the activities of biological processes.

Signal transduction is at the core of many biological processes. For example, cell growth, proliferation, differentiation, and apoptosis are

all controlled by signaling pathways. The cell can respond to multiple signals at a time, and the cell's response to one specific signal could activate multiple signaling pathways. Various signaling pathways within the cell form a signaling network.

In addition to signaling pathways, metabolic pathways are also important for the cell's survival. A metabolic pathway is a series of enzyme-catalyzed biochemical reactions that produce energy storage molecules and biomolecules for the cell. In a metabolic pathway, the product of one reaction is the substrate of the next reaction. Metabolic pathways can share common substrates and enzymes forming a metabolic network.

Gene and Protein Expression Profiling

Most processes within the cell are carried out by proteins and their interactions with other molecules. For example, proteins function as enzymes that catalyze biochemical reactions, receptors that receive and propagate signals, and transcription factors that regulate gene expression. Each process is governed by a specific set of active proteins, and the activities of proteins thereby provide important information about the ongoing processes in the cell. In many cases, protein activities are correlated with their abundances. Since proteins are produced from the mRNAs of genes, protein abundances are often correlated with the mRNA abundances of the genes encoding the proteins. Therefore, gene expression studies have the potential to reveal the active processes in the cell.

Cells affected by diseases often have a set of genes differentially expressed with respect to normal cells, because different cellular processes typically are activated as a response to genetic or cellular changes. Such differentially expressed genes may provide insight into the causes of the diseases or be potential drug targets. However, there are hundreds to thousands of genes in an living organism. It is impossible to know which genes to be examined without detailed prior knowledge. This had been a bottleneck for biomedical research using traditional biotechnologies, which can only measure the expression levels of one or few genes at a time. Fortunately, the invention of DNA microarray technology about a decade ago has made genome-wide gene expression studies possible [2].

Microarray-based gene expression profiling

DNA microarrays can measure expression levels of thousands of genes simultaneously on a single slide. Each slide contains thousands of spatially separated spots on the surface. And each spot contains multiple copies of a short DNA sequence that represents one gene. To measure gene expression levels, the mRNA contents of cells are extracted from samples and reversely transcribed into complimentary DNAs (cDNA). The cDNAs are labeled with a fluorescent dye that absorbs and emits light at specific wavelengths. Then, the cDNAs are hybridized on the array where they bind to their complementary DNA sequences in the spots. Finally, the array is scanned to obtain the emitted fluorescent intensities for all spots. The intensity of each spot indicates the expression level of the gene it represents.

There are many types of DNA microarrays, which differ in array fabrication or choice of dyes [3]. In a two-color DNA microarray [4], which is relevant for this dissertation, the mRNA abundances of genes in two samples are compared directly on the same array by labeling them with different fluorescent dyes. After hybridization, the array is scanned at two different wavelengths, which generates two intensities for each spot corresponding to the expression level of this gene in the two samples. I will refer to the two intensities as red and green in the later context. A schematic overview of cDNA microarray technology is shown in Figure 4.

The mRNA abundance is, however, not always correlated with protein abundance. For example, the rates of mRNA decay, translation, and protein decay can influence this correlation. In addition, protein activity is not always correlated with its abundance either. Proteins could be activated or inactivated by post-translational modifications. Hence, large-scale data at protein level, termed *proteomic data*, are useful. Technologies like protein microarray [6], two dimensional poly-acrylamide gel electrophoresis (2D-PAGE) [7,8], and mass spectrometry [9] are used to generate data for proteomic research. These technologies are still immature and under development though, due to complex features of proteins. So far, the most widely applied approach for proteomic research has been protein expression analysis.

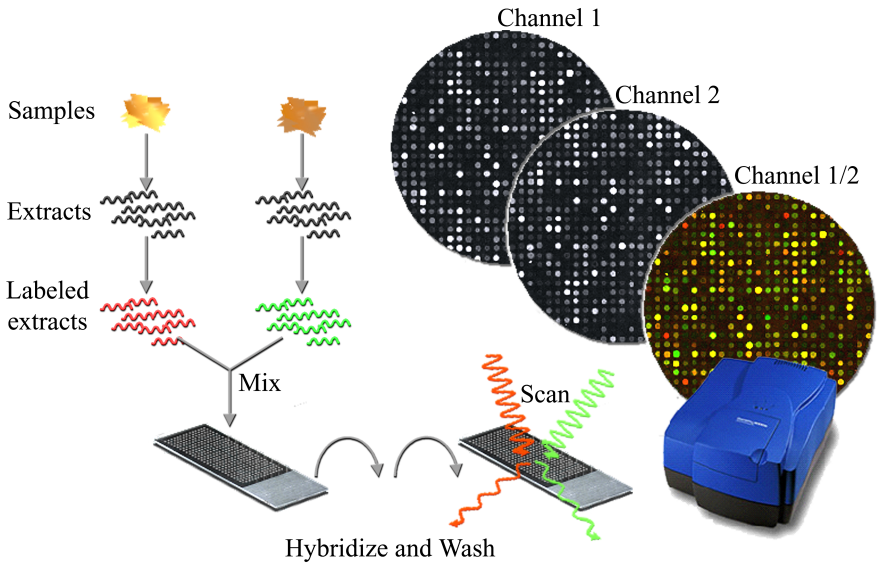

Figure 4: A schematic overview of the two-color DNA microarray technology. Two samples are labeled with a red and a green fluorescent dye respectively and are hybridized on the same array. After hybridization, the array is scanned to obtain two images. The two images are often merged into one image where spots are colored on a scale from red to yellow to green corresponding to the relative gene expression in the two samples. Reprinted with permission from Johan Vallon-Christersson [5].

2D gel-based protein expression profiling

One common way of measuring the abundances of thousands of proteins simultaneously is by means of 2D-PAGE. In 2D gels, proteins extracted from a sample are first separated by isoelectric point using an immobilized pH gradient. Next, proteins are separated by molecular weight, because proteins with different weights move at different speed on the gel. The resulting gel is then stained to visualize the protein spots. Finally, the gel is scanned to obtain the intensities of all spots.

Traditional 2D-PAGE can only deal with one sample on each gel. More recently, differential in-gel electrophoresis (DIGE) [10] was introduced,

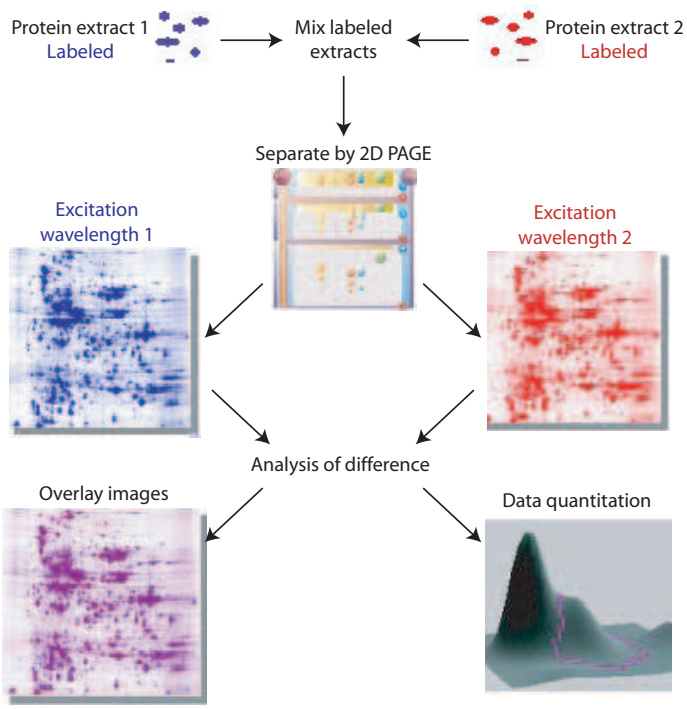

Figure 5: Outline of a DIGE experiment to compare protein abundances in two samples. Protein extractions of two samples are labeled with two different dyes and are resolved on the same gel. Finally, two scanned images are obtained, and the intensities of the spots indicate the abundances of proteins.

which can measure protein abundances of up to three samples on the same gel. In DIGE gels, protein extractions from samples are labeled with different fluorescent dyes and are mixed prior to 2D gel electrophoresis. Finally, the abundances of proteins in these samples are determined by the scanned intensities for the spots. An overview of the DIGE technology is shown in Figure 5.

Genomic and Proteomic Data Analysis

Recent advances in technologies have made a vast amount of data at the genome scale available, including complete genome sequences, genome-wide protein-DNA binding sites, and genome-wide gene and protein expression profiles under various conditions. Such data provide important genetic and cellular information. However, transforming these immense amounts of data into biological information is challenging, especially when there are measurement uncertainties in the data. A successful transformation relies on theoretically-founded methods with deep understanding of the biology. In this chapter, I will discuss some of the challenges: normalization of expression data, identification of differential expression, unsupervised classification, and investigation of functional networks. These are addressed in the appended papers.

Normalization of Expression Data

Expression data obtained by using microarrays are noisy. Differences in the RNA abundances between samples are often mixed with nonbiological variations. Dye bias is the most common nonbiological variation, caused by different labeling or scanning properties of dyes. In particular, the RNA may bind to one dye better than the other, or the same RNA sample labeled with different dyes could have different measured intensities. Except for the dye bias, differences between arrays would exist when the hybridization efficiency on each array is different. Differences between spots would exist when there is different amount of cDNAs printed on each array for the same gene. And different print-tips for different locations on the array would introduce variation as

well. Before applying microarray data for biological studies, the data must be normalized to remove nonbiological variations arising from the technology.

There are many statistical approaches to normalizing two-color DNA microarray data. For instance, normalization can be done separately for each array, using only the red (R) and green (G) intensities for this array, or it can be done using multiple arrays. Overall, many of these approaches aim to have all normalized $\log_2(R/G)$ ratios on each array centered around zero. The underlying assumption is that the numbers of up- or down-regulated genes in each sample are roughly the same, when a random set of genes are printed on each array. Consequently, the mean $\log_2(R/G)$ for each array should be close to zero.

Global normalization is the simplest and most widely used approach, where all the green intensities are multiplied with a constant factor such that the red and green intensities have equal mean or median. This can be done using all the genes on each array, or a selected set of genes, e.g. housekeeping genes [11] or externally spiked genes [12] whose expression levels are constant across multiple conditions. However, dye effects are often dependent on spot intensity and location on the array, so intensity dependent and print-tip based local normalization methods seem more appropriate in this regard [13, 14].

Each of the approaches above is likely to remove only certain nonbiological variations in the data. A more general approach for normalizing DNA microarray data is to use the analysis of variance (ANOVA) models, including fixed-effects ANOVA and mixed-model ANOVA [15–17]. The ANOVA models can be designed to account for variations arising from many sources including arrays, dyes, spots, and their confounding effects, by considering each of them as an unknown parameter of the model. The normalized expression levels of each gene can be obtained by fitting the model using data from all arrays.

In paper I, we introduced a linear mixed model that is able to correct for protein-specific dye effects in DIGE data. DIGE data for protein expression studies have similar properties as microarray data and must be normalized as well.

Identification of Differential Expression

A common task of microarray and DIGE data analysis is to find the genes or proteins differentially expressed between biological groups, for example, disease versus healthy, different cell types, and different conditions. Genes or proteins with expression patterns associated with these groups could provide insight into the causes of diseases, be molecular markers differentiating between cell types, and reflect the active cellular processes under different conditions.

The simplest way of finding differentially expressed genes is the fold change, which considers all genes, whose log ratios between two groups are larger than an arbitrary threshold as differentially expressed [18, 19]. The statistical significance of a gene being differentially expressed is, however, unknown for the fold change. More commonly, differentially expressed genes are identified by performing a statistical test gene-by-gene.

For comparison between two groups, the most widely used tests are t-test [20, 21], modified t-tests (significance analysis of microarrays [22]; the regularized t-test [23]), paired t-test [24], Pearson correlation [25], Wilcoxon rank-sum test [26], and permutation test. Each of these tests emphasizes different aspects of the data. The difference between the t-test and its modifications lies in the calculation of the variance of each gene, so that a gene with too small fold change but small variance by chance will not be selected, or vice versa. In the paired t-test, array and spot effects are taken into account, where, for each gene printed on the array, the expression levels from two groups are compared directly and used as a pair in the t-test. The Pearson correlation measures the association between the expression of a gene with the group label, rather than tests the difference in the mean expression of two groups. Compared with all types of t-tests and Pearson correlation, the Wilcoxon rank-sum test and permutation tests do not require normal distribution of data. As microarray data are noisy and often do not form a normal distribution, the nonparametric tests are appealing in this context. The permutation test is also very flexible. It can be used to test the significance of a score constructed in any way, including the scores used in the different tests.

To identify genes differentially expressed between multiple groups, the ANOVA F test [27] and Kruskal-Wallis test [28] are widely used. In addition, the likelihood ratio test and Wald test are also common, when using models for expression analysis. These two tests can be applied to comparison between any number of biological groups.

Each of these tests has its merits. None of them is superior to others for all microarray data sets. The particular test used depends on the data set under study. Statistical methods for identifying differentially expressed proteins are similar.

In paper I, we applied the likelihood ratio test and the Wald test to identify differentially expressed proteins. We employed the Wilcoxon rank-sum test and the Kruskal-Wallis test, in paper II, to find differentially expressed genes.

Unsupervised Classification

Unsupervised classification refers to revealing unknown biological classes in a collection of samples. In medicine, an important usage of unsupervised classification is to find new subtypes of cancers. Cancer is a complex genetic disease having many subtypes. Classification of cancer is primarily based on the histopathological appearance of the tumor. However, tumors with similar histologic appearance could have developed from different genetic aberrations and have different responses to therapy. For example, different subtypes of breast tumors have different responses to chemotherapy. Cancer treatment based on conventional diagnosis is thus difficult. A major challenge of cancer treatment has been to find specific therapies to pathogenetically distinct tumor types.

Interestingly, recent studies have found that some morphologically similar tumors can be molecularly divided into subclasses with distinct pathogenese. For example, microarray-based gene expression studies have identified subtypes of cutaneous melanoma [29] and four subtypes of breast tumors [30]. These and similar findings have triggered the enthusiasm for unsupervised classification of samples using gene expression data. Many methods have been developed for this purpose.

The most widely used method for unsupervised classification is perhaps agglomerative hierarchical clustering [31, 32]. This method begins by considering each sample as a cluster, and then merges the two closest clusters based on a similarity metric. This process of merging is repeated until there is a single cluster left. Finally, the samples are organized into a tree structure. Classes can be obtained by cutting the tree at a particular height. The k-means method is another commonly used method [33]. Starting from k randomly or carefully chosen data points, called 'centroids', the k-means method iteratively assigns samples to the nearest centroid's cluster and adjusts the centroids to represent the center of the new clusters, optimizing some objective function. Eventually, samples are divided into k clusters. In addition to these two methods, there are a few model-based methods that assume data are sampled from a model distribution, for example a mixture of Gaussian distributions, and seek for parameters that best fit the data [34].

In paper II, we took a strategy different from those described above by using information about differentially expressed genes. Samples from known different classes usually have an overabundance of genes differentially expressed, compared with random classes. These differentially expressed genes are often up-regulated in one class and down-regulated in the other, which forms nice expression patterns characterizing the distinction between compared biological groups. These expression patterns are intrinsic features of the data. Even if we did not know the class labels of the samples, there would still exist such expression patterns. Therefore, in unsupervised classification, we are likely to find the biologically relevant classes in the data, if we could find a partition that exhibits such expression patterns. We applied simulated annealing to find the best partitions.

Investigation of Functional Networks

Finding genes or proteins differentially expressed between biological groups and identifying unknown groups using expression profiles are approaches capable of characterizing biological groups and diagnosing diseases. Furthermore, mapping the differentially expressed genes and proteins to

biological categories, including chromosome location and biological processes, help account for the observed differences between biological groups.

It is important to realize that, in living cells, many genes and proteins function coordinately for complex functions. It is crucial to understand the relationships between them. This understanding has significant impact on drug discovery, because complex diseases often depend on altered interactions between a few genes, rather than changes in a single gene. For example, p53 is a tumor suppressor responsible for DNA repair. It inhibits cell growth in response to DNA damage. But p53 function is controlled by the Mdm2 protein interacting with it. Mdm2 enhances degradation of p53 [35]. If a person has lung cancer, and also has an abnormal overabundance of MDM2 protein in his lung cells, this person can not be cured by simply increasing p53 transcription.

Ideally, we would like to learn the relationships of all genes and proteins in an organism, but biological complexity increases exponentially with the number of genes and the interactions between them. Such large-scale studies are only feasible for simple organisms. In yeast, studies have shown that it is possible to infer molecular pathways [36] and regulatory networks [24,37,38] from gene expression data, using probabilistic models, bayesian or boolean networks. For higher organisms, efforts have been focusing on learning the structure and dynamics of small systems, such as the cell cycle [39] and specific signaling pathways [40]. Such small systems can be studied by perturbation (for example, knocking out interesting genes or overexpressing specific proteins) followed by monitoring the response of each element over time. Finally, the relationships of elements of the system and its response to individual perturbations can be described by mathematical models.

Alternatively, unlike the studies above attempting to learn the detailed relationships, some other studies aim to uncover the coordinate behavior of many genes and proteins in terms of known systems, including metabolic pathways [41, 42] and signaling pathways [43]. Since each pathway usually involves many genes and proteins, and different pathways often share common genes, the set of all known pathways form a complex network of genes and proteins. Active and inactive pathways provide insight into the regularities in the observed gene expression pro-

files with respect to the topology of this gene network.

In addition, gene expression data alone might not be enough for learning functional relationships. More recently, two studies have tried to identify regulatory programs in large-scale transcriptional signatures in cancer, by integrating microarray data with regulatory motifs [44] and DNA copy number [45].

In paper IV, we presented a strategy to study the regulatory mechanisms responsible for the observed gene expression profiles in the context of signaling pathways. We integrated pathway information with regulatory motif data for pathway analysis.

Summary of the Papers

Paper I

In Paper I, we study the dye effects in protein expression data generated by DIGE experiments, where abundances of thousands of proteins in three samples are measured simultaneously on one gel. Each of the samples is labeled with a distinct fluorescent dye. Prior to comparison of protein abundances, differences between dye intensities must be removed. This is usually done by a global normalization within each dye channel. However, we find that dye effects are in fact protein-specific and cannot be removed by any global normalization methods. To address this problem, we introduce an algorithm, a linear mixed model, which incorporates protein-specific dye effects and is applicable to most experimental designs. The algorithm is implemented in a JAVA program called DIGEanalyzer that automatically corrects for protein-specific dye effects and identifies differentially expressed proteins between any linear combination of groups. DIGEanalyzer is available at http://bioinfo.thep.lu.se/digeanalyzer.html.

Conclusion: Dye effects in DIGE data are protein-specific which cannot be corrected for by global normalization methods. We present a program that corrects for protein-specific dye effects and identifies differentially expressed proteins.

Paper II

In Paper II, we present a method to find biologically relevant groups in a set of samples using microarray data. Unlike many methods that cluster experiments based on their distances in gene expression space, we look for partitions of samples that have an overabundance of differentially expressed genes, starting from a predefined number of groups and randomly labeled samples. We evaluate this method using two published microarray data sets: small round blue cell tumors (SRBCT) and breast tumors. The SRBCT data set contains samples belonging to four different SRBCT types and the breast tumors data set contains non-BRCA1/2 familial breast tumors. When applying the method on these data sets, interestingly, we find that the SRBCT data set can be separated perfectly into two groups: tumors and cell lines or into three groups reflecting print batches of microarrays. Our method is able to detect such groups and remove genes discriminating them from analysis, which enables us to find the biologically relevant groups in these data with high success rates. This method is available as a PERL program at http://bioinfo.thep.lu.se/classdiscoverer.

Conclusion: Unknown biological groups in a set of samples could be identified by looking for partitions of samples with an overabundance of differentially expressed genes.

Paper III

In paper III, we study the TGF-β signaling system. Transforming growth factor-β1 (TGF-β) regulates cellular functions, such as proliferation, differentiation, and apoptosis through the TGF-β signaling pathway. It is well-known that the TGF-β signal is transduced through receptor complexes composed of TGF-β receptor type II (TβRII) and activin-like kinase receptor-5 (Alk5) on the cell surface. In this study, we screen for alternative receptors for TGF-β in murine embryonic fibroblast (MEF) cells using gene expression profiling and functional assays. We also identify gene targets of TGF-β signaling in MEF cells.

Conclusion: TGF-β signals exclusively through receptor complexes involving Alk5 in MEF cells.

Paper IV

In Paper IV, we present a method to study the regulatory mechanisms underlying diseases and other biological observations in terms of signaling pathways. We look for active and inactive signaling pathways in the gene expression signatures characteristic of these observations. The method takes a gene signature as input and outputs the signaling pathways whose activation or inactivation might have resulted in the observed expression patterns of these genes. In the analysis, all pathways in the TRANSPATH database are extracted and each is characterized by a set of transcription factors mediating it. The activity of each pathway in the gene signature is inferred based on the enrichment of the downstream target genes of the pathway. Since there are few known target genes, we search for putative target genes by looking for the binding motifs of the transcription factors in the promoter regions of genes. This method is different from many methods in two aspects: First, the activities of pathways are determined by the enrichment of target genes of pathways rather than that of molecular components of pathways. Second, putative target genes that contain the motifs of the transcription factors are used, instead of the few known target genes. We evaluate this method using six human and mouse gene expression signatures.

Conclusion: Regulatory motif analysis of gene expression signatures reveals signaling pathway activation or inactivation.

References

[1] Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD: *Molecular biology of the cell*. Garland Publishing 1994.

[2] Schena M, Shalon D, Davis RW, Brown P: **Quantitative monitoring of gene expression patterns with complementary DNA microarray**. *Science* 1995, **270**(5235):467–470.

[3] Heller MJ: **DNA microarray technology: devices, systems, and applications**. *Annu Rev Biomed Eng* 2002, **4**:129–153.

[4] Xiang CC, Chen Y: **cDNA microarray technology and its applications**. *Biotechnol Adv* 2000, **18**:35–46.

[5] Vallon-Christersson J: **Functional and molecular characterization of BRCA1 and BRCA2 associated breast cancer**. *PhD thesis*, Lund University 2005.

[6] Templin MF, Stoll D, Schwenk JM, Potz O, Kramer S, Joos TO: **Protein microarrays: promising tools for proteomic research**. *Proteomics* 2003, **3**(11):2155–2166.

[7] O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins**. *J Biol Chem* 1975, **250**(10):4007–4021.

[8] Scheele GA: **Two-dimensional gel analysis of soluble proteins. Charaterization of guinea pig exocrine pancreatic proteins**. *J Biol Chem* 1975, **250**(14):5375–5385.

[9] Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198–207.

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

[11] Lee PD, Sladek R, Greenwood CMT, Hudson TJ: **Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies**. *Genome Res* 2002, **12**(2):292–297.

[12] van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FCP: **Monitoring global messenger RNA changes in externally controlled microarray experiments**. *EMBO Rep* 2003, **4**(4):387–393.

[13] Quackenbush J: **Microarray data normalization and transformation**. *Nat Genet* 2002, **32 Suppl**:496–501.

[14] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic Acids Res* 2002, **30**(4):e15.

[15] Kerr MK: **Analysis of variance for gene expression microarray data**. *J Comput Biol* 2000, **7**(6):819–837.

[16] Lee MLT, Lu W, Whitmore GA, Beier D: **Models for microarray gene expression data**. *J Biopharm Stat* 2002, **12**:1–19.

[17] Carter MG, Hamatani T, Sharov AA, Carmack CE, Qian Y, Aiba K, Ko NT, Dudekula DB, Brzoska PM, Hwang SS, Ko MSH: **In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling**. *Genome Res* 2003, **13**(5):1011–1021.

[18] DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680–686.

[19] Draghici S: **Statistical intelligence: effective analysis of high-density microarray data**. *Drug Discov Today* 2002, **7**(11 Suppl):55–63.

[20] Salmon KA, Hung Sp, Steffen NR, Krupp R, Baldi P, Hatfield GW, Gunsalus RP: **Global gene expression profiling in Escherichia**

coli K12: effects of oxygen availability and ArcA. *J Biol Chem* 2005, **280**(15):15084–15096.

[21] Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice**. *Genome Res* 2000, **10**(12):2022–2029.

[22] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116–5121.

[23] Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509–519.

[24] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166–176.

[25] Mansson R, Tsapogas P, Akerlund M, Lagergren A, Gisler R, Sigvardsson M: **Pearson correlation analysis of microarray data allows for the identification of genetic targets for early B-cell factor**. *J Biol Chem* 2004, **279**(17):17905–17913.

[26] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**(11):1454–1461.

[27] Li H, Wood CL, Liu Y, Getchell TV, Getchell ML, Stromberg AJ: **Identification of gene expression patterns using planned linear contrasts**. *BMC Bioinformatics* 2006, **7**:245.

[28] Croonquist PA, Linden MA, Zhao F, Van Ness BG: **Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, N-ras-activating mutations, and coculture with bone marrow stromal cells**. *Blood* 2003, **102**(7):2581–2592.

[29] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V: **Molecular classification of cutaneous malignant melanoma by gene expression profiling**. *Nature* 2000, **406**(6795):536–540.

[30] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747–752.

[31] Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863–14868.

[32] Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS: **Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays**. *Cancer Res* 1998, **58**(22):5009–5013.

[33] Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns**. *J Comput Biol* 1999, **6**(3-4):281–297.

[34] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**(10):977–987.

[35] Piette J, Neel H, Marechal V: **Mdm2: keeping p53 under control**. *Oncogene* 1997, **15**(9):1001–1010.

[36] Segal E, Koller D: **Discovering molecular pathways from protein interaction and gene expression data**. *Bioinformatics* 2003, **19**:i264–i272.

[37] Friedman N, Linial M, Nachman I, Pe'er D: **Using bayesian networks to analyze expression data**. *J Comput Biol* 2000, **7**:601–620.

[38] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**(4):370–377.

[39] Qu Z, Weiss JN, MacLellan WR: **Regulation of the mammalian cell cycle: a model of the G1-to-S transition**. *Am J Physiol Cell Physiol* 2003, **284**(2):349–364.

[40] Wiley HS, Shvartsman SY, Lauffenburger DA: **Computational modeling of the EGF-receptor system: a paradigm for systems biology**. *Trends Cell Biol* 2003, **13**:43–50.

[41] Vert JP, Kanehisa M: **Extracting active pathways from gene expression data**. *Bioinformatics* 2003, **19 Suppl 2**:II238–II244.

[42] Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data**. *Stat Appl Genet Mol Biol* 2004, **3**:Article16.

[43] Breslin T, Krogh M, Peterson C, Troein C: **Signal transduction pathway profiling of individual tumor samples**. *BMC Bioinformatics* 2005, **6**:163.

[44] Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome**. *Nat Genet* 2005, **37**(6):579–583.

[45] Adler AS, Lin M, Horlings H, Nuyten DSA, van de Vijver MJ, Chang HY: **Genetic regulators of large-scale transcriptional signatures in cancer**. *Nat Genet* 2006, **38**(4):421–430.