# Large Scale Analysis of Biological Data using Physically Motivated Techniques

Thesis for the Degree of
Doctor of Philosophy

by

**Gad Getz**

Presented to the Scientific Council of
The Weizmann Institute of Science

December 2003

This work was carried out under the supervision of Professor Eytan Domany, Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot, Israel.

***To Yael and my family***

# Acknowledgments

It is a pleasure to thank my supervisor Prof. Eytan Domany. I have benefited enormously from his devoted guidance, support, knowledge and insight throughout all the stages of the work. His personality and approach have made these years a wonderful and rewarding experience. I was extremely fortunate to explore with him the new and exciting field of computational biology and gene expression analysis.

This field of research is inter-disciplinary and thus requires a close collaboration with biological groups that provide the experimental motivation, data and validation. I have been lucky to work with outstanding groups from all over the world and wish to express my gratitude for their enthusiastic cooperation and the confidence they showed in our work. Dr. Uri Alon's inspiring lecture on colon cancer gene expression data gave the impetus for our entry to his field; he helped us by providing full access to his data. Prof. David Givol was the first who shared his data with us, and has been a most generous source of knowledge and ideas. Prof. Dan Notterman was instrumental in getting us involved in a large-scale project on colon cancer, and has been a close collaborator. Prof. Irun R. Cohen introduced us to the world of antibodies and antigens. Prof. Eli Canaani's insistence on extracting biologically significant and relevant findings from his data forced us to attain the necessary statistical tools. Prof. Gidi Rechavi has shared generously his clinical and biological insights. Special thanks go to Dr. Monika Hegi and her group at University Hospital in Laussane (CHUV) who invited us to become part of their team to work on gliomas; their's was the for were the first large-scale expression data set to which we had exclusive access. Dr. Naama Barkai and Dr. Tzachi Pilpel provided helpful advice and support throughout my work.

I would also like to thank my other colleagues and collaborators from whom I have learned much and enjoyed working with. These are Erel Levine, Noam Shental, Itai Kela, Osnat Ravid-Amir, Hila Gal, Guy Hed, Sophie Godard, Kannan Karuppiah, Michele Vendruscolo, Michael Q. Zhang, Francisco Quintana, Libi Hertzberg, Or Zuk, Michal Mashiach, Omer Barad, Uri Einav, David Sachs and Alina Starovolsky.

My deep gratitude goes to my wife, family and friends for their encouragement and support along the years.

# List of enclosed publications

## Chapter 2

1. Coupled two-way clustering analysis of gene microarray data,
   *Proc. Natl. Acad. Sci.* **97**, 12079–12084 (2000),
   G. Getz, E. Levine and E. Domany.

2. Coupled Two-Way Clustering Server,
   *Bioinformatics* **19**, 1153–1154 (2003),
   G. Getz and E. Domany.

3. Classification using semi-supervised typical cuts,
   *Preprint*,
   G. Getz, N. Shental and E. Domany.

4. DNA microarrays identification of primary and secondary target genes regulated by P53,
   *Oncogene*, **20**, 2225–2234 (2001),
   K. Karuppiah, A. Ninette, G. Rechavi, J. Jakob-Hirsch, I. Kela, N. Kaminski, G. Getz, E. Domany and D. Givol.

## Chapter 3

5. Super-paramagnetic clustering of yeast gene expression profiles,
   *Physica A* **279**, 457–464 (2000),
   G. Getz, E. Levine, E. Domany and M. Q. Zhang.

6. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity,
   *Journal of Autoimmunity* **21**, 65–75 (2003),
   F. Quintana, G. Getz, G. Hed, E. Domany and I. R. Cohen.

7. Coupled Two-Way Clustering Analysis of Breast Cancer and Colon Cancer Gene Expression Data,
   *Bioinformatics* **19**, 1079–1089 (2003),
   G. Getz, H. Gal, I. Kela, D. Notterman and E. Domany.

8. Classification of human astrocytic gliomas on the basis of gene expression: A correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes,
   *Cancer Research* **63**, 6613–6625 (2003),
   S. Godard, G. Getz, M. Delorenzi, P. Farmer, H. Kobayashi, I. Desbaillets, M. Nozaki, A-C. Diserens, M-F. Hamou, P-Y. Dietrich, L. Regli, R.C. Janzer, P. Bucher, R. Stupp, N. de Tribolet, E. Domany and M.E. Hegi.

9. Expression profiles of acute lymphoblastic and myeloblastic leukaemia with ALL-1 rearrangements,
   *Proc. Natl. Acad. Sci.* **24**, 5853–7858 (2003),
   T. Rozovskaia, O. Ravid-Amir, S. Tillib, G. Getz, E. Feinstein, H. Agrawal, A. Nagler, E.F. Rappaport, I. Issaeva, Y. Matsuo, U.R. Kees, T. Lapidot, F. Lo Coco, R. Foa, A. Mazo, T. Nakamura, C.M. Croce, G. Cimino, E. Domany and E. Canaani.

10. Is there a Unique Gene-Expression Signature of Survival in Breast Cancer?,
    *Submitted*,
    I.Kela, G. Getz, L. Ein-Dor, D. Givol and E. Domany.

## Chapter 4

11. Automated assignment of SCOP and CATH protein structure classification from FSSP scores,
    *PROTEINS: Structure, Function, and Genetics* **46**, 405–415 (2002),
    G. Getz, M. Vendruscolo, D. Sachs and E. Domany.

12. FSSP to SCOP and CATH (F2CS) Prediction Server,
    *Bioinformatics*, submitted,
    G. Getz, A. Starovolsky and E. Domany.

# Contents

# Chapter 1

# Introduction

During the past decade high throughput experimental techniques led to an explosion in the amount of data describing biological phenomena. Huge amounts of data are available describing different levels of biology: (i) Genomic data - the Human Genome Project announced its completion on Feb, 2001 [1, 2] unravelling the $3 \times 10^9$ bases of the human DNA and its $\sim 30,000$ genes. Many other genomes are already finished with many more on the way. Differences among individuals are also studied and on the order of $10^7$ single nucleotide polymorphisms (SNPs) are currently known (ii) Gene expression data is measured for tens of thousands of genes in different cell-types at different conditions taken from various organisms. (iii) Amino acid sequences of hundreds of thousands of proteins are known, and (iv) Three dimensional structures, at atomic level, of thousands of proteins are publicly available. The amount of data in each of these fields calls for the development of novel analysis tools that will produce new biologically relevant knowledge [3].

This thesis can be considered as part of the global effort in *functional genomics*, for which Hieter and Boguski [4] provide the following definition: "[Functional genomics] is characterized by high throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The fundamental strategy in a functional genomics approach is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion. Computational biology will perform a critical and expanding role in this area: whereas structural genomics has been characterized by data management, functional genomics will be characterized by mining the data sets for particularly valuable information. Functional genomics promises to rapidly narrow the gap between sequence and function and to yield new insights into the behavior of biological systems."

Our research focused on developing and applying analysis methods to "mine" the data and discover new biologically meaningful information. Two specific problems were addressed; the major one is the analysis of gene expression data obtained from microarray experiments mainly obtained from samples representing various forms of cancer, and the other concerns improving and automating the assignment of protein structures to families of similar folds. Both these research directions may lead to better understanding of cellu-

lar biology (function of genes and their network of interactions) and contribute to cancer research and hopefully to cancer therapy.

## 1.1 Thesis outline

In this Introduction I give an overview of my primary field - gene expression analysis. The aim is to describe the general experimental methods, the biological questions asked and to review the analysis methods used to answer these questions. Chapter 2 is devoted to analysis methods. A detailed description of the methods we developed and used is given. In addition, other methods that were specifically designed to analyze jointly the rows and columns of a matrix, as the one we developed, are reviewed. Chapter 3 describes the problem area that played the central role in my research; analysis of gene expression data. It starts by a brief review of microarray technology and its biological uses and continues with the main biological findings in various projects, that resulted from applications of our methods. We used our techniques also for protein structure analysis, which is presented in Chapter 4. The Thesis ends with a summary which contains a discussion of all the chapters and challenges for the future.

## 1.2 Gene expression experiments

The central dogma in biology [5] describes the information flow in the cell; DNA is transcribed to mRNA which is then translated into protein. The DNA molecule, which resides in the nucleus of all the cells of an organism, stores the entire genetic information. When a certain protein is needed, the information to produce it is copied from the corresponding gene on the DNA to an mRNA molecule. This process is called *transcription* and is catalyzed by the RNA polymerase (RNAP) complex. Transcription is regulated, either enhanced or repressed, by other proteins called *transcription factors* (TFs) that bind to the gene's promoter region (a segment of DNA usually located closely before the protein coding region starts) and affect the binding of the RNAP. The mRNA leaves the nucleus and reaches a ribosome where this information is *translated* to the corresponding amino-acid sequence and the desired protein is assembled. Consequently, the number of mRNA molecules of a specific gene in a cell, the gene's *expression level*, gives a rough estimate of the generated amount of protein it codes for. Note that in higher Eukaryotes, *e.g.* humans, the mRNA is often spliced, pieces of it (*introns*) are removed and the remaining ones (*exons*) are rejoined - not necessarily all of them and not always in the original order. In these cases, a single gene may produce several different mRNA molecules which code for different proteins.

### 1.2.1 Microarray technology

Gene expression research witnessed a revolution with the development of microarray technology. Instead of measuring the mRNA level of a single, or a few, genes at a time, a single microarray experiment can measure the expression levels of tens of thousands of genes simultaneously. There are several technologies to manufacture microarrays all of which are based on the same general scheme: Tiny *probes* are placed on a surface. Each probe binds to mRNA molecules of a specific gene. mRNA is harvested from the examined cells and is labelled with a fluorescent dye. This *target* solution is poured on the surface and the labeled molecules stick to their corresponding probes. A scanning laser microscope generates an image by scanning the surface and collecting the fluorescent emission at every position. The image is then analyzed and the estimated signal from each probe is obtained. The signal of a probe corresponds to the amount of mRNA molecules of that particular gene. See Chapter 3 for further details regarding these technologies.

The two most popular microarray technologies are spotted cDNA microarrays and synthetic oligonucleotide microarrays (produced by Affymetrix$^{\text{TM}}$). From technical reasons, in cDNA microarray experiments one usually pours a mixture target solution which is prepared from two samples, each labelled with a different color (red or green). The outcome of such experiment is the ratio of gene expression levels between the two samples. In synthetic oligonucleotide microarrays, on the other hand, labelled RNA molecules from a single sample are measured. In both technologies, the probes represent either genes of known identity or expressed sequence tags (EST) which are segments of mRNA molecules extracted from cells.

State-of-the-art microarray technology can detect genes of entire genomes of many species, and recently even the human genome. This panoramic view of the cell is believed to represent the "state" of the cell and is referred to as the *molecular profile* of the cell [6]. One should keep in mind that the mRNA levels are only part of the picture since many of the cellular processes occur at the protein level and some of them leave little trace on the genes' expression levels. Nevertheless, this assumption is used and gives rise to many new insights on the cell's behavior.

### 1.2.2 Biological questions

Gene expression experiments are used to answer a variety of biological questions. The questions can be divided to two main categories. Questions of the first kind concern *genes*, *e.g.* assigning function to genes and deciphering gene pathways and regulatory networks; these fall in the realm of functional genomics. Questions of the other category are aimed at reaching better understanding of the molecular biology of the analyzed cells at various *conditions*. Examples of such questions are: which pathways are active, which genes are key players, are cells with different phenotypes also different in their molecular profile.

In a single experiment one can answer questions from both categories. The most basic question asked is a *comparative* one - which genes are differentially expressed between two

or more conditions. Even the answer for this question can be used to study both genes and conditions. On the one hand, one can gain insight regarding the biological mechanisms that operate in the analyzed condition by identifying the pathways and function of the differentially expressed genes for which these are known. On the other hand, one can also suggest function to unfamiliar genes if those act coherently with other genes of known function.

Other questions that focus on the analyzed cells and are particularly common in cancer research are *class prediction* and *class discovery*. In class prediction one tries to predict the class of a sample based on the expression level of a selected set of genes. The potential use of such classifiers is in diagnosis, prognosis and selection of therapy. The goal of class discovery is to identify distinct types and sub-types of a disease based on molecular profiles. These new types can then be studies using comparative analysis and a classifier can be trained to diagnose them. Class discovery is widely used in analysis of tumors [7, 8] since many tumor-types, as they are defined today, are composed of molecularly different sub-types which need to be studied separately and perhaps treated differently [9]. Class discovery and exploratory methods are also used learn about the relationships between the classes; which are close to which, or are they ordered in some particular way. Other questions which are also common in cancer research deal with survival analysis, in which one searches for genes that are indicative of survival or classes which have distinct survival distributions.

Gene expression studies are performed on many organisms and on different types of cells. Most common are yeast, human and mouse tissue samples (both normal or malignant samples) and cell lines. These studies have a major impact on cancer research and our understanding of cell biology understanding [9].

### 1.2.3 Data mining

In a large-scale experiment one measures gene expression in samples of cells taken from ten to hundreds of different conditions. The outcome of these experiments can be arranged in a huge array (typically of more than $10^6$ elements); each row represents one gene's *expression profile* across the conditions and each column is the *molecular profile* of a single condition [6]. Analyzing such vast amounts of data is the goal of "data mining" methods which, in conjunction with understanding of the underlying biology, can extract meaningful new biological information. Data mining in gene expression is particularly difficult since the data are very noisy and the cost of an experiment prevents performance of many replicates, which could have been used to model the noise and to average it out.

The problem of mining vast amounts of data is often initially addressed by *unsupervised* (or clustering) methods in which no explicit assumptions or hypotheses are made; one searches for regular patterns and general structure in the data. For example, when analyzing microarray experiments, each gene is usually represented by its expression level over the samples studied. Clustering genes with correlated expression levels is likely to group together genes that participate in the same or related biological processes without

knowing the nature of these processes in advance. These gene clusters may serve as a first step in the understanding biological processes by focusing the research on these genes.

The unsupervised methods are usually accompanied by supervised approaches in which one searches for patterns that correspond to external labels on the data points, that often are obtained from other sources. For example, when analyzing gene expression of samples taken from normal and diseased tissues, supervised method can find genes whose expression level can differentiate between the two tissue types.

The analysis involves an interplay between supervised and unsupervised methods. Usually, when one performs an experiment one wants to test some hypotheses; these are usually addressed using supervised methods. In addition, one want to "mine" the data and look for signals that can generate new hypotheses, which can then be tested by supervised methods.

Lately, there is growing interest in semi-supervised methods, which are in between the supervised and unsupervised extremes [10,11], which deal with partial knowledge. Actually, most problems in biology belong to this regime since there is already a large body of knowledge regarding the function, structure and interactions of genes and proteins, but this knowledge concerns only a small fraction of known genes and proteins. Therefore, one may benefit from applying such techniques to biological data.

### 1.2.4 A typical experiment

A gene expression experiment is usually conducted as follows: (i) A biological question or hypothesis is raised. (ii) The experimentalist carefully designs an experiment that can answer the question or test the hypothesis, choosing the technology, the cells, their conditions and the number and level of replications to perform. (iii) mRNA is harvested from the cells and the microarray experiments are performed. (iv) Image analysis and quality control are performed in order to extract reliable measurements out of the signals and evaluate their accuracy. (v) The data are analyzed with different data mining methods, supervised and unsupervised, in order to test the raised hypotheses and to extract any valuable information out of the data. Usually at this point one needs to turn to other sources to validate the findings, *e.g.* use other experimental techniques to measure mRNA concentrations, test the results on new samples, literature search, genomic analysis or validation by comparing the results and testing them on other datasets. The following sections are dedicated to the different steps of the experiment.

## 1.3 Experimental design

Designing a microarray experiment is not trivial. Questions arising at this stage are: (i) which microrarry technology to use? (ii) How many examples are needed from each condition? (iii) How many technical replications to perform for each sample? If one decides to use a two-dye experiment (cDNA microarrays) one has to choose which mRNA target to label with each of the dyes. Most cDNA experiments choose a common reference target

(*e.g.* a mixture of normal tissues) as a control and use it in all hybridizations. Kerr *et al.* [12] suggest more efficient labelling schemes which can reach better accuracy in identifying differentially expressed genes. The answers for the questions posed above depend on the biological questions asked, the available samples and on the chosen statistical analysis tools. The choices made at this first stage of the experiment limit the possible statistical significane of the discoveries and may hinder the whole experiment. For further discussion on the experimental design see [6, 12] and references therein.

## 1.4 Data acquisition

Data acquisition is the process of producing the readings from the scanned image of the hybridized microarray. This process depends on the specific technology used, since each technology has its own problems and common experimental errors (see Secsions 3.3 and 3.4) A . The general data acquisition procedure is as follows: first, the image is analyzed and the regions that represent each probe are identified. Next, non-specific hybridization is estimated and subtracted from the signal of each probe; in cDNA microarrays it is estimated by measuring the intensity surounding each spot, whereas in Affymetrix chips probes come in pairs: one containing the exact desired sequence, called a perfect match probe, and the other has one mismatch at the central nucleotide, called a mismatch probe. The level of hybridization to the mismatch probe is used to estimate the part of the signal of the perfect match probe, that is due to non-specific hybridization. Irizarry *et al.* [13] demonstrated that the intensity of a mismatch probe contains various levels of infromation regarding the desired signal (the one of the perfect match probe) and hence subtracting it may yield biased measurements.

Affymetrix provides a software called MicroArray Suite (MAS) [14, 15] that performes the data acquisition and quality assurance for the Genechips. The output of the MAS software are the expression levels of all the probesets[1] (genes or ESTs), called *Average Difference* in MAS 4.0 and *Signal* in MAS 5.0. In addition for each probeset a Present/Absent call is provided. This call takes into account the quality and agreement between the signals of all the probe-pairs that represent a single gene (or EST). A gene is called *Present* if the null-hypothesis that the gene was not expressed at all is rejected (MAS 5.0 also provides a $p$-value for this hypothesis).

Recently other tools were developed to generate the experession levels form the resulting images [13,16–18]. These methods use the images from all the experiments in order to obtain better statistics on systematic fluctuations of the probe signals. The three most common methods to extract the gene expression levels from Affymetrix chips are MAS 5.0 [15], RMA [13] and dChip (based on [16, 17].

Data acquisition from cDNA microarrays can be performed using several software tools, *e.g.* Quantarray™ [19] and Scanalyze [20]. These software packages identify the spots on

[1] A probeset is the set of all probes that represent a single gene or EST.

the image, evaluate the hybridization quality and extract a reading and background for each spot. Next, the background is subtracted from the spot intensity and the signal is obtained. This is performed for both dyes. Since the two dyes have different labelling efficiencies scaling is performed on one of them (usually the Cy5 is normalized). Finally, the ratio between the Cy3-signal and the normalized Cy5-signal is reported (see Sec. 3.4) for more details.

## 1.5 Preprocessing

In general, preprocessing includes transformation of the data in a way that makes the experimental noise independent, identically distributed (i.i.d) and preferably Gaussian and additive. Consequently, a value of $y$ can be considered to be a realization of $y_0+\epsilon$, where $y_0$ is the "true" value and $\epsilon$, the noise, is normally distributed centered at zero with a standard deviation of $\sigma$, *i.e.* $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In microarray experiments, the variation, $\sigma^2$, can be divided into three parts [21]; $\sigma^2_{\text{B}}$ - due to the biological variation within a specific group, $\sigma^2_{\text{A}}$ - variation between technical replicates which includes all aspects of preparing the target as well as array to array variation. The last component is $\sigma^2_e$ which stands for the variation within an array. One can measure each of these components by replicating experiments at the different levels. For example, when finding genes that are differentially expressed between two types of a disease one should collect samples from different individuals with these disease types. The variation across the patients with a certain type is due to all three components of the noise. If one wants to estimate $\sigma^2_{\text{A}}$ one can harvest mRNA from cells of the same patient and repeat the experiment; this will include noise entering from the harvesting step. A more common replication at this level is to separate the harvested RNA into several test tubes and perform a microarray experiment on each of them. Finally, to measure the variance across the array, $\sigma^2_e$, one can place a probe for the same gene at different positions on the array. In addition, part of the preprocessing is to estimate missing values which often represent defects in the experiment which render some values unreliable.

### 1.5.1 Transformation

Many analysis methods, both supervised and unsupervised, assume that the experimental noise surrounding the measurments is normally distributed with a constant variance which is independent of the mean. In this case, similar differences between the readings have the same statistical significance. This assumption, however, is invalid in gene-expression data. Fig. 1.1 depicts a graph of the standard deviation of the *log* of gene expression readings vs. their mean, measured for 8793 genes in 9 repeats[2]. The curve represents the average noise level as a functions of its mean and follows a typical shape. One can clearly see that the

[2]Data taken from a yet unpublished paper by M. Levite and measured using Affymetrix Human Focus chips and MAS 5.0 software.

noise depends on the average expression level; low readings have larger noise. The common practice in gene expresion is to select a threshold, say 6 for this data, which represents the detection level of the experimental system and mark all reading below that value as uncertain. Higher values are analyzed after taking their log and, hence, this preprocessing is called log-transformation. Analyzing the log-transformed data assumes that the variance is constant for values above 6 which is obviously only an approximation. Note that the log-transformation is not mandatory and can be avoided if one applies analysis methods that can incorporate a non-Gaussian model for the noise, *e.g.* a log-normal distribution with intensity dependent parameters.

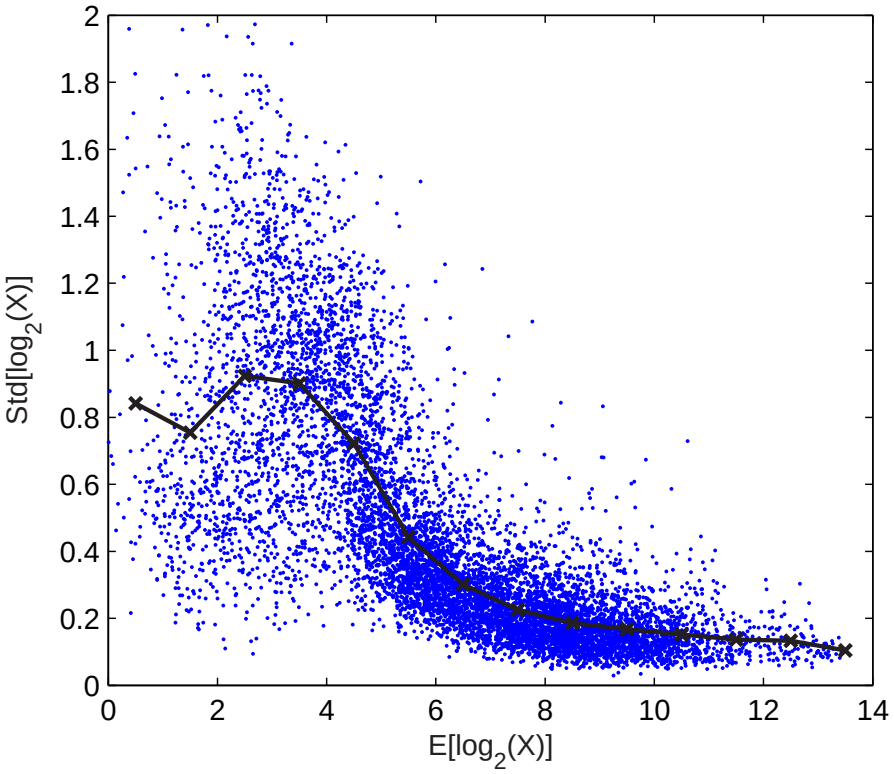



Figure 1.1: Experimental noise of gene expression data. Each point represents a gene (an Affymetrix probeset), on the $x$-axis is the mean of $\log_2(X_{ij})$ and on the $y$-axis is the standard deviation of $\log_2(X_{ij})$ where $X_{ij}$ is the Affymetrix reading for gene $i$ in repeat $j$. The solid line represents the average noise level (Std) obtained at different values of $E[\log_2(X)]$.

### 1.5.2 Scaling

Several of the factors that influence the measured fluorescent emission contribute an overall scaling factor to the readings. Examples of such factors are the labelling efficiency of the

dyes, the mRNA concentration poured on the chip and the scanning sensitivity. The aim of scaling is to estimate the value of this factor and divide the readings by it in order to bring readings from different microarrays to the same scale. Note that if scaling is performed after the log-transformation one should estimate a shifting constant and not a scaling factor.

There are several methods to perform such scaling which are based on different assumptions. The most common scaling methods assume that the total amount of mRNA molecules in each experiment is the same. Consequently, the total expression levels should be the same and, therefore, the scaling factor is estimated by either taking simply the mean, or using a more robust statistic. For example, the Affimetrix software (MAS) uses the trimmed mean (see Sec. 2.4.1) to estimate the scaling factor; others use the median. Other methods use linear regression between the gene expression levels of each sample and the expression levels of a chosen reference experiment (either a single experiment or the average of all experiments). Durbin *et al.* [22] suggest a non-linear method that takes into account the scaling factor together with a model for the noise and performs transfomation and scaling in a single step.

A different type of scaling uses "spike-controls", which are externally added mRNA molecules of known concentrations that hybridize to specific probes on the chip. These spike-controls can help eliminate differences due to the dye efficiency and parameters of the scanner. It cannot, however, correct for differences in the amount of mRNA poured on the chip. There are additional non-linear scaling methods which perform intensity dependent scaling [**?**, 22].

### 1.5.3 Missing value estimation

The data preprocessing includes estimating missing values. These can originate from defective probes or other experimental damages that render the expression level of specific genes untrustable. Troyanskaya *et al.* [23] tested several such imputing methods: (i) Replacing the missing values by zero (the expression levels are assumed to be log ratios and hence zero represents no difference from the control); (ii) Replacing the missing values by the average of the other (non-missing) values of that gene in the other experiments. (iii) Approximating the gene expression matrix by taking the $k$ leading SVD components (see 2.2.6); (iv) Approximating the missing values of a gene by the average of the $k$ most similar genes where the similarity between two genes is based on samples in which the expression levels of both genes are non-missing.

### 1.5.4 Filtering

Gene filtering is perfomed in order to remove unreliable, noisy and uninformative genes prior to the analysis. The two most popular criteria to remove genes are: (i) The number, or fraction, of missing values (or Absent in Affymterix chips) a gene has. For example, genes for which more than 10% of the samples have missing values are removed. (ii) The variance of the expression values of the gene. A gene which is approximately constant across the

samples does not help distinguish between the clusters. Typically, most of the genes are filtered out at this step ($\sim 70\%$). Filtering is higly important for unsupervised analysis since these genes only introduce noise to the differences between the samples. This noise may obscure the structure in the data. Filtering is less crucial for supervised analysis since genes which are noisy or constant will not display considerable difference between groups of samples. One should take into account that the larger is the number of genes, the more likely it is to find specific patterns in random fluctuations (see Sec. 1.7.1 for a discussion on multiple comparisons).

## 1.6 Data mining

Data mining is an active field of research that includes classic and more recent machine learning and statistical methods. The general settings is that data is collected for a set of $N$ *objects* (or *cases*). The objects can be selected according to some of their characteristics, $C_i$. For each of them $M$ additional *features* are measured $X_i = (x_1^i, \ldots, x_M^i)$, thus, the data regarding each object is the pair $(C, X)$. Consider an experiment in which one wants to study gene expression in two different tumor types, $A$ and $B$. The experimentalist designs the experiment and decides to obtain $N_A$ tumor samples of type $A$ and $N_B$ of type $B$. For each sample, the expression levels of $N_g$ genes are measured using a microarray experiment. Here the $N = N_A + N_B$ tumor samples are the objects which were selected according to their tumor type, $A$ or $B$, and each object is represented by $M = N_g$ additional features which are the gene expression levels.

Gene expression data can be organized in a form of a matrix of $N_g$ rows, one for each gene, and $N_s$ columns, one per sample. The matrix element $X_{ij}$ contains the transformed measure (usually log-transformed) of the abundance of the mRNA of gene $i$ in sample $j$. In addition to the expression data, one usually has external information concerning the samples; if, for instance, the samples were extracted from tumors of different kinds, the type of the tumor is an external label. Genes can also carry additional information such as the pathway they are known to participate in, their functional role, their genomic position, binding sites in their promoter region etc.. These external data can be organized in two separate matrices, $\mathcal{L}^G_{i\alpha}$ for genes, where $i = 1 \ldots N_g$ and $\alpha$ goes over the gene attributes and $\mathcal{L}^S_{\beta j}$ for samples where $j = 1 \ldots N_s$ and $\beta$ goes over all sample attributes. The values in $\mathcal{L}^G$ and $\mathcal{L}^S$ can be of any type (numerical, binary, ordinal) depending on the attribute they represent.

The gene expression matrix $X$ can be used to analyze both the genes and the samples. When analyzing the genes, one can interpret the matrix as a collection of $N_g$ row-vectors, each representing a gene[3], $\mathbf{g}_i = X^T_{i\cdot} = (X_{i1}, X_{i2}, \ldots, X_{iN_s})^T$ that reside in $\mathbb{R}^{N_s}$. Each gene is represented by its expression profile across all samples. On the other hand, when analyzing the samples, the matrix can be viewed as a collection of $N_s$ column-vectors,

[3]Throughout the dissertation I will assume vectors are column vectors.

$\mathbf{s}_j = X_{\cdot j} = \left(X_{1j}, X_{2j}, \ldots, X_{N_g j}\right)^T \in \mathbb{R}^{N_g}$. Each sample is represented by its molecular profile, *i.e.* the expression level of all the genes in the specific sample.

The basis of most data mining methods is statistics; one tries to study the joint probability distribution of $C$ and $X$, $P(C, X)$. One should bear in mind that there is an implicit assumption here that there exists some $P(C, X)$ from which the objects are drawn, that does not change with time or experiment. There are many data mining methods that address the questions described above. These can be generally divided to *supervised* and *unsupervised* [24–27].

## 1.7 Supervised methods

In supervised methods a subset of the features are designated as "teachers". The remaining features are used to predict the values of the teacher-features. For example, in the classical problem of supervised analysis a discrete feature which represents the class of an object (*e.g.* tumor-type) is chosen as a teacher. In this case, the prediction is called classification or *class prediction*. In a future scenario, the class of the object will be hidden and the aim of supervised methods is to predict it. In case the predicted features are continuous, these methods are called *regression*. In the example above, one may wish to predict the type of a new tumor sample based on its molecular profile. The teacher-feature is the tumor type and the remaining features, the expression levels of the genes, are used to predict it. There are numerous class prediction methods, many of which are used to analyze gene expression data. The ones more commonly used in gene expression analysis are Fisher's linear discriminant analysis (LDA) [28], perceptrons [29], multi-layer artificial neural networks, k-nearest neighbor classifiers [28], Naïve Bayes classifiers, Bayessian networks [30], classification and regression trees (CART) [31] and support vector machines (SVM) [32, 33] [34]. Linear and non-linear regression methods are also used in analyzing microarrays, especially when normalizing several experiments against each other. Details regarding these methods can be found in Duda *et al.* [24] and references therein. Recently developed methods that combine classifiers (or regressors) in order to achieve superior performance are also used in gene expression analysis, *e.g.* bagging [35, 36] and boosting [36, 37].

Another aspect of supervised analysis is *class comparison* and *feature selection*; in both of them features that best differentiate between classes are sought. The difference between them is that in class comparison one wishes to identify *all* features that are different between the classes, whereas feature selection is usually done in order to improve classification or visualization by discarding noisy features. In Publication (10) [38] we show that one can find classifiers with high performance based on disjoint sets of features (genes); thus, each of these sets is a well-chosen list for feature selection but is only a partial list regarding class comparison.

Class comparison is typically performed by hypothesis testing. For example, if one wants to identify the genes that best differentiate between two tumor types, one goes over the genes, one-by-one. For each gene one poses the null hypothesis, that the expression levels of

that gene, measured for the two tumor types, were drawn from the same distribution, and test whether this hypothesis can be rejected on the basis of the observed measurements. Supervised methods can be applied to study both samples, *e.g.* class prediction based on molecular profile, and genes, *e.g.* to identify genes whose expression profile is different in the various sample types.

The advantage of using supervised methods is that the algorithm can search for a specific pattern which is supplied externally by the "teacher". This enables to pinpoint the exact parameters of a classifier or those features that can predict with high performance the desired attributes. The ability of the external pattern to direct an algorithm to a "gold nugget" inside the large amount of data is crucial for data mining, but this advantage can turn to a disadvantage if not handled with care, since the data are very noisy and many apparent signals may be found due to chance alone. This phenomenon is called *overfitting* and is discussed below.

### 1.7.1 Class comparison - Hypothesis testing

Hypothesis testing is used when one want to test, on the basis of the observed data, whether some hypothesis is correct. Usually, if the tested hypothesis holds, no action needs to be taken; therefore, it is named the *null hypothesis*. A potential discovery is made whenever the null hypothesis is rejected and some action concerning it has to be done (*e.g.* write a paper about it). In any hypothesis testing two types of errors can occur; *false-positives*, in which one falsely rejects the null hypothesis and declares a false discovery, and *false-negatives*, when the null hypothesis should have been rejected but was not. In the latter case a discovery is missed.

Hypothesis testing is performed by first calculating a test statistic on the observed data and then determining the probability to obtain a similar or more extreme value assuming the null hypothesis is correct; this probability is called the *p-value*. Classically, one rejects the null-hypothesis whenever the p-value is less than 0.05, which controls the false-positive rate. At this working condition the probability of true-positives (true discoveries, correctly rejecting the null hypothesis) is called the *power* of the test. Tests with higher power are of greater quality. When comparing two tests at different working points, one usually plots an ROC (Receiver Operating Characteristic) curve which plots the power vs. the false-positive rate; a higher curve reflects a better test.

For example, testing whether the means of two samples are equal is generally performed using a $t$-test. For large samples the means are approximately normally distributed and for equal means the $t$ statistic follows a Student's $T$ distribution which is used to calculate the p-value. In case there are small samples or a different statistic is used for which the distribution under the null hypothesis is unknown (as in non-parametric tests), one can empirically estimate the probability using permutation methods. In these methods, one generates a null distribution in which the groups are equally distributed by randomly shuffling the observed data between the groups. The distribution of the chosen statistic is

estimated by registering the value of the test statistic in many such permutations. Permutation tests do not assume any underlying distribution and thus can be more accurate but then are much more computationally demanding if one wants to accurately measure very low p-values.

Hypothesis testing is very common in gene expression analysis and is used to identify genes that are differentially expressed between two or more classes. Some of the common statical tests are:

- the two sample $t$-test (with unknown but equal variances) testing the null hypothesis that the mean expression levels of two classes are equal.
- A non-parametric test, the Wilcoxon ranksum test (similar to Mann Whitney $U$-test) that tests whether two populations are identical, without making any assumption regarding the underlying distributions.
- The "Threshold Number of Misclassifications" (TNoM) test which was proposed by Ben-Dor *et al.* [39]. In this test one finds a threshold for the expression level of the tested gene that separates the samples into two classes with the least number of misclassifications. Ben-Dor *et al.* [40] describe a method to exactly calculate the p-value for such a separation.
- Comparing more than two classes is usually performed by ANOVA which tests whether the means of several groups are equal [41]. Two-way ANOVA is used whenever one wants to test the effect of two (or more – in higher order ANOVA) factors on a dependent variable, *e.g.* a gene's expression level (see Sec. 2.2.6 for more details).
- Fisher's exact test is used to test the relation between two partitions of set of objects. The null hypothesis is that the two partitions are independent. Fisher's exact test is used both to test partitions of samples and of genes. For samples, it is used to test whether a gene, or a set of genes, separates the samples according to some known classification of the samples. For genes, it is typically used to test if a group of genes contains an unexpected large fraction that belongs to a specific pathway or function. P-values for Fisher's exact test can be calculated analytically (both for one-sided and two-sided tests) using the hypergeometric distribution. In the gene expression literature it is sometimes referred to as the hypergeometric test. See Sec. 2.4.2 for more details.

Another common test deals with comparing survival data between two groups. These test are particularly common in cancer and disease related analyses. Consider an experiment in which one measures for $N$ patients with a certain type of cancer, the time, $T_i$, from diagnosis to the time of death (one can also measure the time to other events, such as appearance of metastases). Such an experiment is usually carried out during a fixed period of time and when it ends some of the patients may still be alive. For those patients, the time from diagnosis to the end of the experiment is recorded and they are labeled as "censored".

Their true value of $T$ is only known to be larger than the recorded $T_i$. Kaplan-Meier plots show the survival function estimated from such data (see Figure 2.7). The non-parametric Mantel-Cox log-rank test [42] tests whether the survival function of two sets of patients is the same. For example, this test was used by Botstein *et al.* [43] to show that novel sub-types of breast cancer, which were found based on their gene expression, have different survival curves and thus a different therapy may be considered.

#### Multiple comparisons

An extremely important issue in gene-by-gene supervised analysis is the problem of multiple comparisons. Applying the same rule, as described above, for the typical number of $N_g = 10,000$ genes and rejecting ones with $P_i < 0.05$, one may end up with a list of genes with many false positives. Even if the null hypothesis is correct for *all* 10000 independent genes, 500 of them will have a p-value less then 0.05 and will be falsely rejected. A strict way to address this issue is to control the "family-wise error rate", the chance of falsely rejecting even a single gene. This can be performed by a Bonferroni correction, *i.e.* rejecting p-values $< 0.05/N_g$, but this can greatly increase the number of false-negatives and miss potential discoveries. Benjamini and Hochberg [44] developed a less stringent method that bounds the "false discovery rate" (FDR), the expected fraction of false-positives. Denote the number of false positives by $V$, the number of rejected hypotheses by $R$ and define $Q = V/R$ for $R > 0$ and 0 otherwise – the FDR is the expectation value of $Q$, $E(Q)$. To bound the FDR by $q$, one orders the $N_g$ genes according to their p-values, $P_{(1)} < P_{(2)} < \ldots < P_{(N_g)}$ and rejects the null hypothesis for those genes whose index, $i$, is less or equal to $i^* = \max_j P_{(j)} < jq/N_g$ . The outcome of this method is a list of genes, for which the expected fraction of false positives is bounded by $q$. In the original proof of the FDR procedure the tests (genes, in this case) were assumed to be independent. Lately, it was also proved to work in case there is positive dependency between the genes [45, 46].

Other, permutation-based methods that address the same problems are SAM by Tusher *et al.* [47], Storey *et al.* [48–51] and Whitehead's GeneCluster software package [7].

### 1.7.2 Class prediction - training and estimating the performance

A classifier is a function, $f$, which maps its input, $\mathbf{x}$, say the molecular profile of a tumor, to one of $C$ classes (*e.g.* tumor types); $f(\mathbf{x}) : \mathbf{x} \mapsto \{1, \ldots, C\}$. The user of the classifier, *e.g.* the diagnosing doctor, defines a loss function, $\lambda(\beta|\alpha)$, which specifies the loss in case a member of class $\alpha$ is predicted to be of type $\beta$. The optimal classifier is the one which attains the minimal expected loss when presented a new case. If the joint distribution of the input and the class is $p(\mathbf{x}, C)$ then the expected loss of a classifier, its *risk*, is given by

$$R(f) = \int \sum_{\alpha} \lambda(f(\mathbf{x})|\alpha) p(\mathbf{x}, \alpha) d\mathbf{x} \ . \tag{1.1}$$

Had one known the joint distribution, the classifier with optimal performance is provided by the Bayes decision rule

$$f^{Bayes}(\mathbf{x}) = \arg\min_{\beta} \sum_{\alpha} \lambda(\beta|\alpha) P(\alpha|\mathbf{x}) = \arg\min_{\beta} \sum_{\alpha} \lambda(\beta|\alpha) \frac{p(\mathbf{x},\alpha)}{\sum_{\gamma} p(\mathbf{x},\gamma)} \ . \tag{1.2}$$

The problem is that $p(\mathbf{x}, C)$ is unknown and only a finite number of examples are available to construct the classifier.

In order to train a classifier one needs first to choose a family of functions $\mathcal{F}$ from which the classifier is selected. The classifier is parameterized by $\theta$; $f(\mathbf{x};\theta) \in \mathcal{F}$. The goal of *training* a classifier is to find a specific set of parameters $\theta^*$ that will minimize its risk. One can either estimate $p(\mathbf{x}, C)$ and then use the Bayes decision rule to perform the classification or, as is usually done, directly estimate the boundaries between the classes.

The *generalization* of a classifier is its performance on yet unseen data. In order to be able to estimate the generalization of a trained classifier, the data is broken into a *training* set and *test* set. The classifier is trained based on the training set by searching for $\theta^*$ which minimizes the error on the training set. Generalization is estimated by the performance of $f(\mathbf{x}, \theta^*)$ on the test set. This yields an unbiased estimate of the true generalization error. But since, in many cases, the test set is small, the variance of this estimator can be very large[4]

The choice of family of functions $\mathcal{F}$ has a large impact on the produced classifier. A low complexity (small) function space[5] will generally yield a poor classifier for which the performance on the training set (which is optimized in the training process) and its generalization are close; this case is called *underfitting*. This is due to the fact that functions in this space cannot precisely describe the true boundaries between the classes and, likewise, cannot fit to fluctuations in the data. On the other hand, in a high complexity (large) function space the training procedure will tend to find a classifier with high performance on the training data whose generalization is poor. In this case, named *overfitting*, the classifier learns fluctuations in the training data that lead to errors on other data sets.

In gene expression data the number of examples are usually very small compared to the diversity of the molecular profiles and size of the function space. The effect of the small sample size is demonstrated in Publication (10).

In order to avoid over-fitting, one needs to choose the appropriate function space. This is commonly done by further dividing the training set and removing from it a *validation*

[4]An estimator $\hat{t}(X^N)$ for the true value $t$ is derived from a sample of size $N$, $X^N$. $\hat{t}$ is said to be consistent if $\hat{t}(X^N) \to t$ when $N \to \infty$. The bias is defined as, $\text{bias}(\hat{t}) = \langle \hat{t}(X^N) \rangle - t$ where the average is over all samples of size $N$ . The $\text{variance}(\hat{t}) = \langle \left(\hat{t}(X^N) - \langle \hat{t}(X^N) \rangle\right)^2 \rangle$. The fluctuations of $\hat{t}$ around its true value $t$ can be decomposed to $\text{bias}^2$ and a variance term; $\langle \left(\hat{t}(X^N) - t\right)^2 \rangle = \text{bias}^2(\hat{t}) + \text{variance}(\hat{t})$.

[5]For binary classifications, the size of a function space can be measured by its Vapnik-Chervonenkis ($VC$) dimension, $d_{VC}$. It is defined as the size of the largest sample that can be shattered by the functions in the space, *i.e.* for any binary assignment of the sample there is a function in the space that can realize the assignment [27]. Larger function spaces, in general, have a larger $VC$ dimension. For example, a linear separator in dimension $d$ has a $VC$-dimension of $d + 1$.

set. The validation set is used to estimate the performance of the classifier using a specific choice of function space. The final function space used for training is the one which attains the best performance on the validation set. To better estimate the performance, especially when the initial training set is small, *cross validation* is used. In this procedure, one divides the training set to $k$ sub-samples and the overall performance is calculated by averaging the performance of $k$ different classifiers trained on $k-1$ sub-samples and tested on the remaining one. When $k = N$, the size of the training set, this is called *leave-one-out* cross validation. Van't Veer *et al.* [52] used leave-one-out cross validation to set the number of genes used in their classifier for good/poor prognosis in breast cancer.

A related method to estimate the generalization of a classifier is bootstrapping which works better in many cases [53]. In this method instead of repeatedly analyzing subsets of the data, one generates and analyzes different *subsamples* of the data. Each subsample is randomly selected, with replacements, from the full sample. Bootstrapping does not assume anything about the underlying distribution $p(\mathbf{x}, C)$ and approximates it by a sum of $N$ delta functions at the observed data,

$$p(\mathbf{x}, C) \approx \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i)\delta(C, \alpha_i). \tag{1.3}$$

The bias and variance of any statistic can be estimated by generating a large number of samples from this approximate distribution which is technically equivalent to randomly sampling with repeats from the observed data.

Many studies using supervised methods do not estimate their performance properly or at least do not publish them in a manner that reflects the noise in these estimates [54]. Common mistakes are making some decisions using the entire training data (*e.g.* choosing the genes to use for prediction) prior to performing the cross-validation, or publishing estimates based on a small test sample without reporting the confidence interval for the success rate.

## 1.8 Unsupervised methods

In contrast to supervised methods, in unsupervised ones no explicit assumptions or hypotheses are made; one searches for regular patterns and general structure that can summarize the data. Due to their exploratory nature these methods are often applied in the first steps of the analysis to give an overview and feeling for the data. Unsupervised analysis includes dimension-reduction and clustering methods.

Unsupervised methods can also be applied both for samples and for genes. Samples are represented by their expression profile over the genes, i.e. by a vector in an $N_g$ dimensional space, and genes are represented by their profile over the samples, a vector of $N_s$ components. The dimensions of both these spaces are too high to allow simple visual display and inspection.

### 1.8.1 Dimension reduction

Dimension reduction methods map the data points which reside in a high-dimensional space to a lower dimensional one, usually to two or three dimensions to allow gaining insight by visual inspection. Mappings can be linear on non-linear; the most common linear mapping is *principal component analysis* (PCA) [34] and is performed by projecting the points onto a lower dimensional hyper-plane chosen such as to maximize the variation captured by the projected points. Technically, the lower dimensional space is defined using the eigenvectors of the covariance matrix of the data that correspond to the largest eigenvalues;

$$\langle \mathbf{x} \rangle = 1/N \sum_i \mathbf{x}_i \tag{1.4}$$

$$\Lambda = \sum_i \left(\mathbf{x}_i - \langle \mathbf{x} \rangle\right) \left(\mathbf{x}_i - \langle \mathbf{x} \rangle\right)^T \tag{1.5}$$

$\Lambda$ is positive definite and therefore can be written as $\Lambda = \sum_{j=1}^{M} \lambda_j \alpha_j \alpha_j^T$ where $\lambda_1 \geq \ldots \geq \lambda_M \geq 0$ and $\alpha_n^T \alpha_m = \delta_{nm}\ \forall n, m$. The components of the projected points, $\{\mathbf{y_i}\}$, are obtained by $y_i^j = \mathbf{x}_i^T \alpha_j$. PCA can also be used to measure the effective (linear) dimension of the data by identifying the number of dimensions needed to capture, say, 80% of the total variation. A relative of PCA is singular value decomposition [34,55] (SVD). This algorithm belongs to a family of algorithms that analyze both samples and genes at the same time - Chapter 2 is devoted to such algorithms (for more details on SVD see 2.2.6). SVD identifies a list of *pairs* of vectors which Alter *et al.* [55] named *eigenarrays* and *eigengenes*; eigenarrays are a linear combination of sample vectors and thus have $N_g$ component, whereas, eigengenes "live" in a $N_s$ dimensional space. An external product of a corresponding pair of an eigenarray and an eigengene generates an $N_g$-by-$N_s$ matrix (same size as the data matrix). The aim of SVD is to decompose the original data matrix to a sum of such external products, $X = \sum_{j=1}^{\min(N_g, N_s)} u_j^T \lambda_j v_j$; each term in the sum is the external product which best approximates, in a least square sense, the residual data matrix (see Sec. 2.2.6 for a detailed description of SVD). The biological interpretation of the SVD procedure is that it separates the expression data into additive contributions from different cellular processes, the degree of activity of a process in each sample is reflected by the components of the eigengene and the effect of the process on each of the genes is manifested in the components of the corresponding eigenarray. Both PCA and SVD are linear projections which are found by optimizing a global property of the data, therefore, the resulting "summary" of the data may hide non-linear effects or ones that involve only small subsets of the data (see Chapter 2 for other algorithms that cope with these problems).

Another dimension-reduction method is multidimensional scaling (MDS) [34] [56, 57] which performs a non-linear mapping of the data points from their high dimensional space to a lower dimensional one while trying to preserve the pairwise distances among them. A cost function is defined, $E(D_{ij}^{\text{low}}, D_{ij}^{\text{orig}})$, that measures how different is the distance matrix between the representatives, in the low dimension, $D_{ij}^{\text{low}}$, from the original distance matrix,

$D_{ij}^{\text{orig}}$. The algorithm starts with some initial mapping (either random or from PCA) and performs gradient descent in the coordinates of the projected points to minimize the cost function. The main advantage of this algorithm that it can reveal non-linear structure in the data while the disadvantages are that it yields a solution which is not unique and that it is computationally heavy. Other available non-linear methods are non-linear PCA (NPCA), projection pursuit and independent component analysis (ICA) (see [34]).

### 1.8.2 Clustering

Clustering is the process of partitioning $N$ points, $\{\mathbf{x}_i\}_{i=1}^N$ into groups or clusters. Points that belong to the same clusters are "closer" (share some properties) to each other and are separated from the rest of the points. As described above, when analyzing genes, the number of points $N = N_g$ and the points reside in an $M = N_s$ dimensional space. When clustering samples, it is reversed; $N = N_s$ and the samples are represented by $M = N_g$ dimensional vectors. Prior to performing the clustering a *distance* or *similarity* measure has to be chosen. Different distance measures have different interpretations and, in general, yield different clustering results. The distance measure should be derived from the problem at hand and should be as invariant as possible to transformations that are irrelevant.

When clustering genes in order to identify co-regulated genes one usually uses the Pearson correlation as a similarity measure since one believes that genes regulated by the same transcription factor directly or indirectly or that belong to the same pathway have the correlated gene expression profiles.

Instead of using the Pearson correlation coefficient as a similarity measure between two expression profiles, $\mathbf{g}_\alpha$ and $\mathbf{g}_\beta$, it is convenient to first *center* (subtract the mean, $Y_{ij} = X_{ij} - 1/N_s \sum_j X_{ij}$) and *normalize* (divide by the norm of the vector, $Z_{ij} = Y_{ij}/\sqrt{\sum_j Y_{ij}^2}$) the gene row vectors. Then, one can use the Euclidean distance between the normalized gene vectors as the distance between genes. This distance corresponds to the Pearson correlation coefficient; $D^2(\mathbf{g}_\alpha^{\text{cn}}, \mathbf{g}_\beta^{\text{cn}}) = \sum_j (Z_{\alpha j} - Z_{\beta j})^2 = 2(1 - \text{Corr}(\mathbf{g}_\alpha, \mathbf{g}_\beta))$. This procedure enables to use any clustering method (even closed packages) since most of the use the Euclidean distance between the points. Of course, this step is unnecessary if one can use the Pearson correlation as a similarity measure directly.

If one wants to cluster together genes that are anti-correlated which might also belong to the same pathway but are down-regulated, one can replace the correlation by its absolute value or add a reversed version for each gene and cluster ordinarily.

There are other measures of similarity that are commonly used to cluster genes:

- Uncentered correlation – which is also known as the *cosine* measure;

$$\text{Uncentered Corr}(\mathbf{g}_\alpha, \mathbf{g}_\beta)) = \frac{\sum_j X_{\alpha j} X_{\beta j}}{\sqrt{\sum_j X_{\alpha j}^2} \sqrt{\sum_j X_{\beta j}^2}} \tag{1.6}$$

- Spearman's rank correlation – is an approximation to the Pearson correlation which uses the non-parametric rank statistic between two variables,

$$r' \equiv 1 - 6 \sum_j \frac{\left(\text{rank}_j(X_{\alpha j}) - \text{rank}_j(X_{\alpha j})\right)^2}{N_s(N_s^2 - 1)} \tag{1.7}$$

  where $\text{rank}_j(X_{\alpha j})$ is the rank of $X_{\alpha j}$ across all the samples. Spearman's rank correlation is a robust measure of monotone association that is used when the distribution of the data make Pearson's correlation coefficient undesirable or misleading [58]. Another rank based correlation measure is Kendall's $\tau$ [41]

- Mutual Information – is a measure taken from information theory [59,60] which tests how much information, on average, can a random varibable $X$ supply regarding the value of the random variable $Y$. In other words, how much of the uncertainty of $Y$ can be explained by $X$. Assume $p(x, y)$ is the joint distribution of $X$ and $Y$ and $p(x)$ is the marginal distribution of $X$ and $p(y)$ of $Y$, then the mutual information between $x$ and $y$ is

$$\begin{aligned} I(X;Y) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} && (1.8) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} && (1.9) \\ &= H(X;Y) - H(X) = H(X;Y) - H(Y) && (1.10) \\ &= H(X) + H(Y) - H(X;Y) && (1.11) \end{aligned}$$

  where $H(X)$ is the *entropy* or information content of $X$,

$$H(X) = -\sum_x p(x) \log p(x) = -E_X \left[\log p(x)\right] \ , \tag{1.12}$$

  and $H(X;Y)$ is the joint entropy of $X$ and $Y$,

$$H(X;Y) = -\sum_{x,y} p(x,y) \log p(x,y) = -E_{X,Y} \left[\log p(x,y)\right] \ . \tag{1.13}$$

  Note that although the definition above gave the impression that mutual information is not symmetric with respect to $X$ and $Y$ it actually is which means that the information supplied by $X$ on $Y$ equals the information $Y$ supplies on $X$. In the case $Y = X$, the mutual information attains its maximal value , $I(X;X) = H(X) = H(Y)$ and if $X$ and $Y$ are independent $I(X,Y) = 0$. Hence, one can use $I(\mathbf{g}_\alpha; \mathbf{g}_\beta)$ as a similarity measure between genes. Mutual information can be defined for discrete and continuous variables but is more straight forward in the discrete case since estimating the distributions is easier. Consequently, one has to discretize the gene expression values

by binning them which can potentially loose information [61]. D'haeseleer *et al.* [61] used mutual information to infer regulatory networds. They defined an assymetric *relative mutual informaton*, $R(\mathbf{g}_\alpha;\mathbf{g}_\beta) = I(\mathbf{g}_\alpha;\mathbf{g}_\beta)/H(\mathbf{g}_\beta)$ and used it to indicate for a possible causal relationship between gene $\alpha$ and gene $\beta$.

For measuring distances between two samples there are several approaches. Many use the Pearson correlation coefficient between also between the samples [**?**, 43, 62–64]. In our analysis we use the Euclidean distance between their columns in the row-normalized matrix. In this distance measure, the contribution of a gene to the distance between two samples is proportional to the number of standard deviations by which the values are apart; $D^2(\mathbf{s}_\mu,\mathbf{s}_\nu) = \sum_i (Z_{i\mu} - Z_{i\nu})^2$.

See Section **??** for a discussion on distance measures used to analyze genes and samples. There are many clustering methods - here I mainly describe ones used in gene expression analysis according to the following categories: representative based methods ($k$-means, SOM), model based methods (EM, Deterministic annealing), agglomerative hierarchical clustering methods (Single-linkage, Complete-linkage and Average-linkage) and density estimation methods (Super-paramagnetic clustering and probabilistic neural networks).

### Representative based methods

The aim of the representative methods is to find a set of $k$ representative vectors, $\{\mathbf{y}_i\}_{i=1}^k$, that can be used to represent each of the $N$ data points, $\{\mathbf{x}_i\}_{i=1}^N$ such that the average distortion is minimal. The distortion measures how dissimilar is a data point from its representative; the square Euclidean distance

$$D^2(\mathbf{x}_i,\mathbf{y}_j) = \sum_{a=1}^{M} \left(x_i^a - y_j^a\right)^2 \tag{1.14}$$

is widely used. Methods of this category are often used in lossy compression when one wants to choose a finite set of symbols to represent as close as possible a much larger repertoire of inputs. The simplest algorithm of this kind is $k$-means in which one guesses the number of representatives, $k$, initializes them to random positions, and then performs a two step iterative procedure until convergence: (i) assign each point $\mathbf{x}_i$ to its closest representative, (ii) update the position of each representative to the mean of the data points which are assigned to it. Each step minimizes the total distortion; for the first step this is obvious since the closest representative is chosen for each point, and in the second step positioning the representative at the mean of its assigned points minimizes the total square Euclidean distances in each cluster. This algorithm is an example of a general optimization algorithm called expectation-maximization (EM) [65] which is discussed below.

Self organizing maps (SOM) [66] is a similar algorithm, but in this case, the representatives are connected to each other by a mesh (usually one or two dimensional). Whenever a representative is moved it pulls its neighboring representatives. The idea behind SOM is that data usually lie on a non linear low-dimensional manifold embedded in the high

dimensional space. The aim of the clustering algorithm is to find it by placing a mesh of representatives that best describe the data.

A clear disadvantage of representative methods is that one needs to specify in advance $k$, the number of representatives to use (for SOM one needs to specify the topology as well). There are heuristic methods to choose the optimal $k$ [34]. Typically, a generalized cost function is created by adding to the average distortion a regularization term, which penalizes having a large number of representatives; $k$ is chosen by searching for the minimum of the generalized cost function. On the other hand, representative methods' great advantage is that these are very fast algorithms since only distances between the data points and their representatives need to be calculated. Since $k \ll N$ and does not depend on it and the number of iterations is usually not too large these algorithms are linear in $N$ with a relatively small coefficient.

### Model based methods

In model based methods one assumes that the data are drawn from some parameterized distribution function. For clustering one usually uses a $k$-mixture model; $p(\mathbf{x}|\Theta) = \sum_{\alpha=1}^{k} \pi_\alpha p_\alpha(\mathbf{x}|\theta_\alpha)$ where the parameters are $\Theta = \{\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k\}$ such that $\sum_\alpha \pi_\alpha = 1$ and each $p_\alpha(\mathbf{x}|\theta_\alpha)$ is a density function that discribes cluster $\alpha$ which is parametrized by $\theta_\alpha$. In order to generate data point $\mathbf{x}_i$ from a given model, one first draws a cluster, *i.e.* a discrete random variable $c_i$ out of $\{1, \ldots, k\}$ according to the probabilities $\{\pi_\alpha\}$, and then chooses $\mathbf{x}_i$ according to the density $p_{c_i}(\mathbf{x}|\theta_\alpha)$.

The aim of clustering is finding those parameters $\Theta^*$ that are most probable given the observed data, $X = \{\mathbf{x}_i\}_{i=1}^N$; this is called *Maximum A-Posteriori* (MAP) estimation, $\Theta^* = \arg\max_\Theta p(\Theta|X)$. If one assumes a uniform apriori distribution over the possible values of $\Theta$ one can use Bayes law and maximize the likelihood of the parameters instead (called ML estimation),

$$\begin{aligned} \Theta^* &= \arg\max_\Theta p(\Theta|X) && (1.15)\\ &= \arg\max_\Theta \frac{p(X|\Theta)p(\Theta)}{p(X)} && (1.16)\\ &= \arg\max_\Theta p(X|\Theta) \equiv \arg\max_\Theta \mathcal{L}(\Theta|X) && (1.17) \end{aligned}$$

where in the last step $p(\Theta)$ is assumed to be uniform[6]. The meaning of maximizing the likelihood is that we search for the parameters that can best explain the observed data. For a set of parameters $\Theta$ the probability to obtain the data, $p(X|\Theta)$, assuming they are an independent identically distributed (i.i.d.) sample, is given by $p(\{\mathbf{x}_i\}_{i=1}^N|\Theta) = \prod_{i=1}^N p(\mathbf{x}_i|\Theta)$. Often it is analytically more convenient to maximize the $\log \mathcal{L}(\Theta|X)$. In many cases, as in the mixture case, this maximization cannot be solved analytically and one has to turn to

[6]In practice, if the data set is large enough, *i.e.* $N \gg 1$, the prior over $\Theta$ is negligable compared to the likelihood and hence MAP and ML are equivalent.

optimization methods. One such method is the EM, *Expectation-Maximization* algorithm [65].

The idea of the algorithm is to simplify the likelihood by adding additional hidden variables which are used to decouple the parameters. The price one needs to pay is that these additional variables need to be averaged over. I will first describe the algorithm in general terms and then return to the mixture model case. Denote by $C$ the hidden (concealed) variables, which together with $X$ are called the *complete-data*, and express the model using a joint density function, $p(X, C|\Theta)$. Using Bayes law one obtains

$$\log p(X|\Theta) = \log p(X, C|\Theta) - \log p(C|X, \Theta) \ . \tag{1.18}$$

Given $\Theta$, the left-hand side of the equality is independent of $C$, whereas on the right-hand side there are two random variables that depend on $C$. Note that we are interested in maximizing $\log p(X|\Theta)$. The EM algorithm is an iterative process which increases the likelihood of the estimated parameters in each iteration by maximizing a lower bound of it. Each iteration is built of two steps; *Expectation* (E) and *Maximization* (M). Suppose $\Theta^{(t-1)}$ is our current estimation of the parameters. Since $C$ are hidden, one can only work with their distribution given $X$ and $\Theta^{(t-1)}$;

$$p(C|X, \Theta^{(t-1)}) = \frac{P(X, C|\Theta^{(t-1)})}{\sum_{C'} P(X, C'|\Theta^{(t-1)})} \ . \tag{1.19}$$

Here enters the contribution of the hidden variables; they are chosen so that the functional form of $p(X, C|\Theta)$ is much simpler compared to $P(X|\Theta)$, and hence Equ. (1.19) can be easily calculated. Next, one can eliminate the dependence on $C$ in Equ. (1.18) by calculating its conditional expectation with respect to $C$ given $X$ and $\Theta^{(t-1)}$. The result is a function of $\Theta$ (and of $X$, which is given) that has to be maximized. In other words, the log-likelihood of $\Theta$, which needs to be maximized, is calculated based on the conditional distribution of $C$ which is evaluated based on $\Theta^{(t-1)}$;

$$\begin{aligned} \log p(X|\Theta) &= \underbrace{E_{C|X,\Theta^{(t-1)}}\left[\log p(X, C|\Theta)\right]}_{Q(\Theta;\Theta^{(t-1)})} - E_{C|X,\Theta^{(t-1)}}\left[\log p(C|X, \Theta)\right] && (1.20) \\ &= Q(\Theta; \Theta^{(t-1)}) + KL\left(p(C|X, \Theta^{(t-1)}) \,\|p(C|X, \Theta)\right) + \\ &\quad H\left(p(C|X, \Theta^{(t-1)})\right) && (1.21) \end{aligned}$$

where the in the last step $E_{C|X,\Theta^{(t-1)}}\left[\log p(C|X, \Theta^{(t-1)})\right]$ is added and subtracted, and, the entropy of $p(x)$, $H(p(x)) = -E_{p(x)}\left[\log p(x)\right]$, and the Kullback-Leibler divergence between $p(x)$ and $q(x)$, $KL(p(x)\|q(x)) = E_{p(x)}\left[\log\left(p(x)/q(x)\right)\right]$, are used. Calculating $\log p(X|\Theta) - \log p(X|\Theta^{(t-1)})$ using Equ. (1.21) one obtains

$$\begin{aligned} \log p(X|\Theta) - \log p(X|\Theta^{(t-1)}) &= Q(\Theta; \Theta^{(t-1)}) - Q(\Theta^{(t-1)}; \Theta^{(t-1)}) \\ &\quad + KL\left(p(C|X, \Theta^{(t-1)}) \,\|p(C|X, \Theta)\right) && (1.22) \end{aligned}$$

since $KL(p(x)\|p(x)) = 0$ and $H\left(p(C|X,\Theta^{(t-1)})\right)$ does not depend on $\Theta$ and hence drops out. Equ. (1.22) together with the fact that $KL(\cdot\|\cdot) \geq 0$ proves that selecting $\Theta^{(t)}$ that increases $Q(\Theta;\Theta^{(t-1)})$ also increases $\log p(X|\Theta)$ (this proof is based on [67] and [68]). In the E-step of the algorithm, the conditional expectation $Q(\Theta;\Theta^{(t-1)})$ is calculated and in the M-step it is maximized with respect to $\Theta$, $\Theta^{(t)} = \arg\max_\Theta Q(\Theta;\Theta^{(t-1)})$, thus increasing $\log p(X|\Theta^{(t)})$ until it reaches some local maxima. Note that this is *not* a gradient ascent method and, in principle, each step can move $\Theta$ far away and not necessarily towards the closest local maximum.

In the case of the mixture model, the hidden variables $C$ are the cluster assignment of each data point. The log-likelihood of the complete data is simply

$$\log p(X,C|\Theta) = \sum_{i=1}^{N}\sum_{\alpha=1}^{k}\delta(c_i,\alpha)\log\left[\pi_\alpha \mathcal{N}(\mathbf{x}_i;\mu_\alpha,\Sigma_\alpha)\right] \tag{1.23}$$

assuming the $p_\alpha(\mathbf{x}|\theta_\alpha)$ are Gaussians. The conditional distribution of $C$ given $X$ and $\Theta^{(t-1)}$ can then be written as

$$\begin{aligned}\log p(C|X,\Theta^{(t-1)}) &= \log p(C,X|\Theta^{(t-1)}) - \log\sum_{C'} p(C',X|\Theta^{(t-1)}) &(1.24)\\ &= \sum_{i=1}^{N}\sum_{\alpha=1}^{k}\delta(c_i,\alpha)\log\left[\frac{\pi_\alpha^{(t-1)}\mathcal{N}(\mathbf{x}_i;\mu_\alpha^{(t-1)},\Sigma_\alpha^{(t-1)})}{\sum_\beta \pi_\beta^{(t-1)}\mathcal{N}(\mathbf{x}_i;\mu_\beta^{(t-1)},\Sigma_\beta^{(t-1)})}\right]. &(1.25)\end{aligned}$$

The EM algorithm for the Gaussian mixture is as follows:

**E-step**

$$\begin{aligned}Q(\Theta;\Theta^{(t-1)}) &= E_{C|X,\Theta^{(t-1)}}\left[\log p(X,C|\Theta)\right] &(1.26)\\ &= \sum_{i=1}^{N}\sum_{\alpha=1}^{k}\underbrace{E_{c_i|\mathbf{x}_i,\theta_\alpha^{(t-1)}}\left[\delta(c_i,\alpha)\right]}_{w_{i\alpha}}\left(\log\pi_\alpha + \log\mathcal{N}(\mathbf{x}_i;\mu_\alpha,\Sigma_\alpha)\right) &(1.27)\end{aligned}$$

$$w_{i\alpha} = \frac{\pi_\alpha^{(t-1)}\mathcal{N}(\mathbf{x}_i;\mu_\alpha^{(t-1)},\Sigma_\alpha^{(t-1)})}{\sum_\beta \pi_\beta^{(t-1)}\mathcal{N}(\mathbf{x}_i;\mu_\beta^{(t-1)},\Sigma_\beta^{(t-1)})} \tag{1.28}$$

**M-step**

$$\pi_\alpha^{(t)} = \arg\max_{\pi_\alpha} Q(\Theta;\Theta^{(t-1)}) \;\Rightarrow\; \pi_\alpha^{(t)} = \frac{1}{N}\sum_i w_{i\alpha} \tag{1.29}$$

$$\mu_\alpha^{(t)} = \arg\max_{\mu_\alpha} Q(\Theta;\Theta^{(t-1)}) \;\Rightarrow\; \mu_\alpha^{(t)} = \frac{\sum_i w_{i\alpha}\mathbf{x}_i}{\sum_i w_{i\alpha}} \tag{1.30}$$

$$\Sigma_\alpha^{(t)} = \arg\max_{\Sigma_\alpha} Q(\Theta;\Theta^{(t-1)}) \;\Rightarrow\; \Sigma_\alpha^{(t)} = \frac{\sum_i w_{i\alpha}(\mathbf{x}_i-\mu_\alpha^{(t)})(\mathbf{x}_i-\mu_\alpha^{(t)})^T}{\sum_i w_{i\alpha}} \tag{1.31}$$

Finally, since the EM algorithm can get caught in local maxima it is important to initialize its parameters with reasonable values. In the case of Gaussian mixture case, it is often initizalized by the outcome of a $k$-means algorithm.

### Deterministic annealing

Deterministic annealing (DA) was introduced by Rose *et al.* [69] and is based on and uses the scheme described in 1.11. For every assignment of the data points to clusters they use the energy functon of the k-means algorithm, see Eq. (1.14). That is, assignment of a point, $\mathbf{x}_i$, to the cluster $C_\alpha$, draws the penalty $E(\mathbf{x}_i \in C_\alpha) = \|\mathbf{x}_i - \mathbf{y}_\alpha\|^2$ where $\mathbf{y}_\alpha$ is the centroid of cluster $C_\alpha$.

The probability of assigning $\mathbf{x}_i$ to any of the $k$ clusters, at a given temperature $T$, can then be calculated by

$$P(\mathbf{x}_i \in C_\alpha) = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{y}_\alpha\|^2/T\right)}{\sum_{a=1}^{k} \exp\left(-\|\mathbf{x}_i - \mathbf{y}_a\|^2/T\right)} \ . \tag{1.32}$$

The most probable set of centroid positions can be obtained by minimizing the free energy, given by

$$F(\{\mathbf{y}_a\}_{a=1}^{k}; T) = -T \log Z = -T \sum_{i=1}^{N} \log \left[ \sum_{a=1}^{k} \exp\left(-\|\mathbf{x}_i - \mathbf{y}_a\|^2/T\right) \right] \ . \tag{1.33}$$

Finding the global minimum of such a function is computationally hard. The deterministic annealing algorithm uses the following heuristic: start at a very high temperature, at which the solution is known; all the k centroids coincide and are located at the center of mass of all the points. Gradually lowering the temperature, follow the minimum of the free energy.

As the temperature is lowered, the free energy passes through a series of phase transitions, at each of which one centroid splits continuously. The problem with this method is that these phase transition are, in general, first-order [70]. This means that the position of the new centroids that exist below a transition can be far away from the position of the single centroid, which they replace, and which minimizes the free energy just above the transition. In other words, just below such a transition the value of the free energy at a distant minimum drops below the one corresponding to the temperature above the transition.

Alon *et al.* [71] use a variant of the deterministic annealing method in which only two centroids are split at each iteration generating a binary tree.

### Agglomerative hierarchical methods

Another family of clustering methods which are widely used in gene expression analysis are the agglomerative hierarchical methods. The most popular members of this family are single-linkage, complete-linkage and average-linkage. A variant of average-linkage is used by M. Eisen's `Cluster` program [62] which became very popular in the gene expression analysis literature [6, 63]. These conceptually simple algorithms are greedy iterative methods in which one starts at the highest resolution where each point is its own cluster. Then, in a series of $N - 1$ iterations, the closest pair of clusters are united until at the end, at

the lowest resolution, one cluster is left. The various members of this family differ by how the distance between two clusters is measured; single-linkage defines the distance between two clusters as the distance between their closest pair. Complete-linkage, takes the other extreme, and measures the distance between clusters by the distance between their furthest pair. Average-linkage, in between, uses the average distance between all pairs. Eisen's variant measures the distance between the center of mass (mean) of each cluster.

The choice of distance between clusters may have a dramatic effect on the clustering results produced for the same data. Complete-linkage tends to yield spherical clusters and is closer to representative based methods, whereas single-linkage can generate clusters of any shape but is susceptible to noise in the data that might generate spurious filaments between clusters or cracks within a cluster. Single linkage is closer to density estimation methods (see below). Average-linkage produces clusters which are somewhat in between [34].

These methods do not produce clusters but rather a *dendrogram* (see Fig. **??**a), a tree that describes the nested partitioning of the data. The height at which two branches connect represents the distance between the joined clusters. Clusters can be identified by cutting the dendrogram at some level or by searching for statistically significant long branches []. Other *internal* scores exist to measure the quality of the cluster, such as the ratio between inner-cluster distances and inter-cluster ones. This score prefers compact clusters compared to elongated ones. There are also *external* methods to identify the clusters; these involve additional information regarding the clustered objects, *e.g.* if one knows the tumor-type of some of the samples one can choose to cut the dendrogram at the level in which the members of a cluster are of the same type.

As seen in Fig. 1.2a the dendrogram imposes some linear ordering on the points, which are actually the leaves of the dendrogram. This order is not unique since for each inner node the left and right branches can be switched. In any such ordering most pairs of close leaves in the one-dimensional ordering are united in one cluster not too far up the dendrogram. Such orderings, produced by two clustering operations, of the genes and of samples, are used to reorder the rows and columns of an expression matrix for the purpose of visualization, see Fig. 1.2b. Nearly every paper analyzing gene expression data includes such a figure.

### Density estimation

A different aspect of clustering is estimating the density of the data points in order to estimate the underlying distribution function which generated the observed points. Particularly, one defines clusters as modes of the distribution, *i.e.* dense regions separated by sparse ones. Classical methods for density estimation are Parzen windows [34], valley seeking [34] and probabilistic neural networks [34]. Another family of density estimation methods are based on cutting graphs. One constructs a graph as follows: each data point is represented by a vertex in the graph and neighboring points are connected with edges. The edges are usually weighted using a positive decreasing function of the distance between the points. The cost of a cut in the graph is usually the sum of weights of the broken edges.

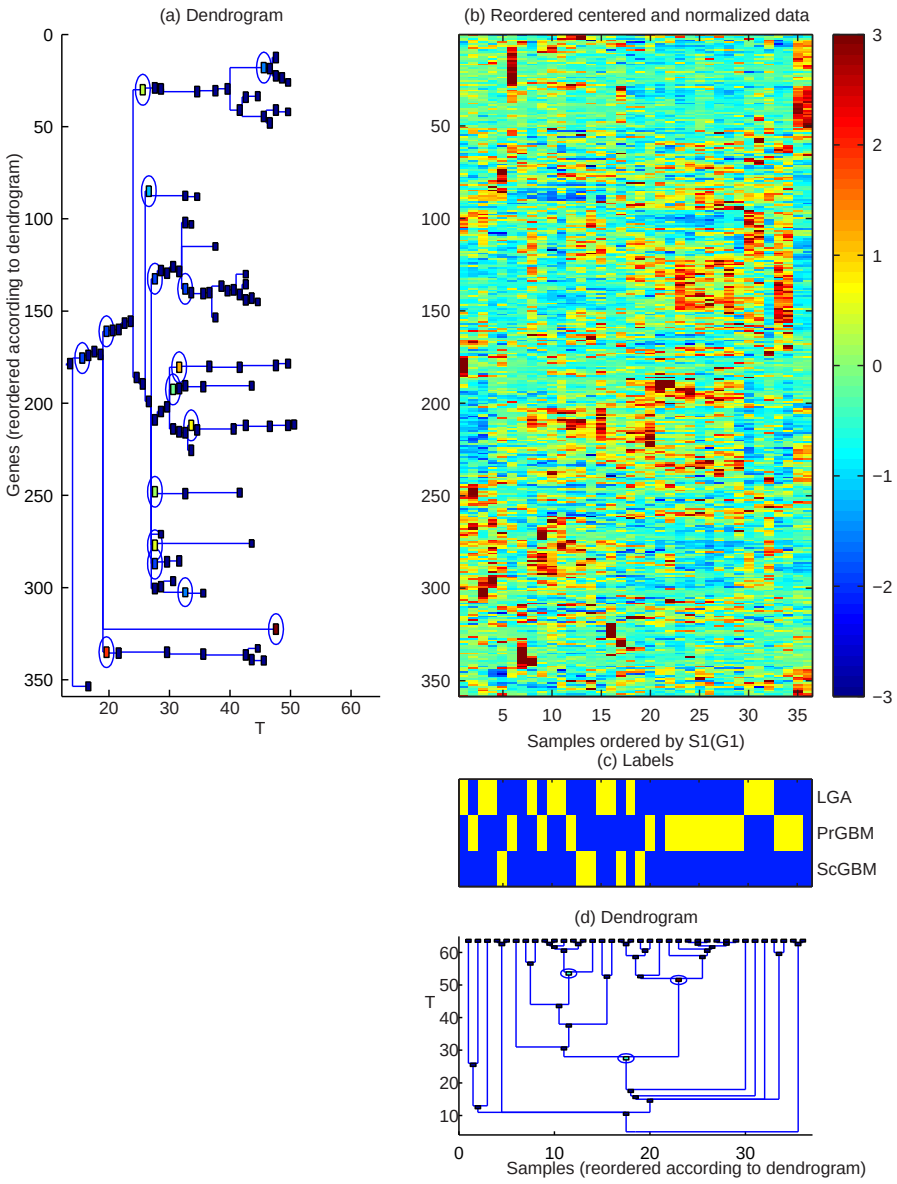



Figure 1.2: Two-way clustering of gene expression data taken from our project on glioblastoma (Publication (8)). (a) The dendrogram of the genes; (b) The reordered data in which the rows are organized according to the genes' dendrogram and the columns according to the samples' dendrogram; (c) The labels of the samples (see 3.6.3 and Publication (8)); (d) The dendrogram of the samples.

The basic idea is to penalize assigning two neighboring points to different cluster. In order to avoid trivial cuts (*e.g.* cutting out a single point) some algorithms incorporate in their costs the sizes of the remaining connected components [73]. Typically, graph-based algorithms search for the cuts with minimal cost. Blatt *et al.* suggested a graph based method

called super-paramagnetic clustering (SPC) [74] which is based on an analogy to a granular ferromagnet. Here I describe SPC, which is the algorithm we chose to perform clustering (see Publications (1)-(9),(11)-(12)); it has several advantages discussed below, which make it most suitable for our needs. The algorithm follows the general scheme described in Sec. **??**; it defines a cost (or energy) function and instead of searching for its minimum, the clustering solution is revealed by "typical" properties of the system at different average energies (controlled by the temperature). The cost function used in SPC is the energy function of a energy of a Potts model of a granular ferromagnet. Each point $i$ is assigned a $q$-state Potts spin, $s_i = 1, \ldots, q$, which represents the point's local [7] cluster assignment. Edges are connected between points using the $K$-mutual neighborhood method, *i.e.* spins $i$ and $j$ are considered neighbors if $i$ is among the $K$ nearest neighbors of $j$ and $j$ is among the $K$ nearest neighbors of $i$ (see Sec. 2.2.5 for discussion on choosing $K$). Neighboring spins which are not assigned to the same cluster pay a penalty of $J_{ij} = J\left(D\left(\mathbf{x}_i, \mathbf{x}_j\right)\right) \geq 0$ that decays with distance; these penalties are the analogs of the ferromagnetic interactions. The configuration of the system, marked here by $\mathcal{S} = \{s_i\}_{i=1}^{N}$, is the state of all the spins. The energy function, $\mathcal{H}(\mathcal{S})$, is the sum of all penalties reflecting "unsatisfied" interactions along cluster borders. The energy is defined as $\mathcal{H} = \sum_{\langle i,j \rangle} J_{ij}\left(1 - \delta(s_i, s_j)\right)$ where $\langle i, j \rangle$ represent a neighboring pair and $\delta(s_i, s_j)$ is the Kronecker delta function, *i.e.* $\delta(s_i, s_j) = 1$ if $s_i = s_j$ and 0 otherwise.

Clusters at temperature $T$ are identified by connecting neighboring points whose probability of being assigned the same label, $C_{ij}(T)$ is greater than 0.5;

$$C_{ij}(T) = \langle \delta(s_i, s_j) \rangle_T = \frac{1}{Z(T)} \sum_{\{\mathcal{S}\}} \delta(s_i, s_j) \exp(-\mathcal{H}(\mathcal{S})/T) \ . \tag{1.34}$$

These probabilities are estimated using a Monte-Carlo simulation or, in a recent work by Barad *et al.* [75], using a mean field approximation. Shental *et al.* [76] showed that in a two-dimensional problem $C_{ij}(T)$ can be approximated using belief propagation and generalized belief propagation (algorithms for inference in graphical models). Since $C_{ij}$ is estimated on the basis of Monte Carlo sampling of the ensemble of possible label assignments (or Potts spin configurations), SPC is not a deterministic algorithm, *i.e.* it can generate slightly different results in each run. This, of course, depends on the length of the Monte-Carlo runs. See Appendix B for the relation between graphical models and statistical mechanics. In Publication (3) we use advanced Monte Carlo methods and graphical models to analyze a similar problem in which some spins are fixed (see Sec. **??**), corresponding to a subset of data points with known class or labels.

SPC usually adds a "growth" step whose aim is to attach the lower parts of the mode in the density function to the cluster. The original "growth" algorithm suggested by Blatt *et al.* [77] is to connect each point to its neighbor with which it has the highest correlation,

[7] Two non-interacting spins with the same Potts state will not necessarily be assigned to the same cluster. The Potts states only represent the *local* relations between the clusters; neighboring points which are not in the same state represent that a cluster border passes between them.

as long as it is greater than some very low threshold. This step is less reproducible than identification of the cluster's core (on the basis of requiring $C_{ij}(T) > 0.5$) since small fluctuations in the estimated correlations can connect different points. Other algorithms were suggested to replace this directed growth step [**?**].

Density estimation methods can be used to generate a hierarchy of clusters by raising a density threshold $\tau$ from 0 to the maximal density. At a certain value of $\tau$, points $\mathbf{x}$ that belong to regions in which the density is below the threshold $\rho(\mathbf{x}) < \tau$, are not assigned to any cluster and each separated region above the threshold is a cluster. This procedure obeys, by construction, the dendrogram criterion; clusters which are separated at a low $\tau$ can never unite at higher value of $\tau$. As in any hierarchical solution, there is no need to specify the number of clusters apriori.

In SPC the temperature at which $C_{ij}(T) = 0.5$ correspond to the density in the region of bond $\langle i, j \rangle$. At $T = 0$, the system is at its lowest energy configuration with all spins assigned to the same state. In this case $C_{ij}(0) = 1$ for all pairs and there is a single cluster. At a very high value of $T$, $C_{ij}(T) \to 1/q$ and no pair is connected; thus each point is its own cluster. The structure of the density of points is revealed by scanning the values of $T$ and constructing the dendrogram.

The advantage of SPC over other density estimation methods is that $C_{ij}$ depends not only on the direct interaction between points $i$ and $j$, which reflects their distance and local density, but rather is a weighted sum of contributions from all the paths that connect the two points [78]. This can reduce the effect of fluctuations in the observed distances by locally averaging the density.

A major difference between density estimation methods and representative methods is their treatment of non-spherical and, in particular, elongated clusters. Density estimation methods will define a dense region, of any shape, as a single cluster; representative methods will scatter the representatives and arbitrarily break the elongated high density region into compact shaped regions, in order to minimize the average distortion.

Density estimation methods have a natural measure for the statistical significance of a cluster. A cluster is said to "live" for a range of densities (temperatures, in SPC) defined by $\tau_1$, the density threshold in which the cluster was created (separated from its environment), and $\tau_2$, the density at which it breaks into small pieces. $\tau_2$ is determined using some arbitrary thresholds on cluster size, such as the number of points lost from its birth etc. The difference between these densities, $\Delta\tau = \tau_2 - \tau_1$, can be used as a test statistic to measure the statistical significance of the cluster. The null hypothesis is that there is actually no cluster and the density is uniform, in which case $\tau_1 = \tau_2$. The statistics of $\Delta\tau$ can be estimated using permutation tests, as those we performed for SPC in Publication (6). A uniform density can be easily identified since there are no large clusters that are stable for a wide range of densities.

Blatt *et al.* suggest to measure another observable, called susceptibility, denoted by

$\chi(T)$, which is proportional to the variance of the size of the largest cluster, $m(\mathcal{S})$;

$$m(\mathcal{S}) = max_\alpha \sum_{i=1}^{N} \delta(s_i, \alpha) \tag{1.35}$$

$$\chi(T) \propto \mathrm{Var}_T(m(\mathcal{S})) = \langle (m(\mathcal{S}) - \langle m(\mathcal{S}) \rangle_T)^2 \rangle_T \ . \tag{1.36}$$

Near the temperature at which the largest cluster breaks up there are large fluctuations in its size, and hence the susceptibility has a peak. At these temperatures the configurations in which the cluster is still intact and ones in which it is broken up are equally probable, yielding a large variation in the largest cluster size. The susceptibility can be used to identify temperatures at which dramatic changes in the clusters' structure occur. Temperature intervals between peaks of the susceptibility mark stable clustering solutions. Note that the susceptibility is a global measure that exposes only changes in the largest cluster. Similar observables can be defined for local regions. For example, one can measure $\chi_i(T)$ for each cluster $i$ found at a lower temperature in order to identify the temperature at which it breaks (an alternative way to define $\tau_2$).

### 1.8.3 Statistical significance of a cluster

Determining the validity (or statistical significance) of a finding is important also for unsupervised methods. Unsupervised methods can also overfit the data by "learning" noise or fluctuations, *e.g.* identifying clusters that were formed by chance, or unnecessarily separating a cluster due to a randomly formed gap in the data. Levine *et al.* [79] suggested a method of measuring the statistical significance of a cluster by measuring its stability against repeatedly generating subsamples from the data, and clustering them. The stability score for a cluster (at a certain value of $T$) is the fraction of pairs of it's points that belong to the same cluster, averaged over the clustering operations performed for the subsamples.

Another way to assign a $p$-value to a cluster is to test the null hypothesis that actually the data is uniformly distribuated and the obtained cluster has been formed by chance. This is often performed using a permutation test in which clusters from the null hypothesis are generated by clustering randomly permuted gene expression matrices. A test statistic is chosen that represents the difference between the cluster and its environment, *e.g.* the density difference between them. The probabilty to obtain clusters with equal or more extreme test statistic than the observed cluster is the assigned $p$-value for the cluster.

### 1.8.4 Biclustering methods - clustering rows and columns simultanously

Biclustering methods are unsupervised methods that search for patterns in a data matrix. Specifically, they search for a submatrix, defined by a set of genes $G$ and a set of samples $S$ such that the expression submatrix, $X(G, S)$, has some desired properties. In the simplest

case, the sought property is flatness, *i.e.* submatrices for which the squared deviations from its mean are small. More comlicated scores are available. Searching for a submatrix with optimal score by exhaustive search is impractical since there is an exponential number of submatrices. Since we developed a biclustering method, called coupled two-way clustering, a detailed description of these methods is presented in Chapter 2.

## 1.9 Supervised vs. Unsupervised methods

Supervised and unsupervised methods complement each other. When an experiment is design one usually has some hypotheses in mind that need to be tested. Naturally, this is done using hypothesis testing and other supervised methods. These should be done with special care in order to avoid false discoveries; multiple comparisons need to be dealt with and one needs to beware not to assume normality (or other distributions) when it is not appropriate. In class prediction over-fitting is the common pitfall - this is particularly dangerous in gene expression where the number of samples is usually very small compared to the dimensionality of the data.

Unsupervised methods are used to discover new patterns in the data that were not anticipated in advance. These can then be tested using supervised methods. Another common use of unsupervised methods is to validate hypotheses which are already known. If, say, a known sample type is found as a cluster by a clustering algorithm, or is distinct from the other types in any projection to low dimensions, one can gain confidence in their separation, even if only a few genes are found to be differentially expressed in this type. The main two reasons for this are: (i) Unsupervised methods are less inclined to overfit the data since there is no specific pattern which they are looking for and (ii) The definition of types is in many cases subjective and error-prone. Identifying the members of a certain type as one cluster substantiates the class assignment. Therefore, unsupervised methods are also used as a data cleansing step prior to any analysis. For example, in our work both on leukemia (Publication (9)) and on colon cancer (Publication (7)) we grouped in an unsupervised manner a few samples of a certain type together with a different class. In the colon cancer project samples which were labelled as tumors joined the normal class and checking with experamentalists revealed that the fraction of tumor cells in those samples was very low. In leukemia we discovered that several patients whose expression profile deviated from their "known" class, and hence were thought to be misclassified, constituted, in fact, a possible new sub-type of leukemia. Perou *et al.* [80, 81] and Sorlie *et al.* [43, 64] analyzed breast cancer patients with unsupervised methods despite the fact that they had external labels which can be used in supervised analysis. This is performed since the known classes of tumors, which were initially thought to be homogenous, turned out to be heterogenous groups which coule be divided further on the basis of gene expression data.

In Publication (4) we used clustering in a novel manner to eliminate the dependence on arbitrary values of filtering parameters and thresholds. We first applied a strict filter on the genes and among them we identified those, whose expression fitted our searched pattern

(in a supervised manner). Both the filter and the supervised test were based on arbitrarily chosen values of various parameters. In the next step, we significantly relaxed the filtering criteria and applied clustering to a much larger set of genes. Finally, we identified those gene-clusters which contained a major fraction of the genes found in the first step. This two step procedure can "fish" the relevant genes while eliminating the arbitrariness of the chosen parameters and placing the thresholds at more natural and data dependent values.

## 1.10 Semi-supervised methods

In semi-supervised problems one has incomplete information regarding the data. The simplest case is of partial labels. In this case, one knows the classification of a (usually small) fraction of the data points and wants to use this knowledge to say something about the structure and classification of the remaining points.

Semi-supervised algorithms is a young field of research. Most algorithms that address this problem define some weighted graph between neighboring points (as done in SPC) and search for the minimal $k$-way cut of this graph [10, 11], *i.e.* the cut with minimal weight that separates the labelled points known to be of different types. This problem is known to be NP-hard for $k \geq 3$ [82]. In Publication (3) we suggested a model, in the same spirit of SPC, based on an inhomogeneous Potts model with the number of Potts states $q$ exceeding the number of known classes. The points with known classification are fixed in the Potts state that corresponds to their label, *i.e.* they "feel" an infinite field to that direction. The unlabelled points are free spins. The statistical mechanics of such a system is not trivial since it may suffer from frustration.

The $k$-way cut problem is equivalent to finding the ground state of this system. In Publication (11) we suggest a heuristic algorithm that tries to find this ground state, the configuration at $T = 0$. The algorithm is a greedy iterative algorithm that freezes one spin at time. In some cases it performs an optimal step and in others the least "damaging" one.

In Publication (3), on the other hand, we analyze this problem at $T > 0$. We follow the scheme described in Sec. 1.11. A toy problem is studied in which the $T = 0$ solution does not yield the correct partition which can be obtained at a higher temperature. This shows that the solution of a computationally hard problem at low $T$ may not only be unnecessary, but can lead to erroneous results, while solving an easier problem gives the correct solution. Further details regarding both methods appear in Sec. **??**.

## 1.11 Relation to statistical physics

Many supervised and unsupervised methods, either parametric or non-parametric, define a merit function which they try to optimize. Methods which are based on a statistical model of the data usually try to find those parameters, $\theta$ that maximize the likelihood of the data, $P(D|\theta)$, (called ML) or following a Bayessian approach, the aposteriori probability,

$P(D|\theta)P(\theta)$ (called MAP). Searching for the optimum has two major disadvantages: First, this procedure ignores the robustness of the found solution, *i.e.* how does the probability drop in the neighborhood of the solution. One can imagine a merit function that has a sharp peak at $\theta_1$ and a lower but much wider peak at $\theta_2$ (see Figure 1.3); which of these is a better solution? The second disadvantage is that finding the optimal solution is in many cases computationally difficult (depending on the specific merit function) - Publication (3) gives an example of a problem in which the correct solution is obtained at a lower but wider peak.

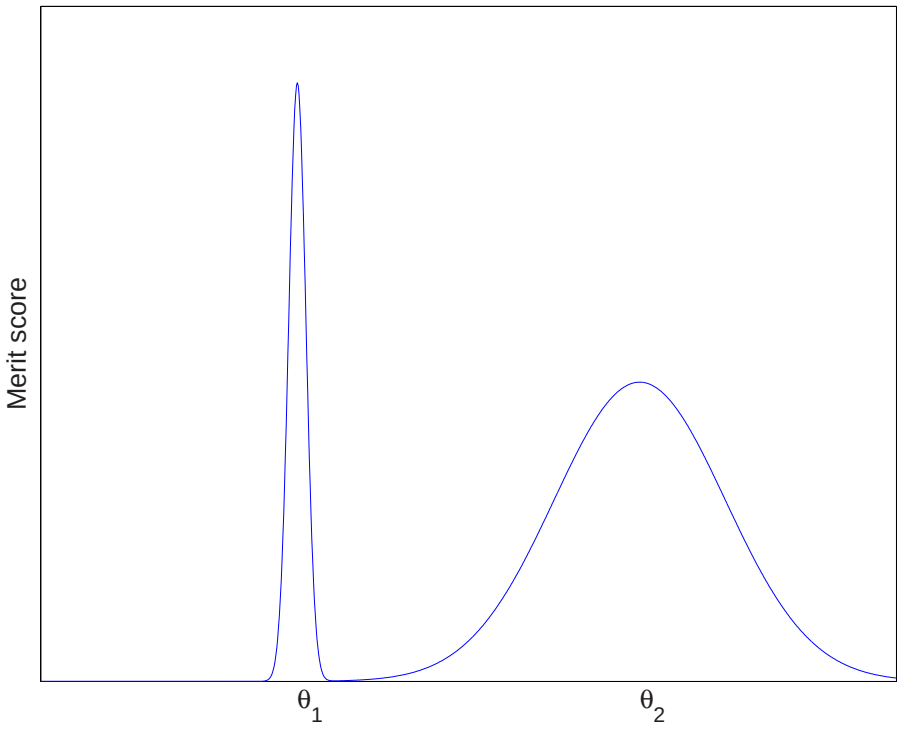



Figure 1.3: Two solutions – one with a higher score but not robust and the other is more robust but with a lower score.

To cope with these problems one can scan the merit function from its optimal value downwards and analyze the average or "typical" solution at different values of the merit function. In case there is a single peak, the average solution will not change by much as the explored value of the merit function is lowered; but in case the merit function behaves as in Fig. 1.3 the "typical" solution will jump closer to $\theta_2$ as soon as the value drops below the lower peak.

This procedure has a direct analog in statistical physics. If one defines the negative of the merit function (or any monotonically increasing function of it) as the *energy*, $\mathcal{H}(\theta)$, the optimal solution is the one with lowest energy. In the case the merit function is the posterior probability one usually uses $\mathcal{H}(\theta) = -log(P(D|\theta)P(\theta))$. The set of all parameters, $\theta$, is called in physics the *configuration*. Averages over all configurations with a certain value

of energy are performed using the microcanonical ensemble which assigns them an equal probability,

$$P(\theta) = \begin{cases} 1/Z & \mathcal{H}(\theta) = E \\ 0 & \text{otherwise} \end{cases} \tag{1.37}$$

where $Z = \sum_\theta \delta(\mathcal{H}(\theta) - E)$ and $\delta(x)$ is the Dirac delta function. Such averages are difficult to calculate and are often performed by replacing the micro-canonical ensemble with the canonical ensemble in which instead of fixing the energy value one fixes the *average energy* value. Applying the maximum entropy principle [83] this leads to a Boltzman distribution in which $\beta = 1/T$, the inverse temperature, acts as a Lagrange multiplier;

$$P(\theta; T) = \frac{1}{Z(T)} \exp(-\mathcal{H}(\theta)/T) \tag{1.38}$$

and $Z(T) = \sum_{\theta'} \exp(-\mathcal{H}(\theta')/T)$. The temperature, $T$, controls the average energy (in a one-to-one relation); low temperatures correspond to low energies. At $T = 0$ only the configuration(s) with lowest energy has a non-zero probability, which corresponds to the MAP solution.

In general, the "typical" solution which is usually defined by averages of some observables (functions of $\theta$) is calculated for a range of temperatures. The average of an observable, $\mathcal{O}(\theta)$, at temperature $T$ is calculated by

$$\langle \mathcal{O} \rangle_T = \sum_{\theta'} \mathcal{O}(\theta') P(\theta', T) = \frac{1}{Z(T)} \sum_{\theta'} \mathcal{O}(\theta') \exp(-\mathcal{H}(\theta')/T) \ . \tag{1.39}$$

This allows to explore the solution-space not only by searching for the optimal solution but also identifying abundant lower performing ones. The field of statistical physics developed many powerful tools to study such systems. In physics, especially of interest are cases and temperatures in which the typical solution has an abrupt change in its properties - these are called phase transitions.

The computational aspect of searching for the "typical" solution is that one only need to calculate averages. These can be estimated by importance sampling using Monte-Carlo methods [84,85] which usually converge much faster than required for optimization methods to find the minimal energy. In case the energy landscape is ragged, *i.e.* has many deep valleys, it may be difficult to obtain accurate estimations at low temperatures. However, in many problems there is no need to find the solutions at such low temperatures, the robust solutions are found at higher ones which are easier to sample. In case one is still interested in the solutions at low temperatures, longer runs and more advanced Monte-Carlo methods (like simulated tempering [86] and multicanonical methods [87–90]) are required (See Publication (3)).

## 1.12 Further analysis

Once a discovery is made either by unsupervised, semi-supervised or supervised methods, one has to validate the results using different experimental techniques (see next Section).

In addition one can further analyze the resulting statements using data from other sources. For example, if one identifies a cluster of genes which can separate a group of samples into two distinct types, the following questions may arise: Do these genes belong to the same biological pathway or process? Are these genes co-regulated by a known or unknown transcription factor? Are these genes located at a specific region of the genome? What are the known functions of these genes? Can we infer function to yet unfamilir genes? In which tissues/cells are these genes known to be active? Are there specific diseases associated with these genes?

There are numerous tools that can be used to start answering these questions. Most of these tools are available as websites in which one can insert a gene (or a list of genes) and obtain details regarding the genes. GeneCards [91] (`http://bioinformatics.weizmann.ac.il/cards`) is a database which integrates information from many other databases and generates a "card" for each gene with information regarding its location in the genome, its sequence, known function, its known participation in diseases and many other details. Recently, Chalifa-Caspi *et al.* [92] added a functionality to GeneCards that can display, for many genes, their expression levels at various normal human tissues.

Another popular source of information is the Gene Ontology (GO) project [93]. This is a collaborative effort whose aim is to describe a gene product by three attributes; its biological process, cellular component and molecular function. Each of these attributes is represented as a node in a tree (an ontology). Affymetrix has a website, NetAffx (`http://www.netaffx.com`), in which one can enter a list of probesets and obtain a table in which each row is a probeset and the columns represent different attributes of the gene, such as: the gene's location, its Gene Ontology, its pathway if it appears in GenMAPP [94], a database of biological pathways (`http://www.genmapp.org/download.asp`) and others.

Other tools that can be used analyze the promoter region of a gene, or a groups of genes, in order to identify transcription binding sites of either known transcription factors (TFs) [95,96] or identify putative binding sites one yet unknown TF (*e.g.* AlignACE [97]).

## 1.13 Validating the results

The aim of validation is to make sure that the conclusions that were obtained from gene expression analysis are indeed correct. Validation can be performed at the mRNA and protein levels.

### mRNA level validation

Since microarray technology is still young and has various uncontrolled sources of noise, it is common to validate the expression leves (or ratios) using different experimental assays. There are sevel alternative methods to measure gene expression; among the most common are Northern Blot, RT-PCR and *in situ* hybridization: In *Northern Blot* one separates fragments of RNA via electrophoresis. Then, the fragments are transferred to a membrane

and hybridized to labelled DNA fragments (usually via radioactive labelling) of the desired gene. This method is less quantitative than microarrays. *RT-PCR* (Reverse Transcription (RT) followed by Polymerase Chain Reaction(PCR)) is a process in which the reverse transcription product[8] is amplified using PCR. Gene specific primers[9] are used to amplify the mRNA of the desired gene. A more accurate extension of RT-PCR is called *real-time quantitative RT-PCR*. In this process the PCR products are fluorescently marked and measured during the PCR exponential accumulation phase [98]. *In situ* hybridization, on the other hand, does not use extracted mRNA but rather identifies the desired mRNA in fresh frozen or paraffin-embeded tissues and can be used to localize the mRNA expression in the tissue. This is important in case the sample contains cells of different types and one wants to know which cells express the gene. This method is based on inserting marked cDNA or cRNA (complementary to the desired gene) to the cells; cells which express the gene are coloured and can be identified under a microscope.

### Protein level validation

The most common experimental techniques to validate the protein levels are immunohistochemistry (IHC) and Western blotting. Immunohistochemistry is based on identification of the desired proteins by marked specific antigens. As in the case of *in situ* hybridization, this method can be used to detect the cell type that expresses a protein and can even detect the cellular location of the the protein (*e.g.* cell membrane, nucleus, or cytoplasm) [99]. Western blotting is similar to Northern blotting (described above) but instead of using labelled DNA fragments one uses labelled antibodies.

[8]DNA created from RNA using a reverse transcription (RT) DNA polymerase originating from retroviruses

[9]Short DNA segments needed to initiate the RT reaction

# Chapter 2

# Data Analysis Methods

## 2.1 Introduction

In this chapter I focus on the main analysis tools that we developed and used; these are divided into unsupervised, semi-supervised and supervised methods. Each method is accompanied by a review of other related methods. The main tools we developed are: (i) Coupled Two-Way Clustering (CTWC) (see Publications (1) and (2)) which is an unsupervised method to "mine" jointly the rows and columns of data that come in the form of a matrix. The algorithm searches for submatrices which yield significant partitions of the data, by iteratively clustering subsets of the genes based on subsets of the samples and vice versa. (ii) Semi-supervised SuperParamagnetic Clustering (SPC), which is used to classify data in which the correct class of a small fraction of the points are known. The algorithm is based on a Potts model with external fields. For this model we have numerical results for all temperature values using advanced Monte Carlo methods and methods for inference in graphical models (See Publication (3)), and we also used a heuristic to find a solution at $T = 0$ (See Publication (11)). In our gene expression analysis we used, in addition to the unsupervised methods mentioned above, also supervised techniques that include standard hypothesis testing methods as well as more recently introduced ones. We also suggested statistical tests to analyze gene expression in conjunction with survival data (see Sec. 2.4.4).

## 2.2 Coupled Two-Way Clustering (CTWC)

For the analysis of gene expression data our aim was to develop a method that can "mine" data that come in the form of a two-dimensional matrix (e.g. genes by samples). We noticed that the gene expression matrix is: (i) very noisy (ii) contains information regarding many processes that occur in the examined cells. Each process involves its relevant genes and may affect only a subset of the samples. Moreover, genes may have several functions and can participate in different processes depending on the biological context. Finding these

potentially overlapping sets of process-specific genes and samples may help infer a function to yet unknown genes and identify samples with common characteristics. In order to identify the signal of a single process, one has to overcome both the noise and the interference from the other processes that "mask" the desired signal. This can be achieved by focusing on small subsets of the data. An exhaustive search of all possible submatrices of a large matrix is a computationally prohibitive task; our goal was to develop a heuristic method that can identify those submatrices that contain the different signals in the matrix, and to apply it to gene expression data. Application of our methods were mainly focused on data obtained from samples representing various forms of cancer; these are described in Chapter 3.

The main idea of the CTWC algorithm is to identify subsets of the genes and conditions, such that when one is used to cluster the other, stable and significant partitions emerge. These partitions can be recognized only when focusing on these small subsets, since all other genes and samples act as noise and mask the hidden structural signal. In this section I first generally describe the CTWC algorithm. Then analysis of a particular data set, done using the CTWC server, is given as an example. I continue with a review other methods aimed at jointly selecting rows and columns of a data matrix and analyzing the resulting submatrix. One should bear in mind that the method is general and applicable to data from other fields. It is a general method to analyze the probability distribution of gene expression matrix; $p(X_{ij})$. Other fields to which it was applied include text mining [100], analysis of glycomolecules [101], antigen reactivity data (Publication (6)) [102] and the low-temperature phase of spin-glasses [103].

A nice metaphor for the problem at hand [104] is that of a football stadium, in which 99,000 spectators scream at random, while 1000 others are singing a coherent tune. These 1000 are, however, scattered all over the stadium – the chance that a listener, standing at the center of the field, will be able to identify the tune are very small. If only we could identify the singers, concentrate them into one stand and point a directional microphone at them – we could hear the signal!

We have put up a CTWC server website (`http://ctwc.weizmann.ac.il`) that enables researchers to submit their gene expression data, execute the CTWC engine that analyzes the data and view the results [105] (see Publication (2)). Resulting clusters of genes and samples are sorted according to different internal and external scores, such as size, stability and overlap with externally supplied sets. Gene clusters can also be transferred to Affymetrix' website (`http://www.netaffx.com`) or to GeneCards (`http://bioinfo.weizmann.ac.il/cards`) [91] for further biological analysis. Analysis web servers are very common in the bioinformatics community and are used on a day-to-day basis. For CTWC a PC stand-alone version also exits, based on a wrapper application written by A. Yitzhaky.

### 2.2.1 The data

In this chapter we assume that the data were already acquired and preprocessed (including scaling, transformation and missing value estimation as described in Sec. 1.5) . Since most

of our applications are on gene expression I will describe the data using terms from this field.

The input to the algorithm is the gene expression matrix $X_{ij}$ of $N_g$ rows (genes) and $N_s$ columns (samples). The matrix element $X_{ij}$ contains the transformed measure (usually log-transformed - See 1.5.1) of the abundance of the mRNA of gene $i$ in sample $j$. Besides the expression data, external information concerning the samples and the genes can be used to evaluate the resulting clusters. This information is organized in two separate matrices, $\mathcal{L}^G_{i\alpha}$ for genes, where $i = 1 \dots N_g$ and $\alpha$ goes over the gene attributes and $\mathcal{L}^S_{\beta j}$ for samples where $j = 1 \dots N_s$ and $\beta$ goes over all sample attributes. In general, the values in $\mathcal{L}^G$ and $\mathcal{L}^S$ can be of any type (numerical, binary, ordinal) depending on the attribute they represent. Note that the CTWC server currently supports only binary or unknown (empty) values.

As described in the Introduction, the gene expression matrix $X$ can be viewed as a collection of $N_g$ gene (row) vectors or $N_s$ sample (column) vectors;

$$X = \begin{bmatrix} \mathbf{g}_1^T \\ \mathbf{g}_2^T \\ \vdots \\ \mathbf{g}_{N_g}^T \end{bmatrix} = [\mathbf{s}_1 \mathbf{s}_2 \dots \mathbf{s}_{N_s}] \tag{2.1}$$

### 2.2.2 Distance measures

The algorithm performs clustering of both genes and samples. Therefore one needs to choose similarity or distance measures between a pair of genes and between two samples.

We are interested in identifying genes that belong to the same biological process; the variation of expression of such genes over the samples is expected to be correlated. Instead of using the Pearson correlation coefficient as a similarity measure between two expression profiles, $\mathbf{g}_\alpha$ and $\mathbf{g}_\beta$, we center and normalize the rows as described in the Introduction (Sec. 2.2.2) and use the Euclidean distance between them:

$$Y_{ij} = X_{ij} - \frac{1}{N_s} \sum_{j=1}^{N_s} X_{ij} \tag{2.2}$$

$$Z_{ij} = \frac{Y_{ij}}{\sqrt{\sum_{j=1}^{N_s} Y_{ij}^2}} \tag{2.3}$$

$$D^2(\mathbf{g}_\alpha^{\mathrm{cn}}, \mathbf{g}_\beta^{\mathrm{cn}}) = \sum_j (Z_{\alpha j} - Z_{\beta j})^2 = 2(1 - \mathrm{Corr}(\mathbf{g}_\alpha, \mathbf{g}_\beta)) \tag{2.4}$$

Distance between two samples is taken as the Euclidean distance between their columns in the row-normalized matrix. As described in Sec. 2.2.2, the contribution of a gene to the distance between two samples is proportional to the number of standard deviations by which the values are apart;

$$D^2(\mathbf{s}_\mu, \mathbf{s}_\nu) = \sum_i (Z_{i\mu} - Z_{i\nu})^2 \ . \tag{2.5}$$

### 2.2.3 The algorithm

Technically, the CTWC method performs iterative clusterings of sub-sets of the genes (samples) based on sub-sets of the samples (genes) and stores the identified stable clusters for subsequent iterations. In order to keep track of the clusters, two lists are recorded; $\mathcal{G}$ for gene clusters and $\mathcal{S}$ for sample clusters. Initially, G1, the cluster of all genes, is stored in $\mathcal{G}$ and similarly S1 (all the samples) is stored in $\mathcal{S}$. In the first step two independent clustering operations are performed, one of all the rows G1 based on all the columns S1, which we denote by G1(S1), and one of all the columns based on all the rows, S1(G1). This first step is known as *two-way clustering* [71]. After each clustering operation one adds the stable clusters that emerge to the appropriate list. We identify as "s table" those clusters that are statistically significant and robust against noise. The algorithm can work with any clustering method for which statistically significant clusters can be identified. We use the SPC clustering method since it has a natural way to select stable clusters (see 1.8.2). Coupling between the two ways of clustering is introduced in the subsequent CTWC iterations. Suppose that by the operation G1(S1) we found 15 stable gene clusters (which we name G2 to G16) and S1(G1) yielded 3 stable sample clusters (S2 to S4). In the second iteration we perform the clustering operations S$I$(G$J$)[1] for $I$=1,...,4 and $J$=2,...,16, as well as the complementary clusterings G$J$(S$I$). The algorithm stops when no more new stable clusters, whose size exceeds a preset threshold, are found. In practice we usually stop after between 2 to 5 iterations.

An important issue when running CTWC is how to assign statistical significance to the clusters we found. A cluster is statistically significant if it is unlikely to be obtained by chance. As discussed in Sec. 1.8.2, $\Delta T$, the difference between the temperature at which a cluster is "born" (separated from its environment) and the temperature at which it "dies" (breaks into smaller clusters) can be used as a test statistic to measure its statistical significance.

The threshold for a significant $\Delta T$ can be estimated using permutation methods. First, one has to choose a random data model to be used to generate date under the null hypothesis, and then test how often clustering data drawn from this random model yields clusters with equal or higher stability ($\Delta T$). We work with a common model that generates matrices by randomly permuting all the elements of the original matrix, producing data matrices whose elements are taken from the same distribution as the ones of the original matrix. Using these probability estimates we can calibrate our thresholds on the stability parameter, to select only significant clusters. Figure 2.1 (taken from Publication (6)) shows a comparison between clustering real antigen data and one of its randomized matrices. One can see that randomized data do not yield stable clusters (ones with large $\Delta T$).

[1] S$I$(G$J$) should be read as "clustering S$I$ based on G$J$".

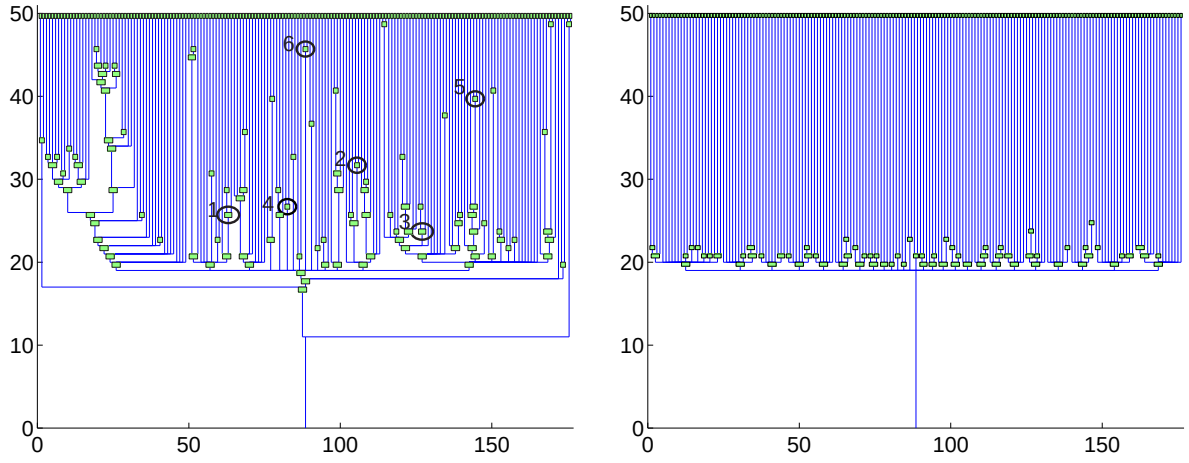



Figure 2.1: Two dendrograms of antigens obtained by clustering. The left dendrogram is of the original data and the right one is obtained from randomized data. In contrast, the randomized data consists of a single large cluster which fragments in few temperature steps into small, short-lived subclusters (see Publication (6)).

### 2.2.4 Interpreting CTWC results

The results of CTWC are the stable gene and sample clusters listed in $\mathcal{G}$ and $\mathcal{S}$, respectively. We register for each cluster the clustering operation that generated it. Other parameters, such as the stability, $\Delta T$, and the cluster size are also stored. Each entry in the lists of stable clusters represents a statistical statement; for example, "cluster S10 is a stable cluster (a mode in the distribution) with stability 5 found by clustering the samples in S5 using the genes in G3 as features, *i.e.* operation S5(G3)". Such a statement means that the genes in cluster G3, which are correlated across some other samples (according to the clustering operation that produced G3) and hence might represent a specific pathway or cellular function, possess a distinct profile in the samples of S10 compared to the other samples in S5 (note that S10 $\subset$ S5).

Next one needs to scan the potential discoveries and identify ones for further analysis. This can be done by sorting the clusters according to *internal* or *external* scores. Internal scores make use of the clustered data alone while external scores use additional information regarding the genes or samples. Naturally, one can use the size and stability of the identified clusters as scores. Clusters with larger stability are more statistically significant, *i.e.* their $p$-value is lower. Note, however, that due to density fluctuations one can find small clusters with relatively high stability even in uniformly distributed data. Tsafrir *et al.* [106] suggested two additional internal merit functions which can be used to sort the resulting clusters; *bimodality* score and *image smoothness* score. Both these scores are based on analyzing the distance matrix, $D_{ij}$, between the clustered objects (genes or samples). The first measures the bimodality of the distribution of distances; well separated clusters have low distances among the cluster members and high ones between the clusters. At first

the distribution of distances is approximated by a mixture of two Gaussians,

$$p(D) \approx a_1 \exp\left[-\frac{(x-b_1)^2}{c_1^2}\right] + a_2 \exp\left[-\frac{(x-b_2)^2}{c_1^2}\right] \tag{2.6}$$

. The bimodality score is then defined as

$$\text{bimodality}(D_{ij}) = \sqrt{\frac{a_1}{a2}}\frac{\max(c_1, c_2)}{|b_1 - b_2|} \tag{2.7}$$

assuming $a_1 > a_2$. A preferable low score is reached when the two modes are well separated and are of the same size. The *image smoothness* score is based an ordered distance matrix. The distance matrix of clusters which form low dimensional manifolds, particularly one dimensional ones, can be ordered along the manifold and thus generate a smooth distance matrix. An algorithm, named *sorting points into neighborhoods* (SPIN), that performs such an ordering is presented in Tsafrir *et al.* [106]. The image smoothness score uses standard image processing techniques to measure the smoothness of the ordered distance matrix;

$$\text{smoothness}(D_{ij}) = \frac{\mathbf{1}^T\left(\frac{\partial J}{\partial x} + \frac{\partial J}{\partial y}\right)\mathbf{1} + \frac{1}{N}\text{Tr}(JW)}{\mathbf{1}^T J \mathbf{1}} \tag{2.8}$$

where $\mathbf{1}$ is a column vector of ones, $J = D - \nabla^2 D$,

$$W_{ij} = \frac{\exp\left[-\frac{(i-j)^2}{\epsilon N}\right]}{\sum_k \exp\left[-\frac{(k-j)^2}{\epsilon N}\right]} \tag{2.9}$$

and $\epsilon$ controls the width of the neighborhood. As in the bimodality score, low values are preferable. In principle, clusters with low values in either of these scores are worth looking at first.

External scores for clusters are based on known labels of genes or samples. For example, if one knows the tumor-type of the samples, and in particular, which are of type $A$, one can check for each cluster S$I$ the extent to which it corresponds to the group of tumors of type $A$. Two quantities are measured; the clusters' *purity* with respect to $A$, $\text{purity}(\text{S}I, A) = |\text{S}I \cap A|/|\text{S}I|$ which measures the fraction of members in the cluster which are of type $A$, and its *efficiency* with respect to $A$, $\text{efficiency}(\text{S}I, A) = |\text{S}I \cap A|/|A|$ which indicates what fraction of the $A$ samples were caught by S$I$. These quantities have different names in other contexts; purity is called *specificity* or *precision* and efficiency – *sensitivity* or *recall*. In information retrieval a combined score is often used called the $F_\beta$-measure [107] which is defined as

$$F_\beta = \frac{(\beta^2 + 1)\,\text{purity} \cdot \text{efficiency}}{(\beta^2\ \text{purity}) + \text{efficiency}} \tag{2.10}$$

where $\beta$ is the relative weight of the purity compared to the efficiency. Using an equal weight, $F$ becomes the harmonic mean of the purity and efficiency;

$$F = 2\ \text{purity} \cdot \text{efficiency}/\left(\text{purity} + \text{efficiency}\right)\ . \tag{2.11}$$

In order to assign a $p$-value for the observed overlap between cluster S$I$ and group $A$, one can test the null hypothesis that the association of a sample with S$I$ and group $A$ are independent events. This can be precisely calculated using Fisher's exact test (sometimes called the hypergeometric test) or approximated by a $\chi^2$ test or $Z$-test (see Sec. 2.4.2).

Internal and external scores can be used also for gene clusters. External scores often compare the clusters to annotated gene sets that have known functions, or that are known to belong to a certain pathway – see Sec. **??** in the Introduction for tools and databases that can be used to search for gene functions.

External classification of genes and samples can be uploaded to the CTWC server as separate tables. The resulting clusters are compared with respect to each of these labels and the purity, efficiency and $Z$-score of each stable cluster are reported (see example in Sec. 2.2.5).

#### Conditional correlation

Analysis of the results can reveal additional discoveries which describe *conditional correlations* in the data. For example, if gene-cluster G3 was found to be stable in G1(S2) and broke into two distinct clusters, G10 and G11, in the analysis of G1(S7), it indicates that the genes in G10 and G11 are correlated only under the conditions of S2. Such a finding can suggest that pathways which are uncorrelated in certain conditions may become correlated in others. We found such a case in the analysis of colon cancer (Publication (1)) [108] when a cluster of growth-related genes was found to be correlated with a cluster of genes related to epithelial cells, but only over the tumor samples. This conditional correlation corresponds to the well-known biological fact that the malignant cells in colon cancer are formed from epithelial cells. Since the tumor cells are epithelial, both clusters, of epithelial genes and of growth genes, are highly expressed in tumor. The different percent of tumor cells contained in the "tumor samples" induces a variation of the measured expression levels of both kinds of genes over these samples, generating a pattern of expression with which both clusters were correlated (when correlations were measured across the tumor samples). In the normal samples, on the other hand, the fraction of epithelial cells is very small and is not correlated to the expression of growth genes.

### 2.2.5 Applying Coupled Two-Way Clustering (CTWC): an Example

In this section I present a step-by-step application of Coupled Two-Way Clustering (CTWC) using the CTWC server site at `http://ctwc.weizmann.ac.il` (See Publications (1) and (2)). This example can be used as a tutorial for using the site which includes all the fine details that can be controlled. If reader which is not interested in this example, the reader may proceed to Sec 2.2.6. The example is based on data published by Armstrong *et al.* [109] in which they analyzed acute leukemia of three different types; acute myeloid leukemia

(AML), acute lymphoblastic leukemia (ALL) and a sub-type of ALL which carries a chromosomal translocation in the MLL gene, located on chromosome 11[2]. An error in DNA replication in which one end of a chromosome $A$ replaces the correct one at the MLL gene's location ($B$ can stand for either a different or the same chromosome than 11), producing an abnormal protein either by fusion to another protein (if $B$ is not chromosome 11 - the alteration is called chromosomal translocation) or through internal rearrangements (if $B = 11$). Armstrong *et al.* show (in a supervised way) that this ALL sub-type has a distinct molecular profile and can be considered a new type of leukemia which they call MLL. For the sake of this example, assume that the MLL type had been unknown and the experiment was planned to study the molecular differences between AML and ALL, without knowing that the MLL sub-type exists. In this example, we will perform class discovery of this "new" sub-type and learn about its relation to the ALL and AML types.

**The preprocessed data**

The data consists of 57 samples taken from leukemia patients: 20 AMLs and 37 ALLs - which consist of 17 MLLs, ALLs with the translocation, and 20 without. The gene expression was obtained using Affymetrix GeneChip Hu95A which can measure $\sim$ 12600 different transcripts. The Affymetrix "average difference" (generated by MAS 4.0 software - see Section 3.3 for details regarding Affymetrix's chips) expression values were linearly scaled (by linear regression of all "Present" genes against the first ALL sample, see supplementary information of Armstrong *et al.* [109]) to correct for overall intensity differences between the experiments. The next step is filtering out genes with low quality and small variation. First, floor and ceiling values are set, by replacing all values below 100 (the level of detection of the technology) with 100 and values above 16000 with 16000 (to reduce the effect of saturation on the analysis). The resulting expression values are log-transformed using $log_2$ and the standard deviation across all samples is calculated for each gene. For the purpose of the example, I chose to work with the top 200 genes; in practice, one usually works with a few thousands of genes. Table 2.2 shows the expression matrix as inserted to the CTWC server.

For each sample, we have an external label indicating whether it is an AML or ALL. We also have a hidden label stating the MLL status. CTWC does not use any of these labels in the analysis, only at the end, to identify which of the resulting clusters correspond to known clinical labels. Table 2.1 presents a "labels" file which is uploaded to the server and used to analyze the results.

Running CTWC involves several steps: preprocessing, choosing a distance measure, setting a clustering method and its parameters and choosing the depth of iterations. All these steps are performed both for the genes and samples. Below I give a brief description of the options. After the analysis is finished one can scan through the results and identify potential discoveries. The next step is either to design further analyses or continue with

[2]The MLL (also called ALL-1) gene is directly involved in 5-10% of ALLs and AMLs

validation of the results.

| LABELS | NAME | ALL1 | ... | ALL20 | MLL1 | ... | MLL17 | AML1 | ... | AML23 |
|---|---|---|---|---|---|---|---|---|---|---|
| ALL | ALL (no MLL trans.) | 1 | | 1 | 0 | | 0 | 0 | | 0 |
| MLL | MLL (ALL w/MLL trans.) | 0 | | 0 | 1 | | 1 | 0 | | 0 |
| AML | AML | 0 | | 0 | 0 | | 0 | 1 | | 1 |
| ORIG ALL | Original ALL | 1 | | 1 | 1 | | 1 | 0 | | 0 |

Table 2.1: The sample labels table $\mathcal{L}^S$; $\mathcal{L}^S_{ij} = 1$ indicates that sample $j$ belongs to group $i$. This table is uploaded in a tab-delimited ASCII format to the server as predefined sets of samples. This information is used when viewing and analyzing the clustering results.

| U95_AFFX | NAME | ALL1 | ... | ALL20 | MLL1 | ... | MLL17 | AML1 | ... | AML23 |
|---|---|---|---|---|---|---|---|---|---|---|
| 31431_at | Hs.160741 Human IgG Fc receptor hFcRn | 12.4 | | 11.2 | 8.6 | | 13.1 | 9.8 | | 9.0 |
| 31506_at | Hs.274463 Human neutrophil peptide-3 gene | 8.3 | | 9.8 | 13.9 | | 9.7 | 9.7 | | 11.1 |
| 31525_s_at | Hs.251577 Human alpha globin: zeta gene | 6.6 | | 13.2 | 13.9 | | 8.5 | 10.2 | | 13.6 |
| 31623_f_at | Hs.203967 Human metallothionein-I-A gene | 6.6 | | 6.6 | 9.5 | | 6.6 | 6.6 | | 13.7 |
| 31687_f_at | Hs.155376 Human sickle cell beta-globin | 9.1 | | 12.8 | 13.9 | | 9.9 | 6.6 | | 11.0 |
| ⋮ | | | | | | | | | | ⋮ |
| 216_at | Hs.8272 Human prostaglandin D2 synthase | 6.6 | | 6.6 | 6.6 | | 11.5 | 9.3 | | 12.8 |

Table 2.2: The expression data table $X$. The data are scaled, thresholded and $\log_2$-transformed. $X_{ij}$ is the expression level of gene $i$ in sample $j$. The first column is the Affymetrix probe-set identifier. A description of the transcript, gene or EST, is in the second column. This table is uploaded in a tab-delimited ASCII format to the server when creating a new project. This is the data analyzed by the coupled two-way clustering method.

### Overview

The CTWC site enables users to register and open an account on the server. For each user the site manages projects, analyses and processes. A *Project* is related to a gene expression matrix $X_{ij}$ and optionally two tables of predefined sets of genes and samples, $\mathcal{L}^G$ and $\mathcal{L}^S$. Every project may contain several *Analyses*; each uses a particular set of running parameters. Within an analysis there are *Processes*; each is a CTWC run defined by its initial gene and sample sets and the desired iteration depth.

### Input Data

The input data are the three matrices as tab-delimited ASCII files. The format of the expression matrix is the same one used by Eisen's `Cluster` program to allow for maximal compatibility. See Tables 2.2 and 2.1 for examples of such files.

### Preprocessing

The CTWC site can perform several common preprocessing steps according to this order:

- Lower threshold of $\theta_{\text{low}}$: if $X_{ij} < \theta_{\text{low}}$, set its value to $X_{ij} = \theta_{\text{low}}$ (and leave unchanged otherwise). Using Affymetrix software with its default parameters a value below 30 (or below 100 for stricter filtering) is considered the level of detection of the system. The meaningless negative numbers produced by the older version of Affymetrix' MAS software (4.0) can also be handled using thresholding.

- Data *transformation*: taking the log, $X_{ij} \leftarrow \log_2(X_{ij})$. Used to transform the log-normal noise to be normally distributed and independent of the value of $X_{ij}$ (see **??** for other methods).

- Handling *missing values*:

  - Identify missing values as those below a threshold $\theta_{\text{m.v.}}$. The Affymetrix software provides an absent/present call that indicates if the transcript can be reliably declared present in the examined sample (the null hypothesis, that it is absent, is rejected). An unreliable measurement is declared absent. If one wants to use these calls, one can make sure all absent genes have negative numbers and then use some positive threshold to identify missing values.
  - Filter out genes with more than $x$ percent of missing values. These genes are either unreliable or not present in many samples and may only introduce noise. This step should be handled with care, since if one compares several types of diseases, the fact that some genes are consistently not present in some of the types can be important and these genes should not be filtered out.
  - Estimation of missing values can be performed in three different ways: $k$NN estimation [23] (see Sec. 1.5.3 and Sec. 2.2.6) which is based on weighted averages of correlated genes, replacement with a constant value or replacement with the gene's average over the other samples.
  - *Scaling* has also several options: substraction or division by the sample's mean or median. If the values are log-transformed then the algorithm subtracts the mean or median, otherwise it divides by them. This is equivalent to finding a scaling factor for each sample. Another option is to use non-linear scaling which uses the algorithm of Greenbaum *et al.* [110] that performs iterative non-linear fits between the samples. Section **??** discusses other methods for scaling. Recently, advanced scaling methods, *e.g.* RMA [13], MBEI (implemented in dChip) [16–18] and MAS 5.0 [15], were developed to perform better scaling of Affymetrix's chips. Consequently, we recommend one of these methods to be used for scaling.
  - *Variance filtering* – keeps the top $x$ genes with largest variation across the samples. This should be performed after the values are transformed.

  In this example, the data were already scaled and log-transformed. Therefore, only variance filtering was applied keeping the top 200 genes.

### Distance measure

Currently, the CTWC works with a specific choice of distance measures. Before any of the clustering operations S$I$(G$J$) or G$J$(S$I$) the sub-matrix that is defined by the G$J$ genes and the samples in S$I$ is centered and normalized. The distance between genes is taken as the Euclidean distances between the rows of the sub-matrix, and between the samples – the Euclidean distances between the columns.

### Clustering method and parameters

The CTWC procedure can utilize any clustering engine for which stability measure can be associated. We use super-paramagnetic clustering (SPC) [**?**, 74] due to various advantages it has (see 1.8.2). SPC has several parameters which can be controlled in the CTWC site; these parameters need to specified for clustering genes and samples:

- The parameter $K$ used for constructing the graph using the $K$ mutual neighborhood algorithm; points $i$ and $j$ are connected by an edge if $i$ is among the $K$ nearest neighbors of $j$ and $j$ is among the $K$ nearest neighbors of $i$. The $K$ mutual neighborhood algorithm gives rise to fairly uniform connectivity in regions where the data are uniformly distributed; in regions where the density of data points has significant gradients, the connectivity is lower [111]. The clustering result is affected by the value of $K$; low values of $K$ generate sparser graphs and tend to produce smaller clusters. Very sparse graphs in which there are only few paths that link any two points in the graph reduce the ability of SPC to average local fluctuations in the density. In the extreme case of a single connecting path, SPC reduces to single-linkage since the pairwise correlations $C_{ij}(T)$ depend only on their direct bond weight $J_{ij}$. Too high values of $K$ generate highly connected regions which usually "dissolve" abruptly and any inner structure can be concealed. As a result of that one might mistakenly declare the points to be uniformly distributed. In practice, values between 10 to 20 work well in most cases. In the example $K = 10$ was used both for genes and samples. Agrawal [112] analyzed the statistics of degrees of vertices in graphs produced for gene expression data. The homogeneity of the histogram of number of neighbors (degrees of vertices) is measured by $\Lambda(K)$ which is defined as

$$\Lambda(K) = \frac{1}{z_{\max} + 1} \left[ \frac{1 - P(z_{\max})}{2} + \sum_{i=0}^{z_{\max}} (i+1) P(i) \right] \tag{2.12}$$

  where $P(i)$ is the fraction of points with $i$ neighbors and $z_{\max}$ is the maximum degree in the network. $\Lambda(K)$ is between 0 and 1 and measures the ratio between the average "star" size (a point and its neighbors) to the maximal star size. Small values of $\Lambda(K)$ represent graphs in which there are only few highly connected points where the rest have only few neighbors, such graphs represent non-trivial structure of the data. On the other end, high values of $\Lambda(K)$ are found in graphs in which most stars are close

to the maximal one (occurs in uniform densities). In gene expression data $\Lambda(K)$ has a flat minimum for a range of $K$ values; $K_1$ to $K_2$. Agrawal suggests to test three different values for $K$; $K_1$, $\lfloor (K_1+K_2)/2 \rfloor$ and $K_2$ which correspond to graphs with low connectivity (referred to as paramagnetic adjacent), mid connectivity (referred to as super-paramagnetic) and high connectivity (referred to as ferromagnetic adjacent). The CTWC server implements this algorithm and uses $K_2$ as a default which can be changed to $K_1$ or $\lfloor (K_1 + K_2)/2 \rfloor$ or overridden by the user by a fixed value for $K$.

Note that on top of the $K$ mutual neighborhood graph the SPC superimposes the minimal spanning tree which makes sure that the graph is a single connected component. This guarantees that at $T = 0$ all the points will form one cluster.

- Temperature range and step: the default range of temperature is $[0, 0.25]$ scanned in steps of 0.004 yielding 62 temperature steps. The temperature values are measured in units of the cost and since the edge weights are defined as

$$J_{ij} = 1/\hat{K} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_i\|^2/(2a^2)\right) \tag{2.13}$$

  where $\hat{K}$ is the average number of neighbors and $a$ is the average distance between neighbors, this range is usually sufficient to capture the structure in the data. The example was generated using the default values.

- Number of Monte Carlo cycles: SPC uses a Swendsen-Wang (SW) Markov chain Monte Carlo (MCMC) method [74] to estimate the pairwise correlations $C_{ij}(T)$. This is an efficient Monte Carlo method which flips the states of a block of many spins in a single move. The default values of 3000 cycles is usually sufficient to estimate $C_{ij}(T)$ at a reasonable accuracy. For the example I used 5000 cycles for better accuracy and reproducibility. Since SPC uses an MCMC the results are not deterministic and can fluctuate. Increasing the number of cycles reduces the fluctuations and thus improves the reproducibility of the results.

- Growth option: As described in Sec. 1.8.2, a "growth" postprocessing step whose aim is to connect low density tails to the cluster's core is applied by default. In this step, each point is connected to its neighbor with which it is most correlated, as long as this correlation is above a very low threshold. The clusters generated by this step are not guaranteed to follow the dendrogram criterion (*i.e.* clusters that separate at low temperatures cannot re-unite at higher ones), and therefore points become connected by the growth step only if they were assigned to the same cluster at lower temperatures. Since this step searches for the maximal correlation among a group of low values, it is highly susceptible to fluctuations in the correlations and thus it contributes to the variation of the results obtained by different Monte Carlo runs on the same data. In the examples the growth option was used.

- Stability parameters - these parameters are used to determine if a cluster is stable. A cluster is considered stable if it is larger than a *minimal size* and does not lose

more than *ignore dropout size* members in each temperature step for at least *stable* $\Delta T$ steps. $T_1$ is defined as the temperature at which the cluster is separated from its environment and $T_2$ is the temperature at which it loses at least *ignore dropout size* of its members. The stability parameter $\Delta T$ is their difference, $\Delta T = T_2 - T_1$. For the example data, different parameters were used for clustering genes and samples; for genes – minimal cluster size is 10, ignore drop out size is 4 and stable $\Delta T = 8$; for samples – minimal cluster size and ignore drop out size are the same and stable $\Delta T = 15$. The relatively large $\Delta T$ for samples was chosen in order to obtain only the few most stable clusters.

### Depth of iterations

The depth of iterations performed by CTWC is controlled separately for genes and samples. Table 2.3 defines which operation are performed at each depth level. Currently, the server accepts depths up to 5. We recommend to start with two way clustering, *i.e.* depth= 1, for genes and samples. If the parameters gave suitable results, we proceed to deeper levels.

| Genes | | | Samples | | |
|---|---|---|---|---|---|
| Depth | Clustering Operation | Resulting clusters | Depth | Clustering Operation | Resulting clusters |
| 1 | G1(S1) | G2,. . .,$GI_{1,1}$ | 1 | S1(G1) | S2,. . .,$SJ_{1,1}$ |
| 2 | G1(S2),. . .,G1($SJ_{1,1}$) | G[$I_{1,1}+1$],. . .,$GI_{2,1}$ | 2 | S1(G2),. . .,S1($GI_{1,1}$) | S[$J_{1,1}+1$],. . .,$SJ_{2,1}$ |
| 3 | G2(S2),. . .,G2($SJ_{1,1}$) | G[$I_{2,1}+1$],. . .,$GI_{3,2}$ | 3 | S2(G2),. . .,S2($GI_{1,1}$) | S[$J_{2,1}+1$],. . .,$SJ_{3,2}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 3 | $GI_{1,1}$(S2),. . .,$GI_{1,1}$($SJ_{1,1}$) | G[$I_{3,I_{1,1}-1}+1$],. . .,$GI_{3,I_{1,1}}$ | 3 | $SJ_{1,1}$(G2),. . .,$SJ_{1,1}$($GI_{1,1}$) | S[$J_{3,J_{1,1}-1}+1$],. . .,$SJ_{3,J_{1,1}}$ |
| 4 | G1(S[$J_{1,1}+1$]),. . .,G1($SJ_{2,1}$) | G[$I_{3,I_{1,1}}+1$],. . .,$GI_{4,1}$ | 4 | S1(G[$I_{1,1}+1$]),. . .,S1($GI_{2,1}$) | S[$J_{3,J_{1,1}}+1$],. . .,$SJ_{4,1}$ |
| 5 | G2(S[$J_{1,1}+1$]),. . .,G2($SJ_{2,1}$) | G[$I_{4,1}+1$],. . .,$GI_{5,2}$ | 5 | S2(G[$I_{1,1}+1$]),. . .,S2($GI_{2,1}$) | S[$J_{4,1}+1$],. . .,$SJ_{5,2}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 5 | $GI_{1,1}$(S[$J_{1,1}+1$]),. . . . . .,$GI_{1,1}$($SJ_{2,1}$) | G[$I_{5,I_{1,1}-1}+1$],. . .,$GI_{5,I_{1,1}}$ | 5 | $SJ_{1,1}$(G[$I_{1,1}+1$]),. . ., . . .,$SJ_{2,1}$($GI_{2,1}$) | S[$J_{5,J_{1,1}-1}+1$],. . .,$SJ_{5,J_{1,1}}$ |

Table 2.3: The operations performed at each depth level and their resulting clusters.

### Initial sets

One has to choose a set of genes and samples that define the initial sub-matrix for CTWC. By default these are all the genes and all the samples. One may, however, want to analyze a specific sub-matrix. For example, if one has samples from several kinds of tumors and normal tissues, it is reasonable to analyze separately the sub-matrix that correspond to the tumor tissues. Most probably, CTWC will perform such a run by itself at deeper levels of the analysis since the tumor tissues usually form a separate and stable cluster when analyzing all the samples based on some tumor/normal separating gene cluster. Other cases in which analyzing a sub-matrix turned out to be convenient had to do with an interesting cluster of genes that was not picked up by CTWC as the feature set to use for clustering the samples, because of the arbitrarily set value of some parameter (such as

the iteration depth). One may wish to perform this step manually by clustering a set of samples based on these genes.

### Viewing the results

The results are stored in HTML files and can be viewed using any web browser. The main results page contains links to tables and to pages that describe the clustering operations and the resulting clusters. Here I include some of the results for the example dataset. In the first iteration of the analysis S1(G1) identifies two stable clusters S2 and S3. Figure 2.2 represents the *card* produced for the clustering operation S1(G1). The card contains several of the figures that are available in the site: (a) Rotated dendrogram – The dendrogram represents the clustering results. The $x$-axis is the temperature (the resolution parameter) and along the $y$-axis the clustered objects, the samples in this case are ordered according to the dendrogram. Only branches that contain more than *ignore drop out size* leaves are shown. Stable clusters are circled and named. (b) The labels of the samples are presented using a binary matrix (blue= 0, yellow= 1). (c) Reordered data – the colored matrix represent the transposed centered and normalized expression matrix, each row is a sample and each column is a gene. The samples are ordered according to the dendrogram on the left and the genes are ordered according to the dendrogram G1(S1). The colors in the matrix represent the value of the matrix elements. (d) Distance matrix – an $N_s \times N_s$ matrix of Euclidean distances between the samples, calculated based on all 200 genes in G1. One can see in the distance matrix three regions of short distances (high densities) represented by green-blue color. The algorithm partitioned the samples according to these three regions, as can be seen in the three main branches, but these were not identified as stable clusters since they did not "live" for the 15 temperature steps required by the selected parameters. The top branch broke up at a higher temperature to two sub-branches, the top of which was large enough and did not break for a large enough $\Delta T$ to be identified as stable, and was labeled S2. The middle main branch qualified as stable (according to the used parameters) and was named S3, while the bottom one did not "live" for the required $\Delta T$ and then broke to clusters smaller than *min cluster size*. Cluster S3 contains 17 samples; 16 are of type MLL and 1 AML. This stable cluster splits the ALL samples to ones with the translocation (MLLs) and one without (other ALLs). Already this operation revealed the distinct molecular profile of the MLL subtype. In general, had there been many other genes which are not related to this separation one would not see this partition when clustering S1(G1), but this example is based only on 200 genes of which a large enough fraction are differentially expressed between the ALL and MLL types.

Figure 2.3 depicts the card generated for clustering the 200 genes based on all the 57 samples, G1(S1). One can see that there are two main branches in dendrogram each of which splits into two other long branches. Looking at the distance matrix one can clearly see two large blue squares representing the main split. Both squares have inner structure which can be understood if one looks at the reordered data 2.2b. The upper branch splits into two; genes in cluster G2 (around 20) are highly expressed on the ALLs (leftmost 20

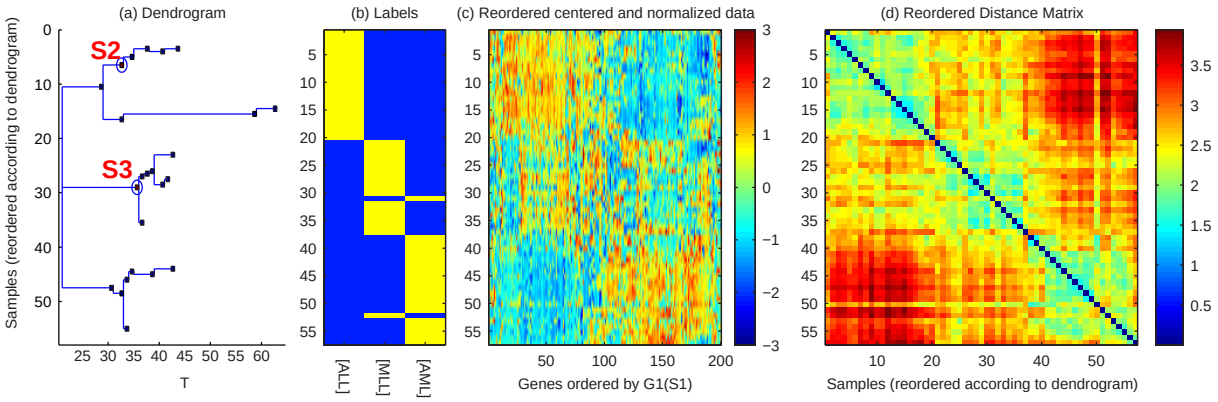



Figure 2.2: The CTWC card for S1(G1).

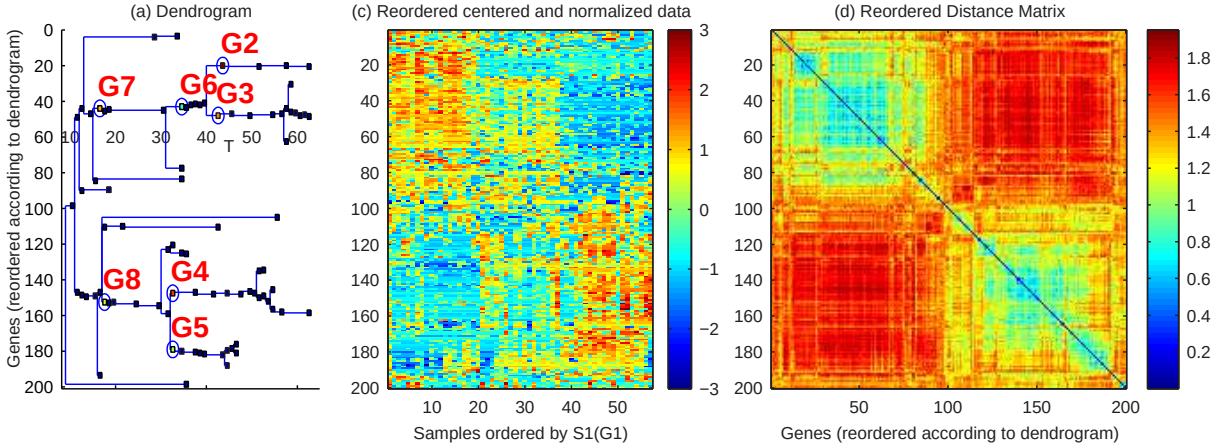



Figure 2.3: The CTWC card for G1(S1).

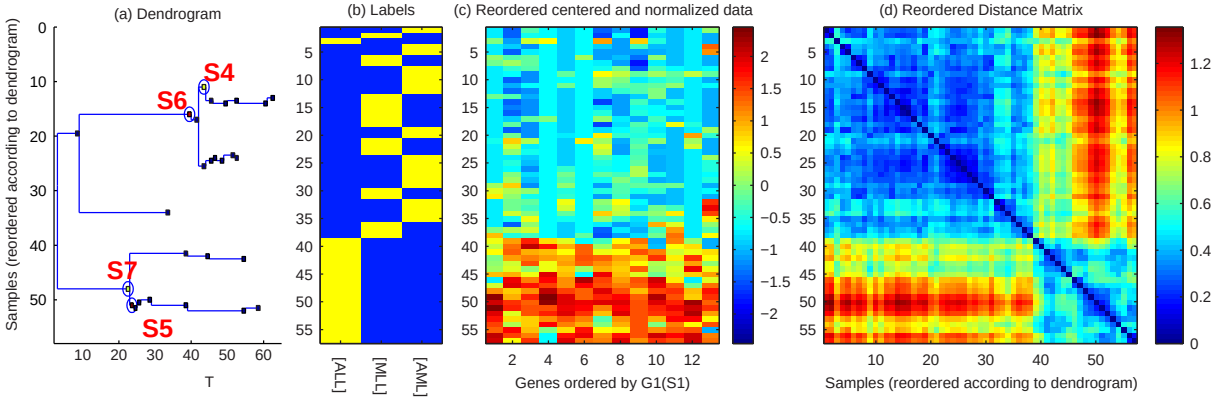



Figure 2.4: The CTWC card for S1(G2).

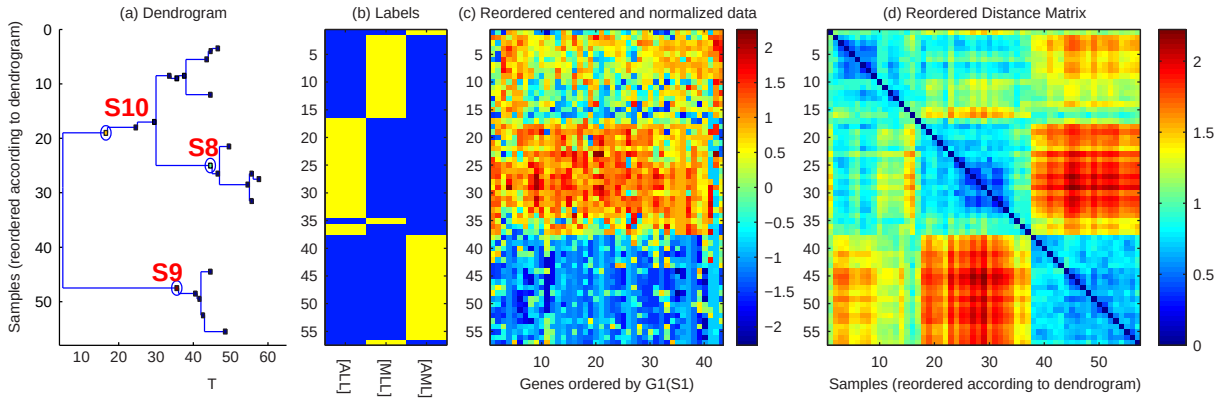



Figure 2.5: The CTWC card for S1(G3).

samples) and low on the remaining samples whereas the genes in cluster G3 (between 30 and 70) are high on the ALLs and intermediate on the MLLs and low on the AMLs.

The result of clustering the samples based on G2 and G3 are seen in Figures 2.4 and 2.5. One can see that G2 separates the samples into S6 and S7; S6 contains the MLLs and AMLs, and S7 contains the ALL samples. This means that regarding the 13 genes in G2, the MLLs behave as AMLs, although the former are actually ALLs with a specific translocation. In each clustering operation the genes and samples are listed, either as clustered objects or as features. A link can transfer the gene list to Affymetrix' web site, `http://www.netaffx.com`, which supplies additional information on each gene, such as its symbol, description, genomic location, gene ontology (biological process, cellular component and molecular function) and if it belongs to a known pathway (using GenMAPP database [94]). Table 2.4 gives an example table produced by NetAffx for the G2 cluster. One of the genes in G3 is CALLA, common acute lymphoblastic leukemia antigen, also known as CD10, is a known marker of ALLs with no translocation since it is expressed at late differentiation stages of the lymphoid lineage. The MLL translocation is believed to occur at early stages of the differentiation and therefore cancer cells with the translocation become malignant before they reach the stage in which the CALLA gene is expressed. In Publication (9) [113] we show that there might be a new sub-type of MLL that has some characteristics of cells in later stages of differentiation. Note that two genes (MADH1 and RAF1) have two different probes in the cluster showing the relatively high correlation of probes of the same gene.

Figure 2.5 shows the results of clustering S1(G3). In contrast to G2, when G3 is used, the ALLs join the MLLs and the AMLs form a separate cluster. The first and major split is to S9 (AMLs) and S10 (ALLs and MLLs) – these genes have very low expression in AMLs compared to the ALLs and MLLs. At higher temperatures S10 splits to two, and one of its branches is steady enough to be called stable and is named S8 (ALLs). Actually there are three levels of expression of this gene-cluster; ALLs have high expression of these genes, MLLs have medium levels and very low in AMLs. G3 contains 43 genes which appear in Table 2.5. Several of these are related to the immune system (marked by boldface in the

table); immunoglobulins, major histocompatibility factors, immune response. This is not surprising since the the ALLs are related to T-cells and B-cells which are the two mature cells in the lymphoid lineage. Another gene is the TCL1A which is known to be activated in chronic T-cell leukemia.

In this example we could identify in an unsupervised fashion the existence of an "intermediate" type of leukemia, the MLL; For the genes in G2 its molecular profile is similar to that of the AMLs whereas according to the "signal" of the G3 genes the MLLs are associated with the ALLs (although at somewhat lower level of expression).

| Probe Set ID | Title | Gene Symbol | Map Location | GO bio process | GO cell component | GO molec function | GenMAPP Pathway |
|---|---|---|---|---|---|---|---|
| 1077_at | recombination activating gene 1 | RAG1 | 11p13 | GO:7516;hemocyte development GO:6310;DNA recombination GO:6955;immune response | GO:5634;nucleus | GO:16787;hydrolase activity GO:3677;DNA binding GO:4519; | |
| 1077_at | recombination activating gene 1 | RAG1 | 11p13 | GO:7516;hemocyte development GO:6310;DNA recombination GO:6955;immune response | GO:5634;nucleus | GO:16787;hydrolase activity GO:3677;DNA binding GO:4519;endonuclease activity | |
| 1325_at | MAD,mothers against decapentaplegic homolog 1 (Drosophila) | MADH1 | 4q28 | GO:6355;regulation of transcription, DNA-dependent GO:7179;TGFbeta receptor signaling pathway GO:7165;signal transduction | GO:16021;integral to membrane GO:5634;nucleus | GO:16563; transcriptional activator activity GO:5057; receptor signaling protein activity | TGF Beta Signaling Pathway |
| 1389_at | **membrane metallo-endopeptidase, enkephalinase, CALLA, CD10** | MME | 3q25.1-q25.2 | GO:7267;cell-cell signaling GO:6508;proteolysis and peptidolysis | GO:5887;integral to plasma membrane | GO:8270;zinc ion binding GO:16787;hydrolase activity GO:4245;neprilysin activity GO:8237; metallopeptidase activity | |
| 32872_at | | | | | | | |
| 35164_at | Wolfram syndrome 1 (wolframin) | WFS1 | 4p16 | GO:6091;energy pathways GO:7605;hearing GO:7399;neurogenesis GO:7601;vision | GO:5624;membrane fraction GO:16021;integral to membrane | | |
| 36536_at | schwannomin interacting protein 1 | SCHIP1 | 3q25.33 | | GO:5737;cytoplasm | | |
| 37280_at | MAD, mothers against decapentaplegic homolog 1 (Drosophila) | MADH1 | 4q28 | GO:6355;regulation of transcription, DNA-dependent GO:7179;TGFbeta receptor signaling pathway GO:7165;signal transduction | GO:16021;integral to membrane GO:5634;nucleus | GO:16563; transcriptional activator activity GO:5057;receptor signaling protein activity | TGF Beta Signaling Pathway |
| 37539_at | RalGDS-like gene | RGL | 1q25.2 | GO:7264;small GTPase mediated signal transduction GO:7218;neuropeptide signaling pathway | | GO:8321;Ral guanyl-nucleotide exchange factor activity GO:5085; RasGEFN; guanyl-nucleotide exchange factor activity | |
| 38124_at | midkine (neurite growth-promoting factor 2) | MDK | 11p11.2 | GO:8283;cell proliferation GO:7267;cell-cell signaling GO:7165;signal transduction GO:30154;cell differentiation GO:74;regulation of cell cycle GO:7399;neurogenesis | GO:5615; extracellular space | GO:8201;heparin binding GO:8083;growth factor activity GO:5125;cytokine activity | |
| 38408_at | transmembrane 4 superfamily member 2 | TM4SF2 | Xq11.4 | GO:9405; pathogenesis GO:6487;N-linked glycosylation | GO:5887;integral to plasma membrane | | |
| 38578_at | tumor necrosis factor receptor superfamily,member 7 | TNFRSF7 | 12p13 | GO:6915;apoptosis GO:6955;immune response GO:7165;signal transduction | GO:5886;plasma membrane GO:16021;integral to membrane | GO:4872; TNFR_c6; receptor activity; GO:4888; transmembrane receptor activity | |
| 41266_at | integrin,alpha 6 | ITGA6 | 2q31.1 | GO:7044;cell-substrate junction assembly GO:7229;integrin-mediated signaling pathway GO:7160;cell-matrix adhesion | GO:8305;integrin complex GO:16021;integral to membrane | GO:4895; cell adhesion receptor activity GO:4872; receptor activity | |
| 41690_at | modulator recognition factor 2 | MRF2 | 10q21.3 | | GO:5622; intracellular | GO:3677;DNA binding | |

Table 2.4: The annotation table for the genes in G2 from the NetAffx website

| Probe Set ID | Title | Gene Symbol | GO bio process |
|---|---|---|---|
| 1470_at | polymerase (DNA directed), delta 2, regulatory subunit 50kDa | POLD2 | GO:6260;DNA replication |
| 266_s_at | CD24 antigen (small cell lung carcinoma cluster 4 antigen) | CD24 | GO:6959;humoral immune response |
| 32168_s_at | Down syndrome critical region gene 1 | DSCR1 | GO:19722;calcium-mediated signaling GO:7165;signal transduction GO:8015;circulation GO:7417;central nervous system development |
| 32238_at | bridging integrator 1 | BIN1 | GO:7268;synaptic transmission GO:6897;BAR;endocytosis;5.5e-130;extended:inferred from electronic annotation GO:6899;nonselective vesicle transport GO:6897;BAR;endocytosis;4.9e-127;extended:inferred from electronic annotation GO:30154;cell differentiation GO:8283;cell proliferation GO:8099;synaptic vesicle endocytosis GO:45786;negative regulation of cell cycle |
| 32773_at | major **histocompatibility** complex, class II, DQ alpha 1 | HLA-DQA1 | GO:9405;pathogenesis GO:19884;antigen presentation, exogenous antigen GO:6955;immune response GO:19886;antigen processing, exogenous antigen via **MHC** class II |
| 33304_at | **interferon** stimulated gene 20kDa | ISG20 | GO:8283;cell proliferation |
| 33999_f_at | hypothetical protein LOC90925 | LOC90925 | |
| 34168_at | deoxynucleotidyltransferase, terminal | DNTT | GO:6260;DNA replication GO:6304;DNA modification GO:6281;DNA repair GO:6960;antimicrobial humoral response (sensu Invertebrata) |
| 34800_at | leucine-rich repeats and **immunoglobulin**-like domains 1 | LRIG1 | |
| 34842_at | small nuclear ribonucleoprotein polypeptide N | SNRPN | GO:6371;mRNA splicing;predicted/computed |
| 35260_at | Mlx interactor | MONDOA | |
| 35350_at | B cell RAG associated protein | GALNA C4S-6ST | |
| 35614_at | transcription factor-like 5 (basic helix-loop-helix) | TCFL5 | GO:42127;regulation of cell proliferation;inferred from expression pattern GO:6355;regulation of transcription, DNA-dependent;inferred from sequence or structural similarity GO:7283;spermatogenesis;inferred from expression pattern GO:45595;regulation of cell differentiation;inferred from expression pattern |
| 35648_at | autism susceptibility candidate 2 | AUTS2 | |
| 36021_at | lymphoid enhancer-binding factor 1 | LEF1 | GO:6355;regulation of transcription, DNA-dependent |
| 36108_at | major **histocompatibility** complex, class II, DQ beta 1 | HLA-DQB1 | GO:6955;**MHC**_II_beta;immune response;1.6e-52;extended:inferred from electronic annotation |
| 36227_at | interleukin 7 receptor | IL7R | GO:18;regulation of DNA recombination GO:6960;antimicrobial humoral response (sensu Invertebrata) GO:7166;cell surface receptor linked signal transduction GO:6955;**immune response** |
| 36239_at | POU domain, class 2, associating factor 1 | POU2AF1 | GO:6355;regulation of transcription, DNA-dependent GO:6959;humoral **immune response** GO:8151;cell growth and/or maintenance GO:6366;transcription from Pol II promoter |
| 36482_s_at | ATPase, Ca++ transporting, ubiquitous | ATP2A3 | GO:6812;cation transport GO:8152;metabolism GO:15992;proton transport GO:6816;calcium ion transport GO:6810;transport |
| 36638_at | connective tissue growth factor | CTGF | GO:1558;regulation of cell growth GO:6928;cell motility;not recorded GO:7155;cell adhesion GO:6259;DNA metabolism GO:9611;response to wounding GO:8544;epidermal differentiation GO:8151;cell growth and/or maintenance |
| 37159_at | hypothetical protein DJ159A19.3 | DJ159 A19.3 | |
| 37251_s_at | glycoprotein M6B | GPM6B | GO:7399;neurogenesis |
| 37467_at | Homo sapiens partial mRNA for IgA1 **immunoglobulin** heavy chain variable region (IGHV gene), clone LIBPA235 | | |
| 37710_at | MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C) | MEF2C | GO:6355;regulation of transcription, DNA-dependent GO:7517;muscle development GO:7399;neurogenesis GO:6366;transcription from Pol II promoter;not recorded |
| 37716_at | antigen identified by monoclonal antibody MRC OX-2 | MOX2 | |
| 37988_at | CD79B antigen (**immunoglobulin**-associated beta) | CD79B | GO:7166;cell surface receptor linked signal transduction GO:6955;**immune response** |
| 38017_at | CD79A antigen (**immunoglobulin**-associated alpha) | CD79A | GO:6952;defense response GO:7166;cell surface receptor linked signal transduction |
| 38018_g_at | CD79A antigen (**immunoglobulin**-associated alpha) | CD79A | GO:6952;defense response GO:7166;cell surface receptor linked signal transduction |
| 38242_at | B-cell linker | BLNK | GO:7242;intracellular signaling cascade GO:6954;inflammatory response;predicted/computed GO:6959;humoral **immune response**;predicted/computed GO:7516;hemocyte development |
| 38604_at | neuropeptide Y | NPY | GO:6928;cell motility GO:7586;digestion;not recorded GO:7273;regulation of synapse GO:6816;calcium ion transport GO:7187;G-protein signaling, coupled to cyclic nucleotide second messenger GO:8283;cell proliferation GO:8015;circulation;not recorded GO:7631;feeding behavior GO:7218;neuropeptide signaling pathway |
| 38994_at | suppressor of cytokine signaling 2 | SOCS2 | GO:1558;regulation of cell growth GO:7242;intracellular signaling cascade |
| 39318_at | **T-cell leukemia/lymphoma 1A** | **TCL1A** | GO:7275;development GO:8151;cell growth and/or maintenance |
| 39878_at | Homo sapiens transcribed sequence with strong similarity to protein ref:NP_065136.1 (H.sapiens) protocadherin 9 precursor; cadherin superfamily protein VR4-11 [Homo sapiens] | | |
| 40191_s_at | KIAA0582 protein | KIAA0582 | |
| 40451_at | polymerase (DNA directed), epsilon | POLE | GO:6281;DNA repair GO:6260;DNA replication |
| 40782_at | short-chain dehydrogenase/reductase 1 | SDR1 | GO:8152;metabolism GO:6631;fatty acid metabolism GO:7601;vision |
| 40936_at | cysteine-rich motor neuron 1 | CRIM1 | GO:7399;neurogenesis GO:1558;regulation of cell growth GO:6508;proteolysis and peptidolysis |
| 41166_at | **immunoglobulin** heavy constant mu | IGHM | |
| 41401_at | cysteine and glycine-rich protein 2 | CSRP2 | GO:16049;cell growth GO:7517;muscle development GO:8283;cell proliferation GO:30154;cell differentiation |
| 41442_at | core-binding factor, runt domain, alpha subunit 2; translocated to, 3 | CBFA2T3 | GO:8283;cell proliferation |
| 41470_at | prominin 1 | PROM1 | GO:7601;vision |
| 41503_at | transcription factor ZHX2 | ZHX2 | GO:6355;regulation of transcription, DNA-dependent |
| 914_g_at | v-ets erythroblastosis virus E26 oncogene like (avian) | ERG | |

### 2.2.6 Other biclustering methods

Coupled Two-way clustering belongs to a family of methods called *biclustering* since they cluster the rows and columns of a matrix simultaneously (the term is attributed to Mirkin [114]). Biclustering methods are dedicated to analyze data which has a block structure, *i.e.* the values in submatrices defined by a subset of the rows and a subset of the columns can be explained using simple relations. In the context of gene expression, groups of genes may belong to cellular processes or pathways which are activated at various degrees in a subset of the samples; these genes and the samples which are affected by a single process form a *bicluster*. Biclustering methods are used in other contexts as well and are common in text-mining [115–118] in which the rows represent words or terms and the columns are the analyzed documents. Focusing only on words that are related to, say, sports can help distinguish between sport-related articles and others.

At first, analysis of large scale gene expression data used only *two-way clustering* [8, 62, 71]. In two-way clustering two independent clustering operations are performed; one for the genes and one for the samples. The linear orderings inferred from the resulting dendrograms are used to reorder the rows and columns of the expression matrix. Blocks of similar expression patters can be identified by visual inspection of the reordered matrix (see Fig. 1.2). This type of analysis is still very common in the gene expression literature. The main drawback of this approach is that it takes a holistic view of the data and blocks, or biclusters, that are present in the data which are either small and thus overwhelmed by the noise or overlap with other blocks can not be identified. This is the motivation for developing CTWC and other biclustering methods.

Biclustering methods that divide a matrix into submatrices with approximately constant values were suggested as early as the 60s and 70s by Morgan and Sonquist [119] and Hartigan [120]. Hartigan's method iteratively partitions the matrix in a greedy way, either by splitting the rows or columns of a submatrix in order to achieve a minimal sum of square deviation in the newly generated submatrices. The stopping criterion is based a $\chi^2$ test.

Two classical methods for simultaneous analysis of rows and columns are singular value decomposition (SVD) (see 2.2.6) and two-way analysis of variance (ANOVA) (see 2.2.6). SVD is very popular in text-mining an information retrieval [121], and was applied to gene expression by Alter *et al.* [55]. Two-way ANOVA was first applied to gene expression by Kerr *et al.* [12]. Many biclustering methods have common features with these algorithms [122]. Therefore, it is convenient to first study SVD and two-way ANOVA (see Sec. 2.2.6) and then compare to other biclustering methods.

An alternative approach to formulate the problem uses the framework of a weighted *bipartite* graph[3] $G = (U, V, E)$; each gene is a vertex in $U$ and each samples is a vertex in $V$. The weight of the edge that connects gene $i$ ($i \in U$) and samples $j$ ($j \in V$) represents the expression level $X_{ij}$. Biclusters are bi-cliques in the graph, *i.e.* a subset of $U$ and a subset

[3]A bipartite graphs $G = (U, V, E)$ has two sets of vertices $U$ and $V$ and edges can connect between any vertex in $U$ and any vertex in $V$; $e = (u, v) \in E$ where $u \in U$ and $v \in V$.

of $V$ and all the edges between them. A cost functions is defined over the bi-cliques and the task of biclustering is to identify maximal bi-cliques whose cost is below some value. Obviously this can be achieved by searching all possible biclusters but this is practically impossible due to the exponential number of subsets of $U$ and $V$. The problem is, in general, NP-hard [123]; thus, one turns to heuristic or approximation methods that usually find only local minima of the cost function.

Cheng and Church [123] define the cost of a bicluster according to its mean squared residue, where the residue of a matrix element is defined as in ANOVA (see Sec. 2.2.6). They use a greedy approach to identify the biclusters by removing and adding sets of genes and samples to the bicluster (see Sec. 2.2.6). Tanay *et al.* introduced a method called SAMBA (Statistical Algorithmic Method for Bicluster Analysis); they developed a probabilistic approach in which each edge in the graph is assigned a weight according to its probability. Next, a hashing algorithm is used to find the $k$ bi-cliques (biclusters) with highest weights ignoring genes with more than $D$ edges. Then, a local improvement of the weight is performed by adding and removing single genes or samples (as in Church *et al.* [123]). Finally, the biclusters are reported, one-by-one, skipping ones which highly overlap with previously reported ones.

Ihmels *et al.* suggested a method called *Signature Algorithm* which starts with a "seed" of genes and identifies the samples in which they are regulated. These samples are then used to identify other genes which are related to the main pattern of the seed genes. Later, Bergman *et al.* extended the algorithm and studied its convergence properties. They define a bicluster as a *transcriptional module* (see Sections 2.2.6 and 2.2.6).

Lazzeroni and Owen introduced *plaid* models [122] which explain the gene expression data by a sum of two-way ANOVA models applied to submatrices. The parameters of the models are found using an EM-like method (see Sec. 2.2.6). Califano *et al.* [124] use an algorithm called SPLASH to identify statistically significant biclusters. They use a set of control experiments to define a distance measure between genes that is based on replacing a gene's expression level with its percentile in the population of the control experiments. Gene Shaving by Hastie *et al.* [125] identifies overlapping sets of genes by iteratively "pealing" off sets of genes which are highly correlated with the main variation component in the data (first principal component of the genes). This method can generate overlapping sets of correlated genes. Ben-Dor *et al.* [126] search for order-preserving submatrices (OPSMs) in which the expression levels of the genes induce the same linear ordering over the samples. They show that finding these submatrices is NP-hard. They suggested a probabilistic model that represents a hidden OPSM within a random matrix and propose a greedy heuristic algorithm that "grows" partial models in order to identify these submatrices. Dhillon *et al.* [127] use an information-theoretic approach which is related to the information bottleneck approach of Tishby *et al.* [128]. They cluster the rows and columns of a matrix such as to preserve the maximal possible mutual information between the rows and columns of the matrix. They applied their algorithm to clustering co-occurrence or contingency tables (such as in text-mining problems). Busygin *et al.* [129] describe an algorithm that performs

any centroid- (or representative-) based method simultaneously on both the genes and the samples. A centroid in the gene space has a conjugate centroid in the sample space which can be found by a linear transformation. Training is performed by alternating the original centroid-based training step between the spaces until is converges.

Below I give an brief summary of some selected methods for biclustering; singular value decomposition (SVD), two-way analysis of variance (two-way ANOVA), Cheng and Church's biclustering method [123], the signature algorithm [130] and its iterated version – the iterated signature algorithm (ISA) [131], Plaid models [122] and double conjugate clustering [129] Finally I comment on differences between these methods and coupled two-way clustering.

### Singular Value Decomposition (SVD)

Alter *et al.* [55] used *singular value decomposition* (SVD) [132] to analyze yeast cell cycle data. Given an $N_g \times N_s$ matrix $X$, assume $N_s \leq N_g$ (which is normally the case in gene expression data), the singular value decomposition of $X$ is defined as

$$X_{N_g \times N_s} = U_{N_g \times r} \Sigma_{r \times r} V^T_{r \times N_s} \tag{2.14}$$

where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ are orthonormal in the sample and gene spaces respectively; $U^T U = V^T V = I$, and $\Sigma = \mathrm{d}iag(\sigma_1, \ldots, \sigma_r)$. $r$ is the rank of $X$ which in general, in noisy data (*e.g.* gene expression), attains its maximal value $N_s$ (below I assume $r = N_s$). Alter *et al.* called the columns of $V$ eigengenes and the columns of $U$ eigenarrays. SVD is obtained using the following steps: (i) Find the eigenvectors (eigengenes), $\{\mathbf{v}_1, \ldots, \mathbf{v}_{N_s}\}$, and corresponding eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_{N_s}$ of the $N_s \times N_s$ matrix $A = X^T X$. (ii) The eigenarrays are constructed by $\tilde{\mathbf{u}}_i = X\mathbf{v}_i, \mathbf{u}_i = \tilde{\mathbf{u}}_i / \|\tilde{\mathbf{u}}_i\| \ \forall i = 1, \ldots, N_s$. (iii) $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, N_s$. The SVD is a unique decomposition up to a sign symmetry; one can negate any pair of $\mathbf{v}_i$ and $\mathbf{u}_i$. Note that the $r$ eigenarrays, $\{\mathbf{u}_1, \ldots, \mathbf{u}_{N_s}\}$, are the eigenvectors of $XX^T$ since $XX^T = U\Sigma V^T \left(U\Sigma V^T\right)^T = U\Sigma^2 U^T$. This implies that if $X$ is centered then $XX^T$ is the covariance matrix of the samples and, therefore, the eigenarrays are the principal components (PCA) of the sample vectors. A $k$-partial sum of the SVD matrices,

$$X^k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{2.15}$$

which has a rank of $k < N_s$, is the best approximation of $X$ by a lower $k$-ranked matrix. A proof given in [132] shows that

$$\min_{Y : rank(Y) \leq k} \|X - Y\|_2 = \|X - X^k\|_2 = \sigma_{k+1} \ . \tag{2.16}$$

This is the basis for the interpretation that the first SVD decompositions describe the major effects in the data and the remaining ones are due to noise.

One method to find the the eigenvectors of the matrix $A = X^T X$ is to start with a random normalized vector $\hat{\mathbf{g}}^{(0)}$ and iteratively apply $A$ to $\hat{\mathbf{g}}^{(t)}$; $\mathbf{g}^{(t+1)} \leftarrow A\hat{\mathbf{g}}^{(t)}$ and $\hat{\mathbf{g}}^{(t+1)} = \mathbf{g}^{(t+1)}/\|\mathbf{g}^{(t+1)}\|$. This mapping converges to an eigenvector of $A$ and if the initial vector $\hat{\mathbf{g}}^{(0)}$ is not perpendicular to the eigenvector that corresponds to the largest eigenvalue, $\mathbf{v}_1$, it will converge to it. Then, one can find the matching singular vector, $\mathbf{u}_1$, as described above and subtract $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$ from $X$. Repeating this procedure continuously removes the eigen-pair with the largest eigenvalue and generates the sequence of singular vectors. A related method is described in the Iterated Signature Algorithm (see 2.2.6 below).

Alter *et al.* used the SVD to perform an approximate two-way centering and normalization. At first the first SVD component was subtracted from the data, which corresponds to subtracting the mean of the rows and the columns; $Y = X - \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T$. Next, instead of normalizing (dividing by the square root of sum of squares) two steps were performed: (i) the *log* of the element-wise square of $Y$ was taken; $\hat{Y}_{ij} = \log(Y_{ij}^2)$ (note that $Y$ contains negative values hence the square inside the log), and (ii) the first SVD component of $\hat{Y}$ was removed; $\hat{Z} = \hat{Y} - \hat{\sigma}_1 \hat{\mathbf{u}}_1 \hat{\mathbf{v}}_1^T$ where $\hat{Y} = \hat{U}\hat{\Sigma}\hat{V}^T$. Subtracting the first SVD component of $\hat{Y}$ corresponds to dividing the rows and columns by the geometric mean of the squared deviations. Finally, the "centered" and "normalized" version of $X$ was calculated by back-transforming $Z = \sqrt{\exp(\hat{Z})}$.

After centering and normalizing using the SVD, Alter *et al.* noticed that the first two SVD components of the resulting matrix are of similar significance, *i.e.* their eigenvalues are close, and contain more than 40% of the overall sum of squares in the expression matrix and then, in the third component, there is a drop to $\sim 8\%$. These two components correspond to $\sin(x)$ and $cos(x)$ at exactly the frequency of the cell cycle of yeast. Next, they sorted the genes according to their phase in the cell-cycle (calculated based on the dot-product with the first two components), which grouped together genes which are up-regulated at the same stage of the cell-cycle.

Troyanskaya *et al.* [23] use SVD to impute missing values in a gene expression matrix. Using the fact that taking a partial sum of the SVD components can be used to approximate the matrix, they first replace each missing value by its rows (gene) mean since SVD can be applied only to full matrices. Next, the $k$ components of the SVD (they found that taking $k = 0.2N_s$ components) are used to approximate the matrix and the elements at the position of the missing values were used as their estimates.

### Two-way analysis of variance (two-way ANOVA)

The Two-way analysis of variance (ANOVA) is a hypothesis testing method that analyzes the effect of two factors, *i.e.* independent variables, concurrently. The analysis can infer whether there is an interaction between the effects of the two variables, $A$ and $B$. Kerr *et al.* [12] introduced higher order ANOVA to gene expression data analysis to analyze the effect of the different cDNA microarrays, the different dye (red of green), the condition and the gene. The data are organized in a two-dimensional table – a dimension for each

factor which is divided according to its possible values; $N_A$ rows and $N_B$ columns. In table element $(i,j)$ are all the observations, $\{y_{ijk}\}_{k=1}^{n}$, which have the same value for the two factors[4].

It is convenient to use the following notation for means of different populations:
$\mu_{ij\cdot}$ the mean of cell $(i,j)$, $\mu_{i\cdot\cdot} = \langle \mu_{ij\cdot} \rangle_j$ mean of row $i$, $\mu_{\cdot j\cdot} = \langle \mu_{ij\cdot} \rangle_i$ mean of column $j$, $\mu = \langle \mu_{ij\cdot} \rangle_{i,j}$ the overall mean, $\alpha_i = \mu_{i\cdot\cdot} - \mu$ is the effect of factor $A$ taking the value corresponding to row $i$, $\beta_j = \mu_{\cdot j\cdot} - \mu$ is the effect due to factor $B$ in column $j$, $(\alpha\beta)_{ij} = \mu_{ij\cdot} - \mu_{i\cdot\cdot} - \mu_{\cdot j\cdot} + \mu$ is the effect of the interaction between factor $A$ taking the $i$-th level and factor $B$ taking the $j$-th level, and $\epsilon_{ijk} = y_{ijk} - \mu_{ij\cdot}$ is the residual or random deviations in each particular observation due to noise or other unexplained sources.

Two-way ANOVA uses the following linear model to explain the data,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \ . \tag{2.17}$$

Three null hypotheses are tested, each generating its own $p$-value: (i) The two factors do not interact with respect to their effect on $y$; $(\alpha\beta)_{ij} = 0 \ \forall i,j$. (ii) factor $A$ has no effect on $y$ and all row means are equal; $\alpha_1 = \ldots = \alpha_{N_A} = 0$ and (iii) factor $B$ has no effect on $y$ and all column means are equal; $\beta_1 = \ldots = \beta_{N_B} = 0$. These hypotheses are tested under the assumptions that the observations in each table element are independent random samples, each of size $n$ from $N_A \times N_B$ populations which are normally distributed with mean $\mu_{ij\cdot}$ and variance of $\sigma^2$; $\{y_{ijk}\}_{k=1}^{n} \sim \mathcal{N}(\mu_{ij\cdot}, \sigma^2)$.

Denote by $\overline{y}_{\cdots}$, $\overline{y}_{i\cdot\cdot}$, $\overline{y}_{\cdot j\cdot}$ and $\overline{y}_{ij\cdot}$ the unbiased estimates of $\mu_{\cdots}$, $\mu_{i\cdot\cdot}$, $\mu_{\cdot j\cdot}$ and $\mu_{ij\cdot}$. The statistics used to test the above hypotheses are based on the sum of squares, $SS$, of different populations:
$SS_{Tot} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \sum_{k=1}^{n} \left(y_{ijk} - \overline{y}_{\cdots}\right)^2$ the total variability of the data,
$SS_A = N_B n \sum_{i=1}^{N_A} \left(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots}\right)^2$ the variability due to the different values of $A$,
$SS_B = N_A n \sum_{j=1}^{N_B} \left(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots}\right)^2$ the variability due to the different values of $B$,
$SS_{AB} = n \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots}\right)^2$ the variability due to the interaction of $A$ and $B$, and finally, $SS_\epsilon = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \sum_{k=1}^{n} \left(y_{ijk} - \overline{y}_{ij\cdot}\right)^2$ the variability due to random noise or unexplained sources. A simple relation holds between these statistics,

$$SS_{Tot} = SS_A + SS_B + SS_{AB} + SS_\epsilon \ . \tag{2.18}$$

Since the $y_{ijk}$ are assumed to be normally distributed these $SS$ statistics are $\chi^2_\gamma$ distributed with number of degrees of freedom, $\gamma$, according to the number of independent terms in the sum. The null hypotheses are tested using the $F$ distribution which is the distribution of the ratio between two $\chi^2$ variables;

$$f = \frac{\chi^2_{\gamma_1}/\gamma_1}{\chi^2_{\gamma_1}/\gamma_2} \ . \tag{2.19}$$

[4]For simplification I assume that the cells have the same number of observations $n$.

$f$ follow an $F_{\gamma_1,\gamma_2}$ distribution. For each hypothesis the $p$-value is calculated by testing the ratio between the corresponding effect $X$ and the error

$$P\left(f \geq \frac{SS_X/\gamma_X}{SS_\epsilon/\gamma_\epsilon}\right) = \int_{\frac{SS_X/\gamma_X}{SS_\epsilon/\gamma_\epsilon}}^{\infty} F_{\gamma_X,\gamma_\epsilon}(x)dx \ . \tag{2.20}$$

For example, to test whether there is an interaction (hypothesis (i) above) one tests if $SS_{AB}/\left((N_A-1)(N_B-1)\right)\left[SS_\epsilon/\left(N_A N_B(n-1)\right)\right]^{-1}$ is large enough to reject the null hypothesis.

Two-way ANOVA is a supervised method and cannot be considered as a biclustering method. But, the model it uses to explain the values, by additive contributions of the rows and columns, is the basis of various scores of biclusters. Note that ANOVA is applied to gene expression after it is log-transformed; thus, the additive contributions can be thought of as scaling factors. The idea is that the expression levels of the submatrix defined by a group of genes and a group of samples can be explained by properties (factors) of the rows and columns alone. For example, genes that belong to a certain biological process may have a certain ratio between them when the process is activated. The absolute level of the expression can be determined by the activity level of the process which may be different in each sample. Across the samples in which the process is active these genes form a submatrix that can be entirely explained by a factor for each row, the genes' ratio, and a factor for each column, the activity of the process in the sample. Had one analyzed this submatrix using two-way ANOVA then one would have accepted the null hypothesis that there are no interaction terms. Moreover, this submatrix can be explained (before taking the log) by a single SVD component.

### Signature Algorithm (Ihmels *et al.* )

Ihmels *et al.* [130] introduce an algorithm to identify transcriptional modules from gene expression data. The algorithm starts with a "seed" set of genes, $G_0$, which is obtained from prior biological information regarding the genes, such as known function or genomic sequence. $G_0$ is used to identify those conditions[5], $C_1$, in which these genes are highly regulated, either up-regulated or down-regulated. The next step returns to the genes and searches for the genes which are highly regulated in the $C_1$ conditions; this yields a new set of genes, $G_1$. This procedure can "fish" out other genes (and may also remove ones from the initial seed) that are co-regulated with a coherent subset in the seed. A later paper by Bergman *et al.* analyzes an iterated version of this algorithm, in which one continues identifying condition sets and gene sets alternatively until this process converges (discussed below).

Technically, two matrices are defined, $X^G$ and $X^C$, based on the log-transformed ex-

[5]I follow Ihmels *et al.* and call the columns of the expression matrix *conditions* which are equivalent to *samples* in my nomenclature.

pression matrix $X$. In $X^G$ the genes are centered and normalized[6], *i.e.* $\sum_j X^G_{ij} = 0$, $1/N \sum_j (X^G_{ij})^2 = 1 \; \forall i$ and in $X^C$ the conditions are centered and normalized. In the first step the conditions are scored by averaging $X^G$ on the input set $G_0$;

$$s_j = 1/|G_0| \sum_{i \in G_0} X^G_{ij} \; . \tag{2.21}$$

The set of conditions in which those genes are regulated is determined by

$$C_1 = \{j : |s_j - \langle s_j \rangle_j| > t_C \sigma_C\} \tag{2.22}$$

where $t_c$ is a condition threshold (the authors used $t_c = 2.0$) and $\sigma_c = 1/\sqrt{|G_0|}$ is the standard deviation expected from random fluctuations; assuming the genes in $G_0$ are uncorrelated the average of $|G_0|$ random variables with expectation of 0 and variance of 1 has a mean of 0 and standard deviation of $1/\sqrt{|G_0|}$.

In the second step, the genes are scored by a weighted average of $X^C$ according to $\tilde{s}_i = 1/|C_1| \sum_{j \in C_1} s_j X^C_{ij}$. The $G_1$ genes that are regulated in the $C_1$ conditions are identified in a similar manner by $G_1 = \{i : |\tilde{s}_i - \langle \tilde{s}_i \rangle_i| > t_G \sigma_G\}$, where $\sigma_G$ is the measured standard deviation of $s_i$ and $t_G$ is some threshold chosen by the authors to be 3.0. There are two roles for the weight $s_j$; one is to handle correctly conditions in which the genes are down-regulated. These have a negative score which makes sure the contribution to the score of genes which are indeed down-regulated[7] will be positive.

This procedure is based on two supervised filtering or hypothesis testing steps; the first tests the conditions one-by-one and identifies those in which the null hypothesis that the genes' expression levels are an independent set drawn from a distribution with zero mean and standard deviation of 1 can be rejected. In the second step one tests which genes have a score which is abnormally high.

The advantage of this algorithm is that it is very fast. This enabled the authors to exhaustively analyze a large yeast gene expression database starting with $\approx 86000$ different gene-seeds defined according to whether the genes contain a certain sequence in their upstream region (all six- to eight-mers). The algorithm has two disadvantages: (i) Genes and conditions are tested one-by-one which prevents identification of sets of genes which are highly correlated but are not extremely up- or down-regulated. (ii) Found clusters have a compact shape since the genes and conditions are selected by comparing then to the mean profile of the cluster.

### Iterated Signature Algorithm (ISA) (Bergman *et al.* )

Bergman *et al.* [131] extend the Signature Algorithm by continuing going back and forth updating genes and conditions until convergence and putting this procedure on more formal

[6]Here the normalization is such that the *mean* of the squared components is unity and not their sum (as defined in **??**).

[7]$X^C_{ij} < 0$ if gene $i$ in condition $j$ has an expression level which is below average compared to the other genes in that condition

grounds; this new algorithm is called *Iterative Signature Algorithm* (ISA). First, a rigorous definition of a *transcriptional module* (TM) is given; a TM is a set of genes and conditions $(G, C)$ which are consistent with each other in the following sense: the conditions $C$ induce a co-regulated expression of the genes $G$, *i.e.* the genes are most similar to each other across conditions $C$. Conversely, the molecular profile of the conditions measured for the genes in $G$ are most similar for the conditions in $C$. The similarity is measured in terms of deviation from the average behavior of the remaining genes or conditions. For example, genes which are coherently up-regulated (or down-regulated) compared to their average expression level across the conditions, $C$, are considered similar. The aim of the ISA algorithm is to identify all TMs in the data. In principle, this can be done by testing all sets of genes against all sets of conditions, but this is, of course, impractical due to the exponential number of these sets. In practice this is done by starting with randomly selected seeds and iterating the signature algorithm until convergence. Mathematically, a TM is defined by two thresholds $t_G$ and $t_C$ and two indicator vectors, $\mathbf{g}$ and $\mathbf{c}$, which are nonzero for the genes in $G$ and conditions in $C$ respectively. The consistency criteria can be written as

$$\exists(t_C, t_G) : \left\{ \begin{array}{rcl} \mathbf{g}(\mathbf{c}) & = & f_{t_C}\left(X^C\mathbf{c}\right) \\ \mathbf{c}(\mathbf{g}) & = & f_{t_G}\left((X^G)^T\mathbf{g}\right) \end{array} \right. \tag{2.23}$$

where

$$f_t(\mathbf{x}) = \left( w(x_1)\Theta\left(\frac{x_1 - \langle\mathbf{x}\rangle}{\sigma(\mathbf{x})} - t\right), \ldots, w(x_{N_x})\Theta\left(\frac{x_{N_x} - \langle\mathbf{x}\rangle}{\sigma(\mathbf{x})} - t\right)\right)^T , \tag{2.24}$$

$\langle\mathbf{x}\rangle$ and $\sigma(\mathbf{x})$ are the average and standard deviation of the components of $\mathbf{x}$ and $\Theta(x)$ is the step function and $w(x)$ is a weighting function. In order to include down-regulated genes in the TM one can take the absolute value of the centered and normalized value $\left[x_i - \langle\mathbf{x}\rangle\right]/\sigma(\mathbf{x})$. A TM is a fixed point of the following mapping (given $t_G$, $t_C$)

$$\mathbf{g}^{(t+1)}(\mathbf{c}) = f_{t_C}\left(X^C\mathbf{c}^{(t)}\right) \tag{2.25}$$

$$\mathbf{c}^{(t+1)}(\mathbf{g}) = f_{t_G}\left((X^G)^T\mathbf{g}^{(t)}\right) . \tag{2.26}$$

Since usually the are a limited number of fixed points the algorithms starts at random gene seeds $\left\{\mathbf{g}_m^{(0)}\right\}$ and collects the distinct fixed points $\left\{(\mathbf{g}_m^{(*)}), (\mathbf{c}_m^{(*)})\right\}$. The set of fixed points depend on the values of the threshold parameters; lower thresholds reveal larger TMs which contain ones found using larger thresholds.

The authors point out that the ISA algorithm is closely related to SVD since if one removes the step functions, plugs $w(x) = x$ as a weighting function and uses $X$ instead of $X^C$ and $X^G$, the iterations procedure coincides with method for SVD and converges to a pair $(\mathbf{g}, \mathbf{c})$ of singular vectors of $X$ (In general it is the pair that corresponds to largest eigenvalue). Applying the step function changes and stabilizes the spectrum of fixed points compared to SVD, which makes the ISA results different and more robust against noise.

A recent paper by Ihmels [133] compared the results obtained on their yeast gene expression database from several clustering methods among which are three biclustering

methods; ISA, Cheng and Church's biclustering [123] and our CTWC [108]. Each found bicluster can be tested whether it is self-consistent according to the ISA definition. Agglomerative average linkage (Eisen *et al.* [62]) and ISA clusters were self-consistent while the clusters produced by other algorithms were not. A *biological merit function* (BMF) defined on the basis of conservation of *cis*-regulatory motifs in four related yeast species was used to evaluate the resulting clusters. The ISA, agglomerative average linkage and CTWC provided the best performance. meaning that CTWC which does not identify self-consistent clusters (in the ISA sense) does yield clusters with biological significance (see 2.2.6 for possible explanation).

### Biclustering (Cheng and Church)

Cheng and Church [123] describe an algorithm which searches for submatrices in the gene expression matrix, $X$, by removing and adding genes and conditions in order to obtain a low cost. $X$, as before, contains the log-transformed expression data. For a set of genes, $G$, and a set of conditions, $C$, a *mean square residue* score is defined,

$$H(G,C) = \frac{1}{|G||C|} \sum_{i \in G, j \in C} \left( X_{ij} - X_{iJ} - X_{Ij} + X_{IJ} \right)^2 \ , \tag{2.27}$$

where the row, column and submatrix means are defined as

$$X_{iJ} = \frac{1}{|C|} \sum_{j \in C} X_{ij} \ , \quad X_{Ij} = \frac{1}{|G|} \sum_{i \in G} X_{ij}, \ , \quad X_{IJ} = \frac{1}{|G||C|} \sum_{i \in G, j \in C} X_{ij} \ . \tag{2.28}$$

A submatrix for which $H(G,C) \leq \delta$ for some $\delta \geq 0$ is called a $\delta$-*bicluster*. Note that these are the same residues analyzed in two-way ANOVA, meaning the a biculster with $H(G,C) = 0$ indicates that all the variation in this submatrix can be attributed to changes in the rows and columns. When comparing to SVD, if one decomposes with SVD the original data before taking the log, a biculster with $H(G,C) = 0$ can be exactly described as an external product of two vectors and has a single SVD component.

The algorithm iteratively searches for low cost biclusters and after each one is found it is "erased" from the expression matrix by replacing it with random numbers drawn from the distribution of all the matrix elements. The search is performed by a *greedy* algorithm that starts with the whole matrix, $G = 1, \ldots, N_g$ and $C = 1, \ldots, N_s$ and first tries to remove multiple rows and columns which are proven to decrease $H$: Define

$$d_g(i; C, G) = 1/|C| \sum_{j \in C} \left( X_{ij} - X_{iJ} - X_{Ij} + X_{IJ} \right)^2 \tag{2.29}$$

and

$$d_c(j; C, G) = 1/|G| \sum_{i \in G} \left( X_{ij} - X_{iJ} - X_{Ij} + X_{IJ} \right)^2 \ . \tag{2.30}$$

For a given $\alpha > 1$ the rows for which $d_g(i; C, G) > \alpha H(G, C)$ are removed and then the columns for which $d_c(j; C, G) > \alpha H(G, C)$ are also removed; this is iterated until no more such multiple deletions are found. Then, the algorithm continues removing single rows or columns that decrease $H$; the row $i$ with largest $d_g(i; G, C)$ is removed and then one removes the column $j$ with the maximal $d_c(j; G, C)$. This stage of the algorithm stops as soon as $H(G, C) \leq \delta$. At this point, the bicluster is at its minimal size and the next step is to add sets of columns and rows which once more decrease $H$: Add the columns for which $d_c(j; G, C) \leq H(G, C)$ and then add the rows for which $d_g(i, G, C) \leq H(G, C)$. This yields the finally reported bicluster whose elements are then replaced with random numbers. Note that after each removal or inclusion of rows or columns $G$ and $C$ are updated and therefore $d_c(j; G, C)$ and $d_g(i, G, C)$ need to be updated. In practice, the algorithm first converges to trivial bi-clusters, ones in which there was no change between the conditions. These are "erased" in the first iterations. The biclusters are sorted according to their mean row variance assuming that high-variance clusters are the most interesting ones.

**Plaid models - Lezzeroni and Owen**

Lazzeroni *et al.* [122] introduced a method called *Plaid models*. The gene expression matrix is described as a sum of $K$ layers;

$$X_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} + \epsilon \tag{2.31}$$

where $\mu_0$ is the background level of gene $i$ and $\mu_k$ the contribution to the basal (average) level from layer $k$. $\alpha_{ik}$ and $\beta_{jk}$ are additive effects of the gene and sample respectively. Finally, $\rho_{ik} \in \{0, 1\}$ indicates whether gene $i$ belongs to layer $k$; similarly, if sample $j$ belongs to layer $k$ then $\kappa_{jk} = 1$.

The cost of a model is its sum of squared residues,

$$E(\{\theta_{ijk}\}, \{\rho_{ik}\}, \{\kappa_{jk}\}) = \frac{1}{2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \left( X_{ij} - \mu_0 - \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} \right)^2 \tag{2.32}$$

where $\{\theta_{ijk}\}$ represent the parameters $(\mu, \alpha, \beta)$. The parameters are found by iterative optimization. In each iteration a different layer is optimized while keeping the parameters of the other layers fixed. Note that the requirement that $\rho$ and $\kappa$ are binary is relaxed and is gradually enforced during the iterative optimization. Suppose that the parameters of $K - 1$ layers are given; then the parameters for the $K$-th model are found by minimizing

$$E^K(\{\theta_{ijK}\}, \{\rho_{iK}\}, \{\kappa_{jK}\}) = \frac{1}{2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \left( X_{ij}^K - (\mu_K + \alpha_{iK} + \beta_{jK}) \rho_{iK} \kappa_{jK} \right)^2 \tag{2.33}$$

where $X_{ij}^K = X_{ij} - \mu_0 - \sum_{k=1}^{K-1} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk}$. This is done as follows:

1. Initialization: set $\theta^{(0)} = 1$, *i.e.* $\mu_K = \alpha_{iK} = \beta_{jK} = 1$. $\rho_{iK}^{(0)}$ and $\kappa_{jK}^{(0)}$ are set to $\mathbf{u_1}$ and $\mathbf{v_1}$ the singular vectors corresponding to the largest eigenvalue.

2. Perform $s = 1, \ldots, S$ iterations in which $\theta^{(s)}$ is updated based on $\rho^{(s-1)}$ and $\kappa^{(s-1)}$, then $\rho^{(s)}$ is updated from $\theta^{(s)}$ and $\kappa^{(s-1)}$ and finally, $\kappa^{(s)}$ is updated from $\theta^{(s)}$ and $\rho^{(s-1)}$. $\theta$ is always one step ahead of $\rho$ and $\kappa$. The final parameters are $\rho^{(S)}$, $\kappa^{(S)}$ and $\theta^{(S+1)}$. Updating of $\theta$ is done by minimizing $E^K$ under the constraints $\sum_{i=1}^{N_g} \rho_{iK}^2 \alpha_{iK} = \sum_{j=1}^{N_s} \rho_{jK}^2 \alpha_{jK} = 0$ using Lagrange multipliers. The same is done in order to update $\rho$ and $\kappa$.

This update procedure is cycled over the $K$ layers several times. The number of layers $K$ is chosen such that the *size* of the last added layer, its sum of squared elements, has a corresponding p-value less than some threshold. The p-value is estimated by searching for layers in a random permutation of the matrix.

The idea of the algorithm is that each cellular process can be described by a single layer of the plaid model. These processes effect only a subset of the genes and samples (represented by $\rho_{ik}$ and $\kappa_{jk}$). The subsets associated with different layers (or processes) can overlap. The degree of activity of the process may vary among the samples which it effects (manifested in $\beta_{jk}$) and the expression of genes that take part in the process can change by various factors[8] (represented by $\alpha_{ik}$). A single layer of the plaid model can be thought of as a "local" two-way ANOVA.

### Double Conjugated Clustering - Busygin *et al.*

Busygin *et al.* [129] suggest an algorithm to perform clustering of genes and samples simultaneously. Their method is a framework that can be based on any centroid clustering algorithm, e.g. k-means or SOM. Here a centroids is actually a pair of *conjugate* centroids; a sample-centroid $\mathbf{y}$ and its corresponding (via a one-to-one mapping) gene-centroid $\mathbf{y}^c$. Clustering is performed by alternating the training steps, of the chosen underlying clustering method, between the gene- and sample-space. In general, training is performed by assigning the points to centroids and then updating the centroids according to their assigned points. Next, the centroids are mapped to the conjugate space (from genes to samples and *vice versa*) and a training step is performed in the conjugate space, *i.e.* points are assigned to centroids and the centroids are updated according to their assigned points. The algorithm terminates as soon as both the number of samples and genes that change their cluster assignment drop below a threshold.

The metric used between a centroid, $\mathbf{y}$, and a vector $\mathbf{x}$ is the angle measure (or the cosine measure as described in 2.2.2), $\theta(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$, *i.e.* the larger the dot product between the vectors after they are normalized to unit length, the closer the vector is to the centroid. Denote by $X_G$ a gene-normalized expression matrix in which each gene-vector is

[8]The model is additive by since the expression data is *log*-transformed the parameters represent fold change

normalized to a unit length and by $X_S$ a sample-normalized expression matrix. In order to transform a sample-centroid $\mathbf{y}$ to its *conjugate* gene-centroid $\mathbf{y}^c$ and vice versa one uses:

$$\mathbf{y}^c = B(X_S^T \cdot \mathbf{y}) \quad (2.34)$$

$$\mathbf{y} = B(X_G \cdot \mathbf{y}^c) \quad (2.35)$$

where $B(\mathbf{y}) = \mathbf{y}/\|\mathbf{y}\|$ projects on the unit sphere.

This algorithm is close to ISA and SVD. It uses similar transformations between the gene-space and the sample-space. In all these algorithms one constructs a vector from the dot-product between a centroid and all of the objects it represents (genes or samples) which measures the similarity between the centroid and each of the objects. For example, a gene-centroid that represents well a cluster of genes is mapped to the conjugate sample-centroid whose components for these genes is close to unity and the remaining ones are close to zero. In the next step, other samples in which these genes are highly expressed will have a large dot-product with the sample-centroid and thus be assigned to the cluster while shifting the sample-centroid. This process is repeated until convergence. Note that these methods are well suited to identify up-regulation (or down-regulation if one uses the absolute dot-product) but will perform poorly if one has only a correlated behavior as can be described by a plaid-model or captured using density based clustering as in CTWC.

**Comments on biclustering methods**

There are two major differences between coupled two-way clustering (CTWC) and other biclustering methods. One advantage of other biclustering methods is that they often define a cost function which they aim to minimize. This aspect is missing in CTWC. CTWC uses the stability measure of SPC to identify statistically significant clusters but does not have an overall score for the pair of gene- and sample- clusters. In that respect CTWC does not yield a classical bicluster, (G$M$,S$N$), which is optimal for both the genes and samples. The statements it produces are of the kind "S$J$(G$I$) $\rightarrow$ S$N$" or "G$I$(S$J$) $\rightarrow$ G$M$" each of which are not symmetric with respect to the genes or samples. Biclusters, on the other hand, aim to be maximal (or optimal) in both sets. Moreover, in CTWC the clusters that are used as features for clustering are themselves stable clusters in some other clustering operation. This requirement enables CTWC to perform its heuristic search for significant partitions but as a result, its ability to identify separations which are based on more than one cellular mechanism is limited.

Another difference which gives CTWC the ability to generate clusters which can not be revealed by other methods is the fact CTWC uses a density estimation clustering method as its engine. Other methods practically use a representative/centroid-type methods since their cost penalizes including genes (or samples) according to whether they fit the average behavior of the other genes (or samples). In general, a cluster that has an elongated (or any other irregular) shape in gene- or in sample- space or in both does not fit the models of other biclustering methods but can form a dense region that is identified by SPC.

An example of a cluster that can be found by CTWC but would elude other methods was found in analysis of leukemia samples in Publication (9) [113]. This cluster contain genes that indicate the stage of differentiation at which the leukemic cells became malignant. Sorting these genes using SPIN, an algorithm by Tsafrir *et al.* [106] that sorts points into neighborhoods, reveals Figure 2.6. One can clearly see that not all genes are correlated with each other, but rather they can be sorted in a linear order such that only close genes according to this order are correlated – meaning they can be ordered along some linear trajectory in gene space. A similar pattern can also be seen in the samples (columns of the matrix). Such a cluster cannot be described by some average behavior and even a plaid model with an additive effect of a gene and sample can not explain the expression levels in this cluster.

In addition, in the work of Ihmels *et al.* [133] they tested if clusters produced by CTWC fit the definition of a transcriptional module and it turned out that many of them do not. On the other hand, many clusters scored highly in a biological merit function that is based on regulatory patterns in the upstream region of the genes (in different yeast species). This demonstrates that irregular shaped clusters do appear in gene expression data and have a biological meaning.

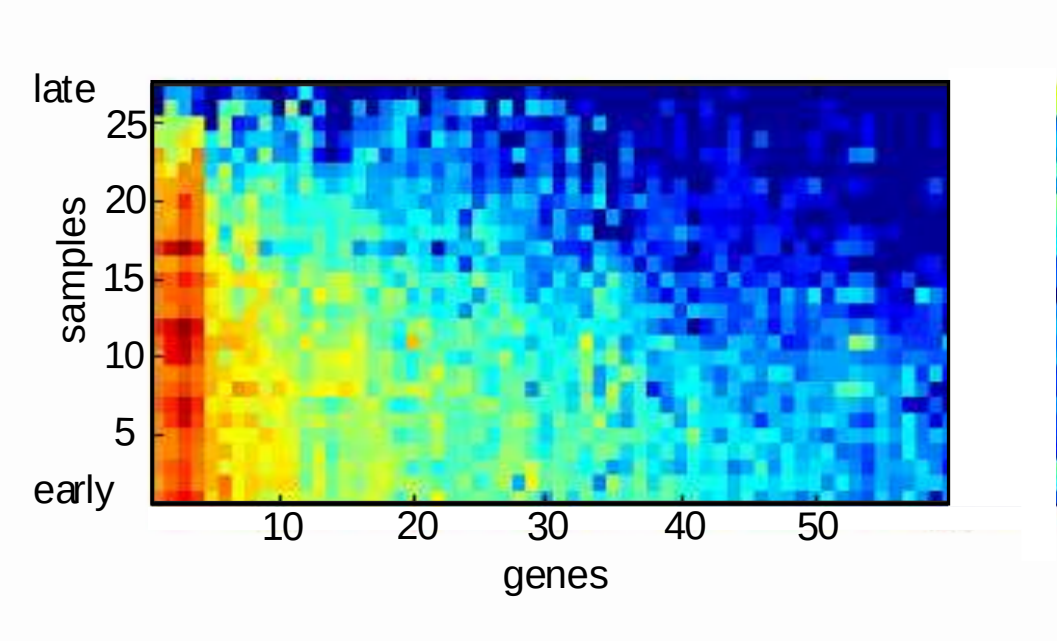



Figure 2.6: Gene expression of 60 genes related to the differentiation stage over 27 leukemia samples (data taken from Publication (9)). The genes are up-regulated at early stages of differentiation. The matrix was ordered using the SPIN algorithm [106]. One can see that the genes are not correlated with the average expression profile and the samples are not close to the average molecular profile, but rather each gene (sample) is close to some other genes (samples) that are positioned near to it in this linear ordering. Such a cluster can be identified by SPC but can not be identified using centroid-based biclustering methods.

## 2.3 Semi-supervised methods

Publication (3) presents a novel method for semi-supervised analysis. This method uses the same scheme described in Sec. 1.11 to identify typical solutions. As described in the Introduction (see Sec. 1.10), semi-supervised methods deal with classification of points in case a large dataset is available but only a fraction (usually small) of the data points are labelled.

Our method is based on an inhomogeneous Potts model (as in SPC) with the number of states q exceeding the number of known classes. The points with known classification are fixed in the Potts state that corresponds to their label, *i.e.* they feel an infinite field to that direction. The unlabelled points are free spins. As in SPC our aim is to find the typical cut at various temperature values and perform the classification based on robust typical cuts (ones which are kept for a large range of temperatures).

We present a toy problem of data selected at random from one of four Gaussian distributions, each corresponding to a different class. Since the underlying distribution is known, we also know the Bayessian classification rule (the optimal classification) which serves as ground truth. Figure 1 of Publication (3) depicts 400 points randomly sampled from this distribution; 7 of these points are labelled. Figures 3 to 5 in the paper demonstrate our main claim: a few labelled points can have a dramatic effect. They decrease the number of classification errors to almost none and at the same time enlarge the temperature range in which this error level is maintained. Moreover, the classification identified by the method is close to the ground truth, as opposed to those found by the commonly applied mincut methods which search for the optimal (minimal energy) classification, instead of the typical one.

In order to calculate the typical solution we apply various sampling methods (Monte-Carlo) and approximation methods (message passing algorithms adopted from graphical models). We show that on our example the Multicanonical Monte-Carlo (MC) method [87–90], Generalized Belief Propagation (GBP) [134] and a new method we present, called Weighted Belief Propagation (WBP), yield similar (correct) results for a wide range of temperatures. Below a certain temperature GBP does not converge, whereas MC and WBP give similar results. There are differences between MC and WBP at very low temperatures, although both methods find the lowest energy configuration, the one that corresponds to the mincut solution.

The Weighted Belief Propagation (WBP) is based on a weighted average of Belief Propagation (BP) results obtained by different initial messages. We weigh the resulting distributions, $P_i(S)$ according to $\exp[-F(P_i)/T]$ where $F(P_i)$ is the free energy of the BP solution. We prove that these weighting coefficients are optimal.

## 2.4 Hypothesis testing

In this Section I describe methods that we used and developed for hypothesis testing. First I discuss the importance of robust statistics in gene expression. Following that, I give a brief description of several methods: Fisher's exact test and other methods for the analysis of contingency tables [41], Threshold number of misclassification (TNoM) [39] to identify genes that can be used to differentiate between tumor classes and methods for survival analysis [135] mainly used in cancer research. As part of the survival analysis I introduce a method we developed which is, in a sense, a combination of TNoM and Kaplan-Meier.

### 2.4.1 Robust statistics

Gene expression data is very noisy and suffers from occasional outliers. Moreover, the common assumption that the data is normally distributed cannot be automatically applied. Therefore, one has to apply *robust* statistical measures and statistical tests which are not affected by outliers or by the possible long-tail of the distribution [136]. For example, the scaling method implemented in Affymetrix' MAS software, which is used to bring measurements from different experiments to the same scale, is based on the *trimmed mean*. The trimmed mean has one parameter, $p$, the percent of top and bottom values which are removed from the analysis. In order to calculate the trimmed mean one first sorts the values $\{x_i\}_{i=1}^{N}$ and obtains $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$. Next, the top and bottom $\lfloor pN \rfloor$ values are dropped and the trimmed mean is simply the mean of the remaining values,

$$\bar{X}_p = \frac{1}{N - 2\lfloor pN \rfloor} \sum_{i=\lfloor pN \rfloor + 1}^{N - \lfloor pN \rfloor} x_{(i)} \ . \tag{2.36}$$

The median is nothing but the trimmed mean for $p = 1/2$.

Another aspect of robust statistics is the use of non-parametric tests. These test do not assume anything regarding the underlying distribution of the values. Such tests are usually based on ranks instead of the values themselves. A simple example of a parametric test and its parallel non-parametric test the the two sample $t$-test and Wilcoxon ranksum test (similar to the Mann Whitney $U$-test) [41]. Consider two sets of numbers, $X = \{x_i\}_{i=1}^{N_x}$ and $Y = \{y_i\}_{i=1}^{N_y}$ where $N_x \leq N_y$. In $t$-test one tests the null hypothesis that the mean of two distributions from which these populations were drawn are equal, *i.e.* $\mu_x = \mu_y$. For large samples one can approximate the distribution of the means by a normal distribution (which is parametric) due to the central limit theorem. In case the means are indeed equal, and assuming the variances are unknown but equal, the $t$ statistic, which is defined as

$$t = \frac{\widehat{\mathrm{E}}[X] - \widehat{\mathrm{E}}[Y]}{\sqrt{\left((N_x - 1)\widehat{\mathrm{Var}}[X] + (N_y - 1)\widehat{\mathrm{Var}}[Y]\right)(1/N_x + 1/N_y)}} \tag{2.37}$$

where $\widehat{\mathrm{E}}[\cdot]$ and $\widehat{\mathrm{Var}}[\cdot]$ are the estimated mean and variance, follows a Student's $T$ distribution and thus is used to calculate the $p$-value. For data in which the samples are small and one cannot assume normality of the mean, one has to turn to the non-parametric Wilcoxon ranksum test. The null hypothesis of this test is that the two groups are drawn from the same distribution. The statistic of the ranksum test, $W$, is the sum of ranks in the smaller sample where the ranks are assigned according to the union of both samples (details of the method can be found at [41]). In case there are no ties in the data, the ranks are the integers $1, \ldots, N_x + N_y$ and the statistics of $W$ is determined by all possible sums of $N_x$ integers randomly chosen out of $\{1, \ldots, N_x + N_y\}$. The $p$-value is the fraction out of all possible sums for which the value of the statistic is more extreme than $W$. The $p$-value can be calculated exactly for small samples or estimated using a gaussian approximation or by sampling permutations for larger samples.

### 2.4.2 Fisher's exact test

Fisher's exact test is a method to test independence in a two-way table, also known as a *contingency* table [41]. Take for example a cluster of samples, S$I$, that was found by clustering gene expression data of $N_s$ samples with known labels; tumor ($T$) and normal. The samples can be arranged in a $2 \times 2$ contingency table in which the rows are named "belongs to S$I$" and "does not belong to S$I$" and the columns "Tumor" and "Normal". Each sample is placed in one of the four cells according to its properties; such an assignment of samples to each cell is shown in Table 2.6. The aim of Fisher's exact test is to determine whether the assignment of a sample to the cluster S$I$ depends on its type; the null hypothesis is that these two binary variables are independent. There are actually three different null hypotheses (model I, II and III) which can be tested, depending on the experiment that generated the data. In model I the total number of samples is fixed but the marginal sums are free to vary in the experiment. The dependency must, therefore, be tested compared to all possible contingency tables with the same total number of samples. In Model II, the marginal sums of one of the variables is fixed and the other is free to change. Finally, in Model III the marginal sums of both variables are fixed and the $p$-value is the probability to obtain the observed values or worse departures from independence out of all possible $2 \times 2$ contingency tables with the same fixed marginals, both for the rows and the columns. Fisher's exact test was designed for Model III. Other statistical tests to analyze contingency tables are the $G$-test and $\chi^2$ test (see Sokal and Rohlf [41]).

There number of possible ways to partition the $N_s$ samples into a $2 \times 2$ contingency table with certain marginals in an independent manner is

$$\binom{N_s}{|\mathrm{S}I|}\binom{N_s}{|T|} = \binom{N_s}{a+b}\binom{N_s}{a+c} = \frac{N_s!}{(a+b)!(c+d)!} \times \frac{N_s!}{(a+c)!(b+d)!} \ . \tag{2.38}$$

Out of these the number of possibilities to obtained the observed values are $N_s!/(a!b!c!d!)$.

Consequently, the probability to obtain exactly the observed table is given by

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N_s!} \ . \tag{2.39}$$

To obtain the $p$-value one has to sum such terms for the observed values and tables which are less probable. As in other many cases, one is usually interested in the two-tailed distribution, *i.e.* one should take into account more extreme cases in the two ends – ones in which S$I$ corresponds to tumor samples and ones in which it corresponds to the normal ones. Note that there is only one degree of freedom in this setting since from the overlap of S$I$ with $T$ $(= a)$ and the fixed marginals one can obtain the remaining variables. If one predicts the type of a sample based on its membership in cluster S$I$, the number of errors is $\min(a+d, b+c)$ since one has the freedom to choose whether S$I$ corresponds to tumor or normal. In practice, more extreme tables are obtained by decreasing $a$ and $d$ by one and increasing $b$ and $c$ by one if $ad - bc < 0$ and switching increasing and decreasing otherwise. Note that this step keeps the marginals. This is continued until one of the table elements vanishes. The other extreme tail can be approximated by taking twice the $p$-value of the single tail result or can be calculated exactly by reversing the described procedure, starting by setting to zero the smaller of $b$ or $c$ (for $ad - bc < 0$) or $a$ or $d$ (for $ad - bc \geq 0$) and adding tables until the accumulated probability exceeds that of the first tail (not including the last table which passed the $p$-value of the first tail).

We applied Fisher's exact test in publication (6) [102] to calculate the $p$-value for the number of errors in predicting Type 1 diabetes based on antigen reactivity data.

| | Tumor | Normal | total |
|---|---|---|---|
| belongs to S$I$ | $a$ | $b$ | $\|SI\| = a+b$ |
| does not belong to S$I$ | $c$ | $d$ | $N_s - \|SI\| = c+d$ |
| total | $a+c = \|T\|$ | $b+d = N_s - \|T\|$ | $N_s$ |

Table 2.6: A contingency table of cluster S$I$ and Tumor/Normal labels.

### 2.4.3 Threshold number of misclassification (TNoM)

The "Threshold Number of Misclassification" (TNoM) score, suggested by Ben-Dor *et al.* [39], is used to test whether thresholding the expression level of a certain gene can be used to predict a given binary classification of the samples. This test is used to identify "relevant" genes with respect to a certain classification and can replace a measure of correlation to a binary vector indicating the class-type. The TNoM score is defined as the minimal number of errors obtained when classifying the samples by using a threshold on the expression level of a gene. To obtain the minimal number of errors, one first sorts the expression levels of the examined gene across the samples from low to high values. Next, besides each expression level one states the label of the corresponding sample. Finally, one

scans the $N_s - 1$ possible partitions into two groups of low and high expression levels and searches for the partition with minimal number of misclassification. Formally, let $\{g_j\}_{j=1}^{N_s}$ represent the expression levels of a certain gene in sample $j$ and assume the samples are ordered such that $g_1 > g_2 > \ldots > g_{N_s}$ (assume there are no equalities). The classification of the samples can be described by $\{v_i\}$ – a vector of $\{+,-\}$ of length $N_s$ with $N_A$ positives and $N_B$ negatives (a $'+'$ represents a tumor of type $A$ and $'-'$ a tumor of type $B$) ordered according to $\{g_j\}$. Using the notation of [40], one can define a path, $\pi_v(i)$, starting from $(0,0)$ and ending at $(N_s, N_A - N_B)$ which is the sum of $v_i$ up to position $i$; $\pi_v(i) = \sum_{j=0}^{i} v_i$. The TNoM score of the gene is given by $\min(N_A - \max_i \pi_v(i), N_B + \min_i \pi_v(i))$ since the first argument is the minimal number of errors obtained for a classifier which associates the lower expression levels with class $A$ (called an $AB$ classifier) and the second argument measures the minimal number of errors if one assigns lower expression levels to class $B$ (called a $BA$ classifier). The best possible gene is one for which either all the $'+'$s precede the $'-'$s and thus $\pi_v(N_A) = N_A$ or all the $'-'$s precede the $'+'$s in which case $\pi_v(N_B) = -N_B$.

The $p$-value for a gene with a TNoM score of $s$ is the fraction of all possible permutations of the labels which attain a score of $s$ or lower. This fraction is equivalent to the fraction of paths going from $(0,0)$ to $(N_s, N_A - N_B)$ whose score is at most $s$ (lower scores are better). For any such path there exists an $i$ for which $\pi_{v'}(i) \geq N_A - s$ or $\pi_{v'}(i) \leq s - N_B$. Ben-Dor *et al.* [40] provides a closed form formula to calculate this p-value. The method is based on repeated reflections (see [137]). Let $U = N_A - s \geq 0$ and $D = s - N_B \leq 0$ the $p$-value is then defined as

$$P(\text{TNoM} \leq s) = \nu(U, D) \binom{N_A}{N}^{-1} \tag{2.40}$$

where

$$\nu(U, D) = \left| \left\{ \pi : \pi(0) = 0, \pi(N_s) = N_A - N_B, \max_i \pi(i) \geq U \text{ or } \min_i \pi(i) \leq D \right\} \right| . \tag{2.41}$$

$\nu(U, D)$ represents the number of paths starting at $(0,0)$ and ending at $(N_s, N_A - N_B)$ which visit $y = U$ or $y = D$ or both. This is equivalent to a 1D random walker that starts at the origin at $t = 0$ and reaches $N_A - N_B$ at $t = N_s$, and the question is how many different walks are there in which the random walker leaves the interval $[D+1, U-1]$. All the paths can be divided into overlapping families; ones which visit $U$, ones which visit $D$, ones which visit $U$ and at some time later $D$ (called $UD$), etc.. Denote by $w$ a $U/D$ alternating sequence of length $l$ and by $\Lambda(w)$ the number of paths for which there exists a set of indices $\{i_1, \ldots, i_l\}$ in which the path visits $U$ and $D$ according to the sequence. Applying the Inclusion-Exclusion principle one can obtain

$$\begin{aligned} \nu(U, D) &= (\Lambda(U) + \Lambda(D)) - (\Lambda(UD) + \Lambda(DU)) + \\ &\quad (\Lambda(UDU) + \Lambda(DUD)) - \ldots \qquad (2.42) \\ &= \sum_{\text{U/D-seq}: w \leq \left\lceil \frac{N_A - N_B}{U - D} \right\rceil} (-1)^{w+1} \Lambda(U/D\text{-seq}) \qquad (2.43) \end{aligned}$$

The number of paths that visit a particular pattern can be calculated using repeated reflections which maps the paths, with a one-to-one relation, to ones which start at $(0, t(w))$ and end at $(N_s, N_A - N_B)$ without any constraints. The number of such paths is simply

$$\Lambda(w) = \binom{N_s}{(N_s + N_A + N_B - t(w))/2} \tag{2.44}$$

where

$$t(w) = \begin{cases} 2\sum_{i=1}^{l} |w(i)| & \text{if } w(l) = U \\ -2\sum_{i=1}^{l} |w(i)| & \text{if } w(l) = D \end{cases} \tag{2.45}$$

For each path identify the sequence of $U$ or $D$ crossing and then identify the first of each type to produce a $U/D$-alternating sequence, *e.g.* $UUDUDDDDUD \to UDUDUD$. Next, reflect the path with respect to $y = U$ or $y = D$ depending on the crossing. This generates a path which starts at $(0, t(w))$ and ends at $(N_s, N_A - N_B)$.

### 2.4.4 Survival analysis

Another supervised technique deals with comparing survival data between two groups. Consider an experiment in which one measures for $N$ patients with a certain type of cancer the time, $T_i$, from diagnosis to the time of death (one can also measure the time to other events, such as appearance of metastases). Such an experiment is usually carried out during a fixed period of time and when it ends some of the patients may still be alive. For those patients, the time from diagnosis to the end of the experiment is recorded and they are labeled as "censored", $C_i = 1$ and $C_i = 0$ otherwise. The existence of censored values is the difference between standard statistical comparison between two groups and the tests used for survival data.

First, I will describe how to estimate the probability to live longer than $t$ and generate Kaplan-Meier estimates (and plots) and then comparison between two sets is presented.

Denote by $T$ the life time random variable; $0 \leq T < \infty$ and has a cumulative distribution $F$ and a density function $f$. The *survivor function*, $S(t)$, is defined as the probability to live longer than $t$, $S(t) = P(T > t) = 1 - F(t)$. The *hazard function* $h(t) = \lim_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t)/\Delta t$ is the instantaneous rate of death. One also defines the *cumulative hazard function* $H(t) = \int_0^t h(s)ds$. There are simple relations between these functions,

$$h(t) = \frac{f(t)}{S(t)}, \quad H(t) = -\log S(t) \ . \tag{2.46}$$

If there are no censored data than estimating the survivor function is simply done by taking one minus the empirical cumulative distribution function; $\hat{S}(t) = 1 - \#(T_i \leq t)/N_s$. In case there are censored data, one includes at time $t$ only the censored patients registered at times $T_i \geq t$ since they are known to be alive at time $t$; the ones for which $T_i < t$ are ignored since their status at $t$ is unknown. It is convenient to define the number of patients

still alive at time $t$, and hence at *risk*, as $r(t)$; $r(t)$ includes censored patients which are still alive but ones who were censored before $t$ and, hence their status is unknown, are not included in $r(t)$. The range $[0,\infty)$ is divided to intervals $I_i = [t_i, t_{i+1})$; The times $\{t_i\}$ are placed just before all death events (not censored ones).

The probability to survive interval $I_i$ is estimated by $s_i = [r(t_i) - d_i]/r(t_i)$ where $d_i$ is the number of deaths in interval $I_i$ which is at least one due to the choice of intervals. The estimated survivor function at time $t$ is thus given by

$$\hat{S}(t) = \prod_{t_i<t} s_i = \prod_{t_i<t} \frac{r(t_i) - d_i}{r(t_i)} \ . \tag{2.47}$$

This is the Kaplan-Meier estimator. Figure 2.7 depicts a Kaplan-Meier curve obtained for breast cancer patients taken from Van't veer *et al.* [52]. We analyzed this data in Publication (10). As for any estimate, the survivor function can be assigned confidence intervals which can be calculated using various methods. We use Greenwood's formula [135] which is based on similar arguments,

$$\mathrm{var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum \frac{d_i}{r(t_i)\left[r(t_i) - d_i\right]} \tag{2.48}$$

The 95% confidence interval is marked in Fig. 2.7 by a gray region[9].

Testing for two sets of patients, $A$ and $B$, the null-hypothesis that their survival function is the same can be performed using the non-parametric Mantel-Cox log-rank test [42] (also called Mantel-Haenszel test). In this test one creates a sequence of $2 \times 2$ contingency tables (see **??**) one for each interval at each (uncensored) observed death . In each contingency table, presented in Table 2.7, the number of "deaths" in set $A$, $d_i^A$ has a hypergeometric distribution. Here we use the $\chi^2$ approximation in which $d_i^A$ is compared to its expected value $\mathrm{E}[d_i^A] = r^A(t_i) d_i^{AB} / r^{AB}(t_i)$ which is based on the marginals of the contingency table ($d_i^{AB}$ and $r^{AB}(t_i)$ are defined in Table 2.7). The variance of $d_i^A$ is given by $\mathrm{Var}[d_i^A] = \mathrm{E}[d_i^A] r^B(t_i)(r^{AB}(t_i) - d_i^{AB}) / \left(r^{AB}(t_i)(r^{AB}(t_i) - 1)\right)$. The variable $X_i^2 = \left(d_i^A - \mathrm{E}[d_i^A]\right)^2 / \mathrm{Var}[d_i^A]$ is approximately $\chi^2$ distributed with one degree of freedom. Assuming the $\{d_i^A\}$ are independent, the variable $X^2 = \left[\sum_i \left(d_i^A - \mathrm{E}[d_i^A]\right)\right]^2 / \sum_i \mathrm{Var}[d_i^A]$ can also be approximated by $\chi_1^2$ since a sum of independent Gaussian variables, $\{d_i^A\}$, is also Gaussian. Therefore, the $p$-value of the Mantel-Cox is calculated by

$$P(S^A(t) = S^B(t)) = \int_{X^2}^{\infty} \chi_1^2(x) dx \ . \tag{2.49}$$

[9] In practice, in order that the confidence interval will be confined to $(0,1)$ we use $\exp\{\hat{H}(t) \exp[\pm 1.96\, \mathrm{std}(\hat{H}(t))/\hat{H}(t)]\}$ where $\hat{H}(t) = \sum_i d_i / r(t_i)$ and $\mathrm{var}(\hat{H}(t)) = \sum d_i / [r(t_i)\,(r(t_i) - d_i)]$.

| Table at time $t_i$ | Deaths | Survivals | total |
|---|---|---|---|
| set $A$ | $d_i^A$ | $r^A(t_i) - d_i^A$ | $r^A(t_i)$ |
| set $B$ | $d_i^B$ | $r^B(t_i) - d_i^B$ | $r^B(t_i)$ |
| total | $d_i^{AB} = d_i^A + d_i^B$ | $r^{AB}(t_i) - d_i^{AB}$ | $r^{AB}(t_i) = r^A(t_i) + r^B(t_i)$ |

Table 2.7: The contingency table that corresponds to time $t_i$.

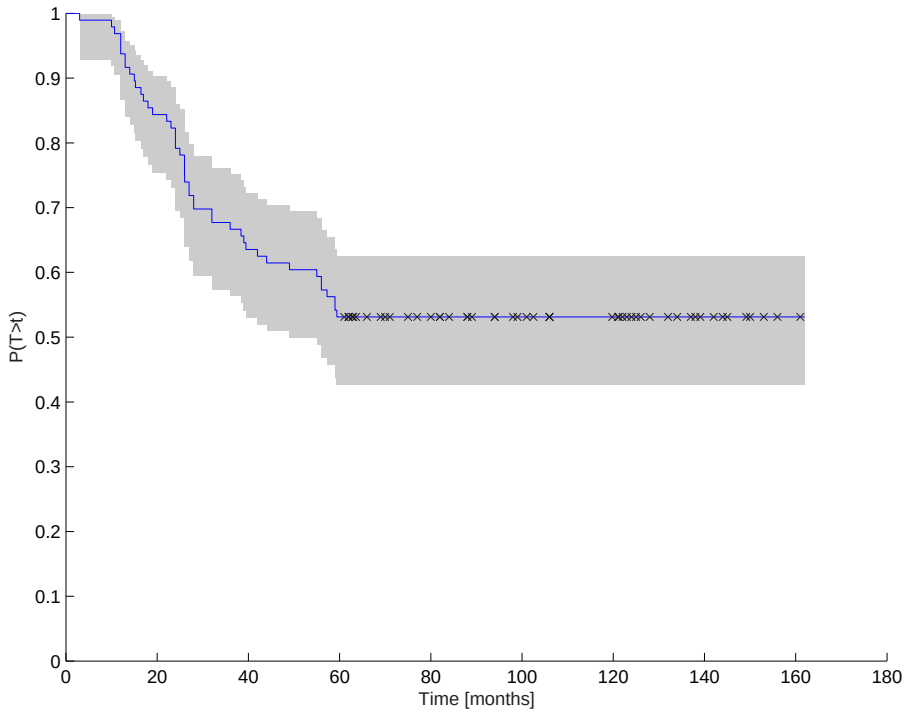



Figure 2.7: Kaplan-Meier plot for metastasis-free time interval (MFTI) of breast cancer patients taken from Van't Veer *et al.* [52]. The Kaplan-Meier plot estimates the survivor function, *i.e.* the probability to survive (or in this case be metastasis-free) longer than $t$. The $\times$'s mark censored events
and the gray region represents the 95% confidence interval of the survivor function.

### Most Different Survivor Curves (MDSC)

Part of our analysis of breast cancer expression and survival data, we introduced a method which follows the same spirit as the TNoM test (see 2.4.3) for classification but instead of searching for the threshold that generates the least number of errors we search for a threshold that partitions the patients into two groups which have the most different survival functions (have the lowest p-value). We call this method the *Most Different Survivor Curves* test (MDSC). Then, in order to assign a p-value for such a separation, that compensates for the selection of the best separation, we performed a permutation test. We randomly generated $10^6$ permutations of the expression levels of the tested gene (this analysis is done

only once since the ranks and not the expression levels themselves enter the calculations) and searched for the separation with lowest p-value and recorded it. The distribution of these $10^6$ optimal p-values was used to assign a true p-value for the examined genes and clusters. For example, Figures 2.8 and 2.9 depict the MDSC partition of 96 breast cancer samples taken from Van't Veer *et al.* [52]. The partition is performed by thresholding the expression level of Cyclin E2, a known marker for poor prognosis in breast cancer [138]. The $p$-value for this partition (using Mantel-Cox log-rank test) is $4.3 \times 10^{-5}$ and its corrected value is $1.5 \times 10^{-3}$.

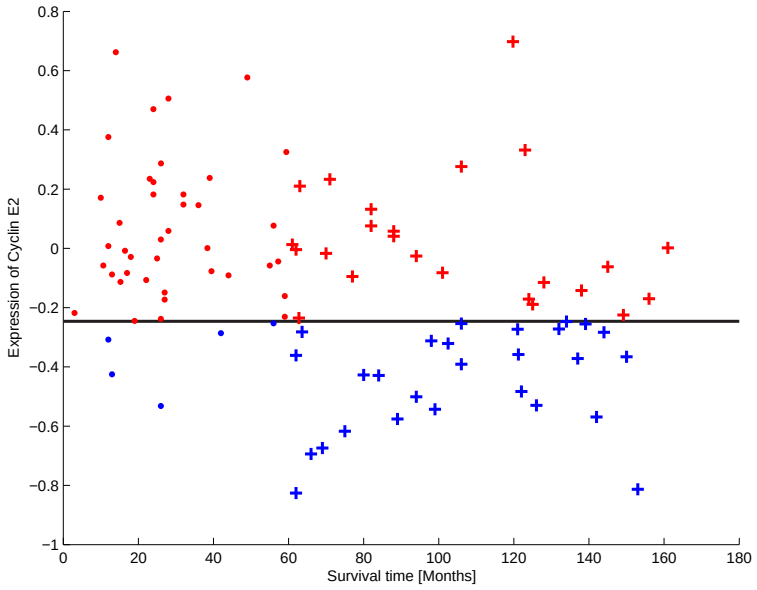



Figure 2.8: Log-transformed expression levels of Cyclin E2 (CCNE2) vs. metastasis-free time interval (MFTI) of 96 breast cancer samples from Van't Veer *et al.* [52]. Pluses represent censored data points, which in this case are all observations at $\geq 60$ months. The horizontal line depicts the optimal partition of the samples, based on their expression level of Cyclin E2, to two groups, red and blue, for which the Mantel-Cox log-rank test attains its minimal value. The Kaplan-Meier plots of these two groups appear in Figure 2.9.

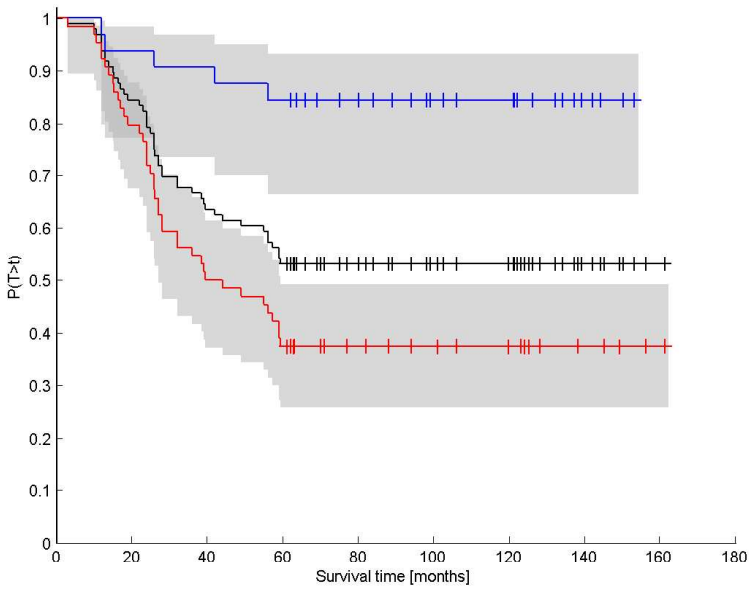



Figure 2.9: Kaplan-Meier plots of the optimal partition found by the Most Different Survivor Curves (MDSC) method (red and blue). The Mantel-Cox log-rank test between the two curves yields a $p$-value of $4.3 \times 10^{-5}$. The corrected $p$-value which compensates for selecting the optimal partition is $1.5 \times 10^{-3}$. The black line represents the Kaplan-Meier curve of the whole data set. The gray regions are 95% confidence intervals.

# Published Works

The following pages contain the published works related to the preceding chapter.

# Publication 1:

## Coupled two-way clustering analysis of gene microarray data

Authors: G. Getz, E. Levine and E. Domany
Published in: *Proc. Natl. Acad. Sci.* **97**, 12079–12084 (2000).

# Coupled two-way clustering analysis of gene microarray data

**Gad Getz, Erel Levine, and Eytan Domany***

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Bradley Efron, Stanford University, Stanford, CA, and approved August 14, 2000 (received for review March 27, 2000)

**We present a coupled two-way clustering approach to gene microarray data analysis. The main idea is to identify subsets of the genes and samples, such that when one of these is used to cluster the other, stable and significant partitions emerge. The search for such subsets is a computationally complex task. We present an algorithm, based on iterative clustering, that performs such a search. This analysis is especially suitable for gene microarray data, where the contributions of a variety of biological mechanisms to the gene expression levels are entangled in a large body of experimental data. The method was applied to two gene microarray data sets, on colon cancer and leukemia. By identifying relevant subsets of the data and focusing on them we were able to discover partitions and correlations that were masked and hidden when the full dataset was used in the analysis. Some of these partitions have clear biological interpretation; others can serve to identify possible directions for future research.**

In a typical DNA microarray experiment, expression levels of thousands of genes are recorded over a few tens of different samples† (1, 3, 4). This new technology gave rise to a computational challenge: to interpret such massive expression data (5–7). The sizes of the datasets and their complexity call for multivariant clustering techniques (8, 9), which are essential for extracting correlated patterns and the natural classes present in a set of *N objects*, represented as points in the multidimensional space defined by $D$ measured *features*.

Gene microarray data are fairly special in that it makes good sense to perform clustering analysis in two ways (1, 2, 8). The first views the $n_s$ *samples* as the $N = n_s$ objects to be clustered, with the $n_g$ genes' levels of expression playing the role of the features, representing each sample as a point in a $D = n_g$-dimensional space. The different phases of a cellular process emerge from grouping samples with similar or related expression profiles. The other, not less natural, way looks for clusters of genes that act correlatively on the different samples. This view considers the $N = n_g$ *genes* as the objects to be clustered, each represented by its expression profile, as measured over all of the samples, as a point in a $D = n_s$-dimensional space.

In previous work (1, 2, 10), samples and genes were clustered completely independently; here we introduce and perform a *coupled* two-way clustering (CTWC) analysis (8).‡

Our philosophy is to narrow down both the features that we use and the data points that are clustered. We believe that only a small subset of the genes participate in any cellular process of interest, which takes place only in a subset of the samples; by focusing on small subsets, we lower the noise induced by the other samples and genes. We look for pairs of a relatively small subset $\mathcal{F}_i$ of features (either genes or samples) and of objects $\mathcal{O}_j$, (samples or genes), such that when the objects in $\mathcal{O}_j$ are represented using only the features from $\mathcal{F}_i$, clustering yields stable and significant partitions. Finding such pairs of subsets, $(\mathcal{O}_j, \mathcal{F}_i)$, is computationally hard; the CTWC method produces such pairs in an iterative clustering process.

CTWC can be performed with any clustering algorithm. We tested CTWC in conjunction with several clustering methods, but present here only results that were obtained by using the *superparamagnetic* clustering algorithm (SPC) (11, 12), which is especially suitable for gene microarray data analysis (13) because of its robustness against noise and its "natural" ability to identify stable clusters. By "stable" we mean those clusters that are statistically significant according to some criterion (see below).

CTWC was applied to two data sets, one from an experiment on colon cancer (1) and the other on leukemia (3). From both datasets we were able to "mine" partitions and correlations that have not been obtained in an unsupervised fashion by previously used methods. Some of these new partitions have clear well-understood biological interpretation. We do not report here discoveries of biologically relevant, previously unknown results. The main point of our message is twofold: (*i*) we were able to identify biologically relevant partitions in an unsupervised way, and (*ii*) other, not less natural, partitions were also found (http://www.weizmann.ac.il/physics/complex/compphys), which may contain new important information and for which one should seek biological interpretation.

## CTWC

**Motivation and Algorithm.** The results of every gene microarray experiment are organized in an *expression level matrix* $\mathcal{A}$. A row of this matrix corresponds to a single gene, while each column represents a particular sample. In a typical experiment simultaneous expression levels of thousands of genes are measured. Gene expression is influenced by the cell type, cell phase, external signals, and more (14). The expression level matrix is therefore the result of all these processes mixed together. Our goal is to separate and identify these processes and to extract as much information as possible about them. The main difficulty is that each biological process on which we wish to focus may involve a relatively small subset of the genes; the large majority of those present on the microarray constitute a noisy background that may mask the effect of the small subset. The same may happen with respect to samples. A straightforward approach to finding *pairs* of subsets, $(\mathcal{O}_j, \mathcal{F}_i)$, that lead to "meaningful" (see above) clusters, could be to take all possible submatrices of the original data and apply the standard (uncoupled) two-way clustering procedure to every one of them. By keeping track of all stable clusters that are formed in this process, and storing the identity of both genes and samples that define the particular submatrix, one is guaranteed to find every possible stable

---

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: CTWC, coupled two-way clustering; SPC, superparamagnetic clustering; SOM, self-organizing map; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia.

*To whom reprint requests should be addressed. E-mail: fedomany@wicc.weizmann.ac.il.

†By "sample" we refer to any kind of living matter that is being tested—e.g., different tissues (1), cell populations collected at different times (2), etc.

‡Hartigan (8) performed a coupled reordering of the rows and columns of a matrix, to identify submatrices that fit best a particular expected model. The aim of CTWC is in some sense the opposite; to identify submatrices that reveal some *unknown* structure.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.210134797. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.210134797

partition in the data. This approach is, of course, impossible to implement, because the number of such submatrices grows exponentially with the size of the problem. CTWC provides an efficient heuristic to generate such pairs of object and feature subsets by an iterative process that severely restricts the possible candidates for such subsets; we consider and test only those submatrices whose rows (columns) belong to genes (samples) that were identified (in a previous iteration!) as a stable cluster.

The iterative process is initialized with the full matrix—i.e., the sets of all genes ($g^0$) and of all samples ($s^0$) are used as (both) features and objects, to perform standard two-way clustering. Denote by $g_i^1$ and $s_j^1$ stable clusters of genes and samples found in this first step.

Every pair ($g_i^n$, $s_j^m$) (made of clusters obtained so far, $n, m = 0, 1$) defines a submatrix of the expression data; for every such submatrix we perform two-way clustering. The resulting stable gene (or sample) clusters are denoted by $g_k^2$ (or $s_l^2$). Each cluster is stored in one of two "registers of stable clusters"; gene clusters in register $\mathcal{G}$ and sample clusters in $\mathcal{S}$. Together with each new cluster we also store pointers that identify the pair of "parent" clusters ($g_i^n$, $s_j^m$) that were used as the object and feature sets in the clustering process that generated it. These steps are iterated further, using pairs of all previously found clusters. We make sure that every pair is treated only once; the process is terminated when no new clusters that satisfy some criteria (such as stability, critical size, or the criterion used in ref. 8) are found (unpublished work).

**Analyzing the Clusters Obtained by CTWC.** The output of CTWC has two important components. First, it provides a broad list of gene and sample clusters. Second, for each cluster (of samples, say) we know which subset (of samples) was clustered to find it, and which features (genes) were used to represent it. We also know for every cluster, say $s$, which other clusters can be identified by using $s$ as the feature set. We present here some of the possible ways one can use this kind of information. Particular implementations are described in *Applications*.

*Identifying genes that partition the samples according to a known classification.* This is a supervised test of clusters that were obtained in an unsupervised way. Denote by $C$ a known classification of the samples, say into two classes, $c_1$ and $c_2$. CTWC provides an easy way to rank the clusters of genes in $\mathcal{G}$ by their ability to separate the samples according to $C$.

First we evaluate for each cluster of samples $s$ in $\mathcal{G}$ two scores, *purity* and *efficiency*, which reflect the extent to which assignment of the samples to $s$ corresponds to the classification $C$. These figures of merit are defined (for $c_1$, say) as

$$\text{purity}\ (s|c_1) = \frac{|s \cap c_1|}{|s|};\ \text{efficiency}\ (s|c_1) = \frac{|s \cap c_1|}{|c_1|}.$$

Once a cluster $s$ with high purity and efficiency has been found, we can use the saved pointers to read off the cluster (or clusters) of genes that were used as the feature set to yield $s$ in our clustering procedure. Clustering, being unsupervised, as opposed to classification, discovers only those partitions of the data that are, in some sense, "natural." Hence by this method we identify the most natural group of genes that can be used to induce a desired classification.

One can test a gene cluster $g$ that was provided by CTWC also by more standard statistics, such as the $t$ test (15) or the Jensen–Shannon distance (16). Both compare the expression levels of the genes of $g$ on the two classes of samples, $c_1$ and $c_2$. Alternatively, one can also use the genes of $g$ to train a classifier to separate the samples according to $C$ (3) and use the success of the classifier to measure the relevance of the genes in $g$ to the classification.

*Discovering new partitions.* The members of every cluster $s$ have been linked to each other and separated from the other samples on the basis of the expression levels of some coexpressed subset of genes. It is reasonable therefore to argue that the cluster $s$ has been formed for some biological or experimental reason.

As a first step to understand the reason for the formation of a robust cluster $s$, one should try to relate it to some previously known classification (for example, in terms of purity and efficiency). Clusters that cannot be associated with any known classification have to be inspected more carefully. Useful hints for the meaning of such a cluster of samples may come from the identity of the cluster of genes that was used to find it. Similarly, sample clusters can be used to interpret clusters of genes that were not previously known to belong to the same process.

*CTWC is a sensitive tool to identify subpartitions.* Sample clusters that emerged from clustering a subset $s$ of the samples reflect a subpartition of $s$. When clustering the full sample set, this subpartition may be missed.

*CTWC reveals conditional correlations among genes.* The CTWC method collects stable gene clusters in $\mathcal{G}$. In many cases the same groups of genes may be added to $\mathcal{G}$ more than once. This is caused by the fact that some genes are coregulated in all cells, and therefore are clustered together, no matter which subset of the samples is used as the feature set. For example, ribosomal proteins are expected to be assigned to the same cluster for any set of samples that is not unreasonably small.

Some gene clusters, however, are different; they are coregulated only in a specific subset of samples. We call this situation conditional correlation. The identity of the sample cluster that reveals the conditionally correlated gene cluster is clearly important to understand the biological process that makes these genes correlated.

All of the features listed above were tested on artificially generated expression data into which correlations, partitions, and subpartitions were incorporated and masked. CTWC successfully unraveled all of the hidden structure from these "toy data" (17).

## Clustering Method, Statistical Significance, and Similarity Measures

Any reasonable clustering method can be used within the framework of CTWC. The optimal algorithm for analysis of gene expression data should have the following properties: the number of clusters should be determined by the algorithm itself and not externally prescribed [as is done when using self-organizing maps (SOMs) and K-means]; stability against noise; generating a hierarchy (dendrogram) and providing a mechanism to identify in it robust stable clusters; and ability to identify a dense set of points, which form a cloud of an irregular nonspherical shape, as a cluster. SPC, a hierarchical clustering method recently introduced by Blatt *et al.* (11), is the algorithm that best fits these requirements. The intuition that led to it is based on an analogy to the physics of inhomogeneous ferromagnets. Full details of the algorithm and the underlying philosophy are given in refs. 12 and 18.

The input for SPC is a distance or similarity matrix $d_{ij}$ between the objects $\mathbb{O}$, calculated according to the feature set $\mathcal{F}$. A tunable parameter $T$ ("temperature") controls the resolution of the performed clustering. One starts at $T = 0$, with a single cluster that contains all the objects. As $T$ increases, phase transitions take place, and this cluster breaks into several subclusters that reflect the structure of the data. Clusters keep breaking up as $T$ is further increased, until at high enough values of $T$ each object forms its own cluster. As opposed to most agglomerative algorithms, SPC has a natural measure for the relative stability of any particular cluster: the range of temperatures, $\Delta T$, over which the cluster remains unchanged. The more

stable a cluster is, the larger the range $\Delta T$ through which it is expected to "survive." For a stable cluster $s$, the corresponding $\Delta T_s$ constitutes a significant fraction of $T_{max}$, the temperature at which the data break into single-point clusters. Inspection of the gene dendrograms of Fig. 3 reveals stable clusters and stable branches.

In this work we chose the value of $\Delta T_c$, above which a cluster is considered as stable, in the following way. We permuted at random elements of the expression matrix under investigation, and applied SPC to the randomized matrix. $\Delta T_c$ was selected so that for 500 different random permutations no clusters that survived for $\Delta T > \Delta T_c$ were found. This gives a bound on the probability that clusters that we labeled as stable were in fact an artifact of noisy data.

**Normalization of the Gene Expression Array.** The Pearson correlation is commonly used as the similarity measure between genes or samples (1, 2, 19). This measure conforms with the intuitive biological notion of what it means for two genes to be coexpressed; it captures similarity of the "shapes" of two expression profiles, and ignores differences between their magnitudes (2). The correlation coefficient is high between two genes that are affected by the same process, even if each has a different gain due to the process, over different background expression levels (caused by other processes). Note, however, that a positive correlation between two highly expressed genes is much more significant than the same value between two poorly expressed genes. By using correlations one ignores this dependence of the reliability on the absolute expression level.

As to samples, correlations do not always capture their similarity. Consider two samples, taken at different stages of some process, with the absolute expression levels of a family of genes much below average in one sample and much higher in the other. Even if the expression levels of the two samples over these genes are correlated, one would like to assign them to different clusters.

We therefore used the following normalization scheme. Denote by $\mathcal{B}$ the matrix of the raw data. $\mathcal{B}$ is an $n_g \times n_s$ matrix, where $n_g$ is the number of genes and $n_s$ is the number of samples.

We normalize $\mathcal{B}$ in two steps. First, divide each *column* by its mean: $\mathcal{B}'_{ij} = \mathcal{B}_{ij}/\bar{\mathcal{B}}_j$; $\bar{\mathcal{B}}_j = (1/n_g)\Sigma_{i=1}^{n_g} \mathcal{B}_{ij}$, and then normalize each row, such that its mean vanishes and its norm is one:

$$\mathcal{A}_{ij} = \frac{\mathcal{B}'_{ij} - \bar{\mathcal{B}}'_i}{\|\mathcal{B}'_i\|},$$

where $\bar{\mathcal{B}}'_i = (1/n_s)\Sigma_{j=1}^{n_s}\mathcal{B}'_{ij}$ and $\|\mathcal{B}'_i\|^2 = \Sigma_{j=1}^{n_s} (\mathcal{B}'_{ij} - \bar{\mathcal{B}}'_i)^2$.

For both genes and samples we used the Euclidean distance as the dissimilarity measure. For two genes (rows of $\mathcal{A}$) the Euclidean distance is closely related to their Pearson correlation.

## Applications

We applied CTWC to data from two experiments. Here we report only the results that were obtained by CTWC and could not be found by using a straightforward clustering analysis. We highlight a small subset of the partitions that we were able to extract from the data and for which satisfactory biological explanation was found. We do *not* report here new discoveries of biologically relevant, previously unknown results. Rather, we claim to have discovered a method that is capable to mine such information out of the available data. New, relevant information may be contained in the new partitions that were found, to which we were not yet able to assign biological meaning. Some new, uninterpreted results are also reviewed briefly; full lists of the corresponding clusters and their constituent samples or genes can be found at our website (http://www.weizmann.ac.il/physics/complex/compphys).

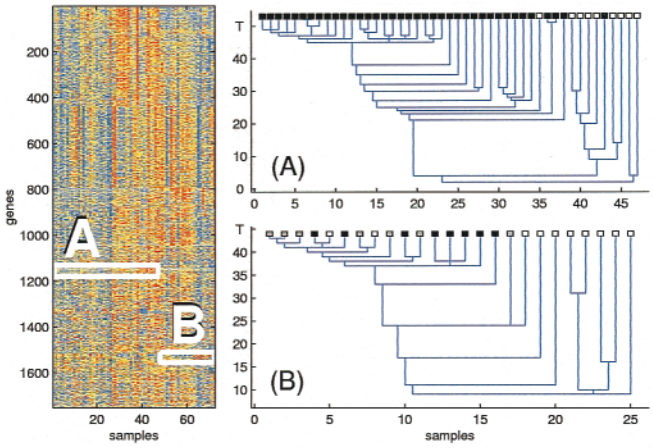



**Fig. 1.** The expression level matrix of the leukemia experiment is shown on the *Left*. Rows correspond to different genes, ordered by clustering them using all of the samples. The two boxes contain expression data from ALL patients (A) measured on one gene cluster and AML patients (B), on another gene cluster. On the *Right*, clustering the ALL samples, using the data in box A, yields good separation between T cell ALL (black) and B cell ALL (white). Clustering of AML samples, using the data in box B, yields a stable cluster, which contains all patients who were treated, with results known to be either success (black) or failure (gray). The vertical axis is the "temperature" parameter *T*, and on the horizontal axis the samples are ordered according to the dendrogram.

**Analysis of Leukemia Samples.** Golub *et al.* (3) obtained data from 72 samples collected from acute leukemia patients at the time of diagnosis. Forty-seven cases were diagnosed as acute lymphoblastic leukemia (ALL) and the other 25, as acute myeloid leukemia (AML). RNA prepared from the bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix (Santa Clara, CA), containing 6,817 human genes.

After rescaling the data in the manner described in ref. 3, we selected only those genes whose minimal expression over all samples is greater than 20. Only 1,753 genes survived this thresholding operation (http://www.weizmann.ac.il/physics/complex/compphys). The resulting array was then normalized as described above, to give a 1753 × 72 expression level matrix $\mathcal{A}$ (see Fig. 1).

Two iterations of CTWC sufficed to converge to 49 stable gene clusters (LG1–49) and 35 stable sample clusters (http://www.weizmann.ac.il/physics/complex/compphys) (LS1–35). We highlight here four of our findings, which demonstrate the power of the method to solve problems listed above.

*Identifying genes that partition the samples according to a known classification.* First we use the known ALL/AML classification of the samples to determine which gene clusters can distinguish between the two classes. We found a single cluster (LG1) of 60 genes that, when used as the feature set, induces a stable separation of the samples into AML/ALL clusters. (A cluster is identified with a certain class if both its purity and efficiency exceed 3/4.) This finding demonstrates the idea behind CTWC and its power. When SPC was applied, using the entire set of 1,753 genes, we did not find robust clusters that could be identified as AML or ALL tissues. Apparently, the two clouds of points in the 1,753-dimensional space, which contain the two groups of tissues, are displaced relative to each other, but they do have a region of overlap—the data in fact form a single cloud! In such a case SPC will not identify ALL and AML as separate clusters. Using only the genes of LG1 apparently eliminates this overlap of the points.

In such a situation, methods such as K-means and SOM may

assign two centroids to the data, so that proximity to these centroids can be used to characterize the two different kinds of tissues. In such cases, however, it is important to preset the number of clusters into which one wishes to break the data. Indeed, Golub *et al.* (3) showed that two-cluster SOM analysis on a subset of the data did separate the AML and ALL tissues in a robust manner. They also identified two groups, of 25 genes each, whose expression levels differ between these two clusters. Of the 25 genes that had higher expression levels in the AML patients, only 12 survived our thresholding and were included in our set of 1,753. Of these 12 genes, 5 indeed reside in our separating cluster LG1.

*Discovering new partitions.* Next, we search the stable sample clusters for unknown partitions of the samples. We focus our attention on sample clusters that were repeatedly found to be stable. One such cluster, denoted LS1, may be of interest; it includes 37 samples and was found to be stable when either a cluster of 27 genes (LG2) or another unrelated cluster of 36 genes (LG3) was used to provide the features. LG3 includes many genes that participate in the glycolysis pathway. Because of lack of additional information about the patients we cannot determine the biological origin of the formation of this sample cluster.

*Identifying subpartitions.* Using a 28-gene cluster (LG4) as features, we tried to cluster only the samples that were identified as AML patients (leaving out ALL samples). A stable cluster, LS2, of 16 samples was found (see Fig. 1, box B); it contains most of the samples (14/15) that were taken from patients that underwent treatment (with known results—success or failure). For none of the other AML patients was any information about treatment available in the data. Some of the 16 genes of this cluster, LG4, are ribosomal proteins and some others are related to cell growth. Apparently these genes can partition the AML patients according to whether they did or did not undergo treatment.

This result demonstrates a possible diagnostic use of the CTWC approach; one can identify different responses to treatment, and the groups of genes to be used as the appropriate probe.

We repeated the same procedure, but discarding AML and keeping only the ALL samples. We discovered that when any one of five different gene clusters (LG4–8) are used to provide the features, the ALL samples break into two stable clusters; LS5, which consists mostly of T cell ALL patients and LS4, which contains mostly B cell ALL patients (see Fig. 1, box A). When all of the genes were used to cluster all samples, no such clear separation into T cell ALL vs. B cell ALL was observed. One of the gene clusters used, LG5, with T/B separating ability, contains 29 genes, many of which are T cell related. Another gene cluster, LG6, which also gave rise to T/B differentiation, contains many HLA histocompatibility genes.

It is important to understand the difference between our results and those of ref. 3, where Golub *et al.* applied the SOM algorithm to a subset of 38 mixed AML and ALL samples. The number of desired clusters $K$ has to be used as an input to SOM. Setting $K = 2$, Golub *et al.* report finding AML/ALL separation; results for $K = 3$ were not reported; for $K = 4$ the clusters were identified as a single AML cluster, a T cell ALL cluster and two B cell ALL clusters. Our method, on the other hand, is completely unsupervised; it identified the T cell ALL/B cell ALL as a robust partition of the ALL samples, and also revealed that the genes that induce this partition are connected to the immune system.

These results demonstrate how CTWC can be used to characterize different types of cancer. Imagine that the nature of the subclassification of ALL had not been known. On the basis of our results we could predict that there are two distinct subclasses of ALL; moreover, by the fact that many genes that induce separation into these subclasses are either T-cell-related or HLA genes, one could suspect that these subclasses were immunology related.

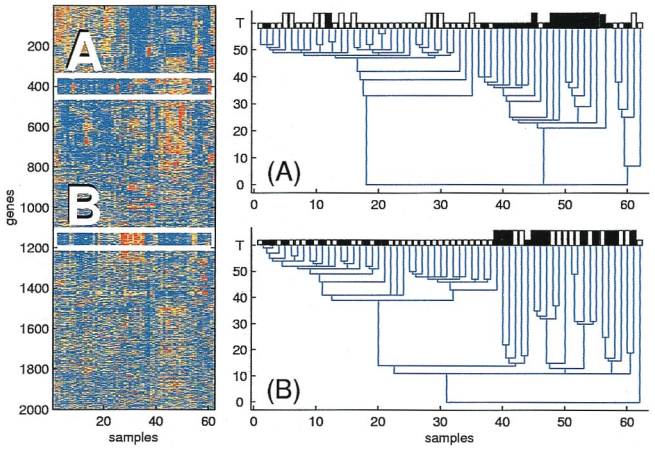



**Fig. 2.** The expression level matrix of the colon experiment is shown on the *Left*. Rows correspond to different genes, ordered by clustering them using all of the samples. The two boxes contain expression data of all samples for two gene clusters. On the *Right*, when the genes of the first cluster (A) are used, clear separation between tumor samples (white) and normal ones (black) is obtained. Another separation of the samples is obtained by using the second gene cluster (B). This separation is consistent with two distinct experimental protocols, denoted by short and long bars. The vertical axis is the "temperature" parameter $T$ and on the horizontal axis the samples are ordered according to the dendrogram.

As a different possible use of our results, note that some of the genes in the T-cell-related gene cluster LG5 have no determined function, and may be candidates for new T cell genes. This assumption is supported both by the fact that these genes were found to be correlated with other T cell genes and by the fact that they support the differentiation between T cell ALL and B cell ALL.

**Analysis of Colon Cancer Data.** The data set we consider next contains 40 colon tumor samples and 22 normal colon samples, analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes and expressed sequence tags (ESTs). Following Alon *et al.* (1), we chose to work only with the 2,000 genes of greatest minimal expression over the samples. We normalized the data to get a $2000 \times 62$ expression level matrix $\mathcal{A}$.

CTWC was applied to this data set. Seventy-six stable sample clusters (CS1–76) and 97 stable gene clusters (CG1–97) were obtained (http://www.weizmann.ac.il/physics/complex/compphys) in two iterations. One of the latter was a cluster of ribosomal genes, similar to the one identified in ref. 1.

*Identifying genes that partition the samples according to a known classification.* Again we search first for gene clusters that differentiate the samples according to the known normal/tumor classification. We found four gene clusters (CG1–4) that partition the samples this way (CG4 contains CG1). The genes of these clusters can be used if one wishes to construct a classifier for diagnosis purposes (see Fig. 2, box A). Alon *et al.* (1) calculated a muscle index, which can distinguish normal from tumor tissues. Of the 17 smooth muscle genes that contributed to their index, only 4 were included among the 2,000 that we used in our analysis. All of these were included in CG1 (and CG4).

*Discovering new partitions.* Five clusters of genes (CG2, CG4-CG7) generated very stable clusters of samples. Two of the five (CG2 and CG4) differentiated tumor and normal; two others were less interesting because the clusters they generated con-

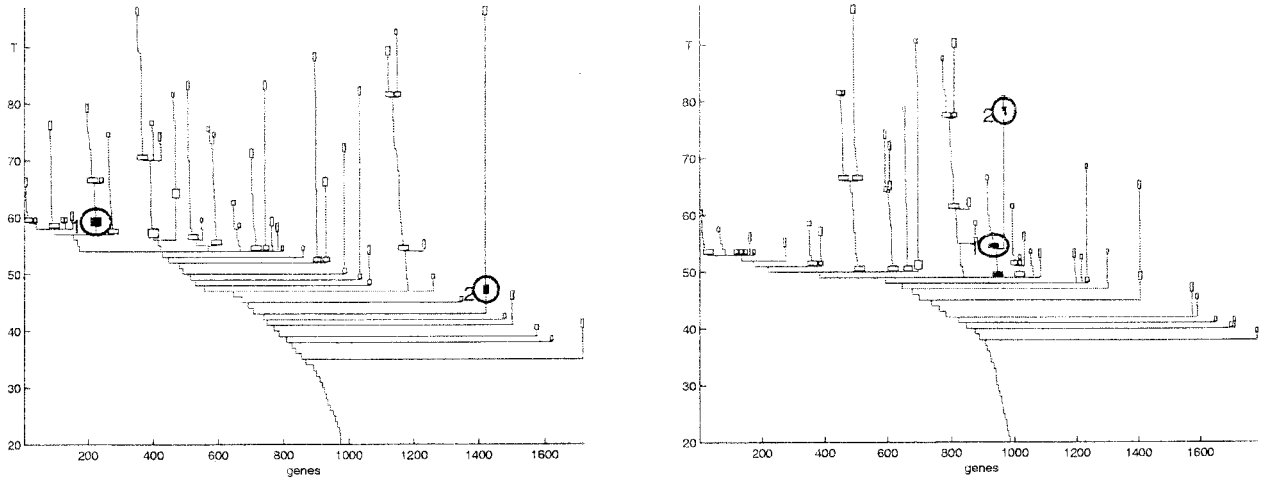



**Fig. 3.** Clustering genes of the colon cancer experiment, using all samples (*Left*) and using only tumor samples (*Right*) as the feature sets. Each node of this dendrogram represents a cluster; only clusters of size larger than 9 genes are shown. The last such clusters of each branch, as well as nonterminal clusters that were selected for presentation and analysis, are shown as boxes. In each dendrogram the genes are ordered according to the corresponding cluster analysis. The two circled clusters of the *Left* dendrogram are reproduced also in the *Right*, but there the two share a common ''parent'' in the tree. Note that the stability of a cluster is easily read off a dendrogram produced by the SPC algorithm.

tained most of the samples. The gene cluster CG5, however, gave rise to a clear partition of the samples into two clusters, of 39 and 23 tissues (see Fig. 2, box B). Checking with the experimentalists (U. Alon, K. Gish, D. Mack, and A. Levine, personal communication), we discovered that this separation coincides almost precisely with a change of the experimental protocol; 22 RNA samples were extracted by using a poly(A) detector (''protocol A''), and the other 40 samples were prepared by extracting total RNA from the cells (''protocol B''). Cursory examination did not yield any obvious common features among the 29 genes of the cluster CG5 that gave rise to this separation of the tissues.

*Identifying conditionally correlated genes and subpartitions.* Finally, we turn to identify conditionally correlated genes by comparing stable gene clusters formed when using different sample sets as features. We found that most gene clusters form irrespectively of the samples that are used. We did find, however, four special groups of genes (CG8–11) that formed clear and stable clusters when we used only the tumor samples as features, but were relatively uncorrelated—i.e., spread across the dendrogram of genes—when clustering was based on all of the samples or only the normal ones.

One of these four clusters (CG9), breaks up, at a higher resolution, into two subclusters, as shown in Fig. 3 *Right*. One of these subclusters (CG12), consists of 51 genes, all of which are related to cell growth (ribosomal proteins and elongation factors). The other subcluster (CG13), contains 17 genes, many of which are related to intestinal epithelial cells (e.g., mucin, cathepsin proteases). Interestingly, when the genes are clustered on the basis of either all samples or only the normal ones, both clusters (CG12 and CG13) appear as two uncorrelated distinct clusters, and their positions in the dendrogram are quite far from each other (Fig. 3).

The high correlation between growth genes and epithelial genes, observed in tumor tissue, suggests that it is the epithelial cells that are rapidly growing. In the normal samples there is smaller correlation, indicating that the expression of growth genes is not especially high in the normal epithelial cells. These results are consistent with the epithelial origin of colon tumor.

Two other groups of genes formed clusters only over the tumor cells. One (CG11, of 34 genes) is related to the immune system (HLA genes and immunoglobulin receptors). The second (CG10, of 62 genes) seems to be a concatenation of genes related to epithelial cells (endothelial growth factor and retinoic acid), and of muscle- and nerve-related genes. We could not find any common function for the genes in the fourth cluster (CG8).

Clustering the genes on the basis of their expression over only the normal samples revealed three gene clusters (CG14–16) that did not form when either the entire set of samples or the tumor tissues were used. Again, we could not find a clear common function for these genes. Each cluster contains genes that apparently take part in some process that takes place in normal cells, but is suppressed in tumor tissues.

## Summary and Discussion

We proposed a new method for analysis of gene microarray data. The main underlying idea of our method is to zero in on small subsets of the massive expression patterns obtained from thousands of genes for a large number of samples. A cellular process of interest may involve a relatively small subset of the genes in the dataset, and the process may take place only in a small number of samples. Hence when the full data set is analyzed, the ''signal'' of this process may be completely overwhelmed by the ''noise'' generated by the vast majority of unrelated data.

We are looking for a relatively small group of genes, which can be used as the features used to cluster a subset of the samples. Alternatively, we try to identify a subset of the samples that can be used in a similar way to identify genes with correlated expression levels. Identifying pairs of subsets of genes and samples that produce significant stable clusters in this way is a computationally complex task. We demonstrated that the CTWC technique provides an efficient method to produce such subgroups.

The CTWC algorithm provides a broad list of stable gene and sample clusters, together with various connections among them. This information can be used to perform the most important tasks in microarray data analysis, such as identification of cellular processes and the conditions for their activation, establishing connection between gene groups and biological processes, and finding partitions of known classes of samples into subgroups. CTWC is applicable with any reasonable choice of clustering algorithm, as long as it is capable of identifying stable clusters. In this work we reported results obtained by using the SPC algorithm, which is especially suitable for gene microarray

data analysis because of its robustness against noise, which is inherent in such experiments.

The power of the CTWC method was demonstrated on data obtained in two gene microarray experiments. In the first experiment the gene expression profile in bone marrow and peripheral blood cells of 72 leukemia patients was measured by using gene microarray technology. Our main results for these data were the following: (*i*) The connection between T-cell-related genes and the subclassification of the ALL samples, into T cell and B cell ALL, was revealed in an unsupervised fashion. (*ii*) We found a stable partition of the AML patients into two groups: those who were treated (with known results), and all others. This partition was revealed by a cluster of cell-growth-related genes. This observation may serve as a clue for a possible use of the CTWC method in understanding the effects of treatment.

The second experiment used gene microarray technology to probe the gene expression profile of 40 colon tumor samples and 22 normal colon tissues. Using CTWC, we find a different, less obvious, stable partition of the samples into two clusters. To find this partition, we had to use a subset of the genes. The new partition turned out to reflect two different experimental protocols. We deduce that the genes that gave rise to this partition of the samples are the ones that were sensitive to the change of protocol.

Another result that was obtained in an unsupervised manner by using CTWC is the connection between epithelial cells and the growth of cancer. When we looked at the expression profiles over only the tumor tissues, a cluster of cell growth genes was found to be highly correlated with epithelial genes. This correlation was absent when the normal tissues were used.

These features, discovered in data sets that were previously investigated by conventional clustering analysis, demonstrate the strength of CTWC. We find CTWC to be especially useful for gene microarray data analysis, but it may be a useful tool for investigating other kinds of data as well.

We thank N. Barkai for helpful discussions. The help provided by U. Alon in all stages of this work has been invaluable; he discussed with us his results at an early stage, provided us his data files, and shared generously his understanding and insights. This research was partially supported by the Germany–Israel Science Foundation (GIF).

1. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 6745–6750.
2. Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
3. 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., *et al*. (1999) *Science* **286,** 531–537.
4. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A.,, Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., *et al.* (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9212–9217.
5. Lander, E. (1999) *Nat. Genet.* **21,** 3–4.
6. Zhang, M. (1999) *Comput. Chem.* **23,** 233–250.
7. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999) *Nature (London)* **403,** 83–86.
8. Hartigan, J. (1975) *Clustering Algorithms* (Wiley, New York).
9. Kohonen, T. (1997) *Self-Organizing Maps* (Springer, Berlin).
10. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* (2000) *Nature (London)* **403,** 503–511.
11. Blatt, M., Wiseman, S. & Domany, E. (1996) *Phys. Rev. Lett.* **76,** 3251–3255.
12. Domany, E. (1999) *Physica A* **263,** 158–169.
13. Getz, G., Levine, E., Domany, E. & Zhang, M. (2000) *Physica A* **279,** 457–464.
14. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994) *Molecular Biology of the Cell* (Garland, New York).
15. Wadsworth, P. & Bryan J. (1960) *Introduction to Probability and Random Variables* (McGraw-Hill, New York).
16. Cover, T. & Thomas, J. (1991) *Elements of Information Theory* (Wiley-Interscience, New York).
17. Domany, E., Getz, G. & Levine, E. (2000) Israel Patent Appl. 134994.
18. Blatt, M., Wiseman, S. & Domany, E. (1997) *Neural Comput.* **9,** 1805–1842.
19. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10614–10619.

# Publication 2:

## Coupled Two-Way Clustering Server

Authors: G. Getz and E. Domany
Published in: *Bioinformatics* **19**, 1153–1154 (2003).

BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 9 2003, pages 1153–1154
DOI: 10.1093/bioinformatics/btg143

# Coupled two-way clustering server

*Gad Getz* and Eytan Domany*

*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*

Received on December 25, 2002; revised on November 4, 2002; accepted on November 6, 2002

**ABSTRACT**

**Summary:** The CTWC server provides access to the software, CTWC1.00, that implements **C**oupled **T**wo **W**ay **C**lustering (Getz *et al.*, 2000), a method designed to mine gene expression data.

**Availability:** Free, at http://ctwc.weizmann.ac.il.

**Contact:** ctwc_support@weizmann.ac.il

**Supplementary information:** The site has a link to an *example* which provides figures and detailed explanations.

A DNA chip experiment provides expression levels, $E_{gs}$, of thousands of genes $g$ for up to 100 samples $s$, summarized in an *expression table* of $\approx 10^6$ entries. Analysis of such data has several aims: (1) identify genes whose expression levels reflect biological processes of interest (such as development of cancer); (2) group the samples (e.g. tumors) into classes, possibly in a clinically relevant way, and (3) provide clues for the function of genes (proteins) of yet unknown role.

First one filters the genes (Alon *et al.*, 1999), leaving a set $G1$ to work with. Next, cluster all genes of $G1$ on the basis of their expression levels over the set of all samples, $S1$ [an operation denoted by $G1(S1)$], and cluster $S1$ using all the genes of $G1$ [$S1(G1)$]. In general, however, only a small subset of $N_r$ genes are relevant for one particular biological process of interest. Since usually $N_r \ll |G1|$, the 'signal' of these genes may be masked by the 'noise' generated by the (much more numerous) other genes. Furthermore, to assign samples into two clinically meaningful classes (e.g. adenoma and carcinoma), we may have to remove first a previously identified group of samples (e.g. healthy tissue), and cluster only the remaining $N_s' < N_s$ tumors. Thus one should analyze, one at a time, special *submatrices* of $E_{gs}$. CTWC (coupled two way clustering) is a heuristic, iterative method to search for informative $N_r \times N_s'$ submatrices among the exponentially many possible ones. In the first two steps, $G1(S1)$ and $S1(G1)$, we identify and *register* stable, statistically significant clusters of genes, $GI$ with $I = 2, 3, \ldots$ and of samples, $SJ$, $J = 2, 3, \ldots$. Next, we cluster every one of the stable sample groups $SJ$ (including $S1$), using the expression levels of every stable gene group $GI$, one at a time. Such a clustering operation, denoted by $SJ(GI)$, may generate new stable sample subgroups. Similarly, one reclusters every gene group $GI$ on the basis of every sample group $SJ$. New stable gene and sample clusters that emerge are added to the respective registers and used in the next iterative step, until the emergent new clusters are smaller than some preset threshold. A typical positive finding of the method is such a statement (Getz *et al.*, 2000): 'A particular group of samples $SJ$ (e.g. patients suffering from ALL leukemia) breaks into two clear subgroups (e.g. T-ALL and B-ALL) on the basis of the expression levels of a group of genes $GK$'.

CTWC uses as its 'clustering engine' an algorithm called superparamagnetic clustering (SPC) (Blatt *et al.*, 1996). SPC places in the same cluster objects that are 'close' to one another, producing a dendrogram, as a parameter $T$, that controls resolution, is varied. SPC is stable against addition of noise to the data and can identify irregular shaped clouds of points as clusters. Most importantly, SPC provides for each cluster a 'stability' index, whose value is indicative of the extent to which the cluster is 'real', and not due to noise in the data. The index is based on the physical intuition that underlies SPC; a stable cluster behaves as an independent 'ordered magnetic grain' for a wide range of values of $T$ (Blatt *et al.*, 1997). Using this index we exhaustively scan (and cluster, one at a time) those submatrices, whose genes and samples constitute stable clusters. CTWC has been used successfully to study data from experiments on colon cancer, leukemia (Getz *et al.*, 2000), breast cancer (Kela, 2002; Getz *et al.*, 2003), glioblastoma (Godard *et al.*, 2003), skin cancer (Dazard *et al.*, 2003) and antigen chips (Quintana *et al.*, 2003).

**THE SERVER** is frequently updated. Here we present a detailed, step by step 'roadmap' of the server, from data entry to viewing the results. We recommend that the instructions be read while viewing the *example (ES)* found at the CTWC site.

**Data preparation and Entry:** Filter the genes down to $|G1| < 3000$ (in our example we kept 2000 genes). The resulting matrix $E_{gs}$ is uploaded in the format used in `Cluster` (Eisen *et al.*, 1998), of an ASCII table separated by tabs (see *ES* links 1,2). Three optional preprocessing

*To whom correspondence should be addressed.

operations, can be performed after uploading the data matrix; *Scaling*, *Thresholding* and *Log* (see *ES* link 3, where only the first two were performed). Optionally the user may upload also a $P \times N_s$ table of $P$ 'predefined sets', whose entry $L_{is}$ can be 1/0/blank, indicating that sample $s$ belongs to set $i$/does-not-belong to set $i$/has unknown assignment (the example includes $P = 4$ categories; tumor,normal, protocols A and B—see *ES*, links 4,5). One can upload a similar table for genes.

**Creating Projects and Analyses:** Each user creates projects in his account. A *Project* is related to a dataset $E_{gs}$ and to two tables of predefined sets. Every project may contain several *Analyses*; each uses a particular set of running parameters. Within an analysis there are *processes*; each is a CTWC run defined by its initial gene and sample sets and the desired iteration depth.

*An analysis* is specified by its clustering parameters for SPC (try first the default values!), which are explained in the site. Here we mention only *Min T, Max T* and $\Delta T$, that govern the range and step size that specify the parameter $T$, which controls the resolution. At '*Min T*' there should be a single cluster, and at '*Max T*' many small clusters.

Another set of parameters, used by CTWC, define a stable cluster: (a) a '*minimal cluster size*' must be exceeded; (b) the number of cluster members lost, when $T$ increases by $\Delta T$, must be less than '*ignore dropout size*'; (c) conditions (a,b) must hold for at least '*stable delta T*' temperature steps. Clusters that qualify as stable are used in subsequent CTWC iterations.

**Execution of Analysis:** $G1/S1$ are the default for the initial gene/sample clusters used. In subsequent runs one can apply CTWC to a sub-matrix, defined either by a stable cluster that was found in a previous *Process*, or by one of the predefined sets. Specify the iteration depth of CTWC: try first *'depth'*=1 for samples and genes, performing $G1(S1)$ and $S1(G1)$. If the parameters gave suitable results, proceed to deeper levels (see *ES* link 8). Starting the analysis invokes a run; upon completion it generates output files and notifies the user by e-mail.

**Results:** Each execution generates *results* pages. The main one lists all stable gene ($GI$) and sample ($SJ$) clusters. Our example uses depth 1 for genes and 2 for samples (see *ES* link 9), showing for each stable cluster its stability index, the clustering operation in which it was found, and a table of all the clustering operations that were applied to it, and the clusters found by them.

Additional tables (for genes and for samples) relate stable clusters to the predefined labels. Each stable cluster is represented by a row and each predefined label by a column. The table element of cluster $Cx$ and set $Py$ contains the purity ($|Cx \cap Py|/|Cx|$) and efficiency ($|Cx \cap Py|/|Py|$) indices that measure the extent to which $Cx$ captures $Py$, and a score that measures the likelihood to obtain such overlap by chance. Significant overlaps are linked to the clustering operation that found them, allowing an easy search for clusters that capture known sets in the data. In the example $S7$ overlaps with normal samples and $S5$ with protocol B. The links show that $S7$ was found in $S1(G5)$ whereas $S5$ was identified in $S1(G4)$. This example demonstrates how different sets of genes (e.g. $G4$, $G5$) can yield very different separations of the samples ($S1$).

Links from the main *results* page point to two kinds of pages. (a) A cluster page contains a list of its members and whether they belong to predefined sets (see *ES* links 10,11). (b) A page describing a clustering operation, containing tables and figures (see *ES* links 12–14), such as a dendrogram, depicting hierarchical partitioning of the data, and the distance matrix, which shows the distances between the clustered objects (genes or samples), after reordering them according to the dendrogram.

## ACKNOWLEDGEMENTS

Funding for the server was provided by Yeda and the Levine Fund.

## REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Blatt,M., Wiseman,S. and Domany,E. (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.*, **76**, 3251–3254.

Blatt,M., Wiseman,S. and Domany,E. (1997) Data clustering using a model granular magnet. *Neural Comp.*, **9**, 1805–1842.

Dazard,J.-E., Gal,H., Amariglio,N., Rechavi,G., Domany,E. and Givol,D. (2003) Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer. *Oncogene*, in press.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Getz,G., Gal,H., Kela,I. and Domany,E. (2003) Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, **19**, 1079–1089.

Godard,S., Getz,G., Delorenzi,M., Kobayashi,H., Farmer,P., Nozaki,M., Diserens,A.-C., Hamou,M.-F., Dietrich,P.-Y., Regli,L. *et al.* Taxonomy and Classification of Human Astrocytic Gliomas on the basisof gene expression, submitted.

Kela,I. (2001) Clustering of gene expression data, M.Sc. Thesis, Weizmann Institute.

Quintana,F., Getz,G., Hed,G., Domany,E. and Cohen,I.R. (2003) Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity, submitted.

# Publication 3:

## Classification using semi-supervised typical cuts

Authors: G. Getz, N. Shental and E. Domany
*Preprint.*

# Classification using semi-supervised typical cuts

G. Getz[1], N. Shental[1], and E. Domany

December 28, 2003

## Abstract

In this work we introduce a novel semi-supervised algorithm which is an extension of the super-paramagnetic clustering method (SPC) [1] for the partially labelled case. Former work in the field of semi-supervised clustering classify the points according to the $k$-way cut. In contrast to these methods, we search for the typical cut which yields more accurate and robust results. Calculating the typical cut is performed using various sampling and approximation methods. Sampling is performed by several Monte-Carlo methods including the Multicanonical algorithm. We also use approximate inference methods adopted from the field of graphical models, such as Generalized Belief Propagation and suggest an extension to Belief Propagation. The performance of the different methods is evaluated on a two-dimensional data set.

## 1 Introduction to Semi-supervised learning

Situations which have many unlabelled points and a few with known labels call for semi-supervised learning methods. The goal of semi-supervised learning is to classify the unlabelled points, taking into account the assignment of labels to a subset of points and the distribution of the unlabelled ones. This problem is also known as clustering with partial labels. The partially labelled case can be considered intermediate between classification, where all the points of the training set are labelled, and clustering, in which there are no labels and only the distribution of the points is used. Such problems occur in many fields, in which obtaining data is cheap but labelling is expensive. Therefore, using supervised methods is impractical but, on the other hand, presence of a few labelled points can significantly improve the performance of unsupervised methods.

The basic assumption in unsupervised learning, *i.e.* clustering, is that points that belong to the same cluster actually originate from the same class. Density-based clustering methods define a cluster as a mode in the distribution, *i.e.* a relatively dense region

[1]These authors contributed equally.

surrounded by lower density regions. Hence each mode is assumed to consist of a single class; a certain class may, however, be dispersed over several modes. In case the modes are well separated, they can be easily identified by clustering methods and then classification can be performed as a postprocessing step, by assigning the points of each cluster according to the label of a single point in it. In cases when the modes are closer and the density between them is not significantly lower, unsupervised methods may encounter difficulties in separating the modes. In this case semi-supervised methods, which use a few labelled points, may be of help. Since points of different labels cannot be assigned to the same cluster, semi-supervised algorithms must place a border between them. Most probably the border will pass through the lower density regions that were difficult to identify without the information provided by the labelled points.

Throughout this paper we follow a toy-problem presented in Fig. 2. The points were generated from four Gaussians according to the distribution depicted in Fig. 2A, where each Gaussian corresponds to a class. We randomly sampled 400 points from this distribution which are shown in Fig. 2B. Knowing the distribution from which the points are sampled, we can calculate for each point the probability it originated from the four classes. The Bayesian separator, which corresponds to the optimal classification in this problem, is depicted as dashed lines in Fig. 2B. Two points in each class are labelled and are marked by circles. Note that in this example the modes are close to each other and density between them is not extremely low.

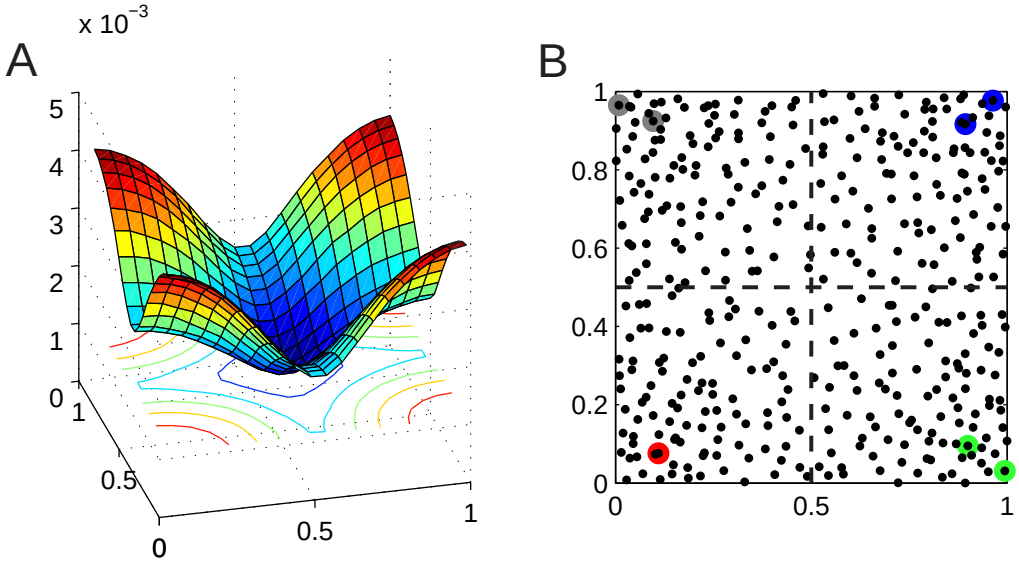



Figure 1: (A) The four Gaussians distribution from which the data was drawn. (B) The sample data with the labelled points colored according to their class.

Our semi-supervised algorithm is an extension of the Super-Paramangetic Clustering algorithm (SPC), which is a graph-based clustering method. We first describe related graph-based clustering methods and their shortcomings, and then turn to the semi-supervised

case.

## 1.1 Clustering by minimal and typical cuts

Graph-based clustering methods partition the data points, $\{\mathbf{x}_i\}_{i=1}^N$, by first constructing a weighted graph in which each vertex, $v_i$, represents a data point, $\mathbf{x}_i$, and edges are connected between neighboring points. The edge weights, $J_{ij} > 0$, are inversely related to the distance between the points. A clustering solution is a cut (partition) of the graph. Each cut is assigned a cost according to the weight of the edges it disconnects. Most previous work on graph-based methods perform clustering by searching for the partition of minimal cost.

Let $S = (s_1, \ldots, s_N)$ define a partition where points in each cluster are assigned the same value. The straightforward cost function, $E(S)$, is the sum of weights of the disconnected edges;

$$E(S) = \sum_{\langle i,j \rangle} J_{ij}(1 - \delta(s_i, s_j)) \ ; \tag{1}$$

this cost is known as the min-cut cost function. Using this cost, cutting the graph in dense regions (in which the edges have larger weights) has a much higher cost than a cut passing through sparse regions. Shi and Malik [2], pointed out that clustering using this cost function often produces trivial partitions in which single data points are separated. There are various ways to overcome this problem; Shi and Malik define a new cost function, termed *normalized-cut*, which takes into account the size of the separated clusters and penalizes generating small clusters.

Searching for the minimal cost solution has another disadvantage, common to all cost functions: it ignores the robustness of the found solution, *i.e.* how does the cost increase in the neighborhood of the solution. In a finite sample there are two types of fluctuations which may affect the mincut solution; a low density "crack" within a high density region, and the opposite case, where a high density "filament" appears in a low density region. Both types of fluctuations may drive the minimal cost solution far from the true and desired one – a crack may cut through a cluster and break it in two; on the other hand, or a filament may unify two clusters (of different class). Although the cut along a crack attains the minimal cost, slight deviations from it (which pass through high density regions) may increase the cost dramatically. On the other hand, the true cut passes through a "true" low density region and, therefore, deviations from it will have similar costs.

Blatt *et. al.* [1] use the min-cut cost function and perform clustering using a different approach which overcomes these problems. Instead of minimizing the cost, Blatt *et. al.* search for the "typical" cut – a typical cut is a weighted average of many cuts of similar costs. Taking a set of cuts into account increases the robustness of the solution, as many close cuts of similar cost can be preferred over the single, and isolated, cut of minimal value. Typical cuts have another advantage over searching for a cut with minimal cost, since they provide a "soft" solution to the clustering problem, as opposed to a "hard" solution. Typical cuts were introduced by Blatt *et. al.* in the development of SPC, using

the framework of statistical physics. Inspired by their work, Gdalyahu *et. al.* [3] suggested a probabilistic graph-partitioning approach for typical cuts and, recently, Shental *et. al.* [4] reformulated SPC in terms of graphical models.

## 1.2 Semi-supervised learning

A natural extension of clustering methods which minimize a cost function, in case partial labels are provided, is to search for the cut with minimal cost that complies with all label constraints. This is called the $k$-way cut problem. Unfortunately, finding the minimal cut in case there are more than two classes is NP-hard [5] and therefore one uses approximate methods. Boykov *et. al.* [6] suggested a method called *Graph Cuts* which can be applied to other types of costs as well, and heuristically minimizes the cost function even if many labeled points are given. Another method introduced by Zhu *et. al.* [7] minimizes a quadratic cost function based on Gaussian random fields and harmonic functions. Fig. **??**A shows the minimal cut solution of our toy problem.

In this work we join the two approaches and introduce a semi-supervised extension to typical-cut clustering methods, which calculate the typical cut under the constraints of labelled points. Fig. **??**B shows that the typical-cut solution for this problem is in nearly perfect agreement with the Bayesian separator.

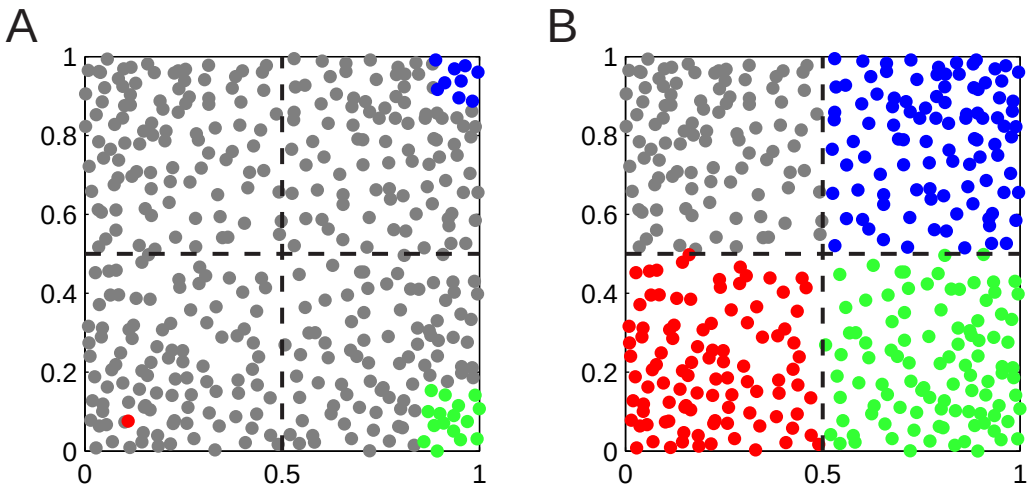



Figure 2: (A) The min-cut solution. (B) The Bayesian solution.

## 1.3 Paper outline

This paper is organized as follows: In the next sections we present the statistical physics view of typical cuts and its translation to the formalism of graphical models. Following that, we discuss the implication of adding labelled points to the SPC algorithm. In section

4 we present the algorithm, followed by a section describing its results. Section 6 discusses the difficulties and some directions for future research.

# 2 Relation to statistical physics

The SPC algorithm is based on the physical properties of a model granular ferromagnet. In this model, each point is assigned a "Potts spin" [8], $s_i = 1, \ldots, q$, which is a $q$-state random variable. Neighboring spins interact with ferromagnetic interactions, $J_{ij} > 0$ which decrease with distance (as defined in [1]). Every configuration of the system, $S = (s_1, \ldots, s_N)$, represents a particular partition of the graph. The cost of a configuration, called the energy in physics, is defined as in Equ. (1).

Statistical physics deals with calculating typical properties of a system by averaging over different configurations, each having its statistical weight, $P(S)$. The choice of weight function defines the statistical *ensemble* used to calculate averages. In the *microcanonical ensemble* all configurations with a certain value of the energy are assigned an equal probability,

$$P(S) = \begin{cases} 1/Z(E_0) & E(S) = E_0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the partition function $Z(E_0) = \sum_S \delta(E(S) - E_0)$ and $\delta(x)$ is the Dirac delta function. In order study the system at different costs, one has to vary the parameter $E_0$ over the range of possible costs and calculate average properties at each $E_0$. Microcanonical averages are difficult to calculate and often the microcanonical ensemble is replaced by the *canonical ensemble* in which the *average energy* is fixed, instead of the energy itself. Applying the maximum entropy principle [9] this leads to the Boltzmann distribution, in which $\beta = 1/T$, the inverse temperature, acts as a Lagrange multiplier;

$$P(S;T) = \frac{1}{Z(T)} \exp\left[-E(S)/T\right] \tag{3}$$

and $Z(T) = \sum_S \exp\left[-E(S)/T\right]$. The temperature, $T$, controls the average energy (in a one-to-one relation); low temperatures correspond to low energies. At $T = 0$ only the configuration with lowest energy has a non-zero probability, which corresponds to the mincut solution. As in the microcanonical case, the properties of system are studied at different values of $T$. For large systems with local interactions the two ensembles coincide and, therefore, for computational reasons one can use the canonical ensemble.

In order to calculate the "typical" solution, one has to define an observable, $O(S)$, which represents some attribute of the configuration and whose average value defines the typical property of the system. In our case, we are interested in two observables; the single spin statistics, $o_i^\alpha(S) = \delta(s_i, \alpha)$, and the pairwise spin statistics, $o_{ij}^{\alpha\beta}(S) = \delta(s_i, \alpha)\delta(s_j, \beta)$ where $i, j = 1, \ldots, N$ and $\alpha, \beta = 1, \ldots, q$.

The average of an observable $O(S)$, at temperature $T$, is calculated by

$$\langle O \rangle_T = \sum_S O(S) P(S;T) = \frac{1}{Z(T)} \sum_S \mathcal{O}(S) \exp\left[-E(S)/T\right] \ . \tag{4}$$

The average of the chosen observables is given by

$$b_i^\alpha(T) = P(s_i = \alpha; T) = \langle o_i^\alpha(S)\rangle_T = \frac{1}{Z(T)} \sum_S \delta(s_i, \alpha) \exp\left[-E(S)/T\right] \tag{5}$$

and the pairwise joint distributions of neighboring spins,

$$\begin{aligned} b_{ij}^{\alpha\beta}(T) &= P(s_i = \alpha, s_j = \beta; T) = \langle o_{ij}^{\alpha\beta}(S)\rangle_T = \\ &= \frac{1}{Z(T)} \sum_S \delta(s_i, \alpha)\delta(s_j, \beta) \exp\left[-E(S)/T\right] \ . \end{aligned} \tag{6}$$

In many physical systems, the temperature range can be divided into intervals, or *phases*, each of which has its own global properties. Granular-magnets are known to have three phases [10]; a low temperatures phase in which the system is ferromagnetic, *i.e.* most of the spins are assigned the same value; an intermediate temperature phase, called super-paramagnetic phase, in which spins that belong to the same grain are assigned the same value which varies among the grains; a high temperature phase in which the system is paramagnetic, *i.e.* the values assign to the spins are independent.

The field of statistical physics developed many powerful tools to study such systems, both analytical and computational. Since the number of configurations is exponential with the size of the system one can only compute approximations to the sums in Equations (5) and (6). A common approximation method is based on Monte-Carlo sampling, *e.g.* Metropolis [11]. SPC uses the Swendsen-Wang method [12] which is more suitable for sampling in models of granular magnets.

## 2.1 Equivalence to inference in a undirected graphical model

Equ. (3) can also be interpreted, using the terminology of graphical models, as an undirected graphical model. In this model, the joint distribution of the random variables, $S = \{s_i\}_{i=1}^N$, is described as

$$P(S; T) = \frac{1}{Z(T)} \prod_{\langle i,j\rangle} \psi_{ij}(s_i, s_j; T) \prod_i \phi_i(s_i) \tag{7}$$

where the product is taken over neighboring spins $\langle i, j\rangle$ as in Equ. (1),

$$\psi_{ij}(s_i, s_j; T) = \begin{cases} 1 & \text{for } \ s_i = s_j \\ \exp(-J_{ij}/T) & \text{for } \ s_i \neq s_j \end{cases} \tag{8}$$

and $Z(T)$ is as defined above. The $\phi_i(s_i)$ terms in Equ. (7) is called the local evidence and represent external fields. The cost function defined in Equ. (1) does not have external fields and, thus, all the $\phi_i(s_i) = 1$.

The field of graphical models provides ways to estimate marginal distributions over small sets of variables. Therefore, we can use this machinery as an additional procedure to estimate the average values of the observables defined in Equations (5) and (6).

The Belief Propagation (BP) introduces variables $m_{ij}(s_j)$ which can intuitively be understood as a message from spin $i$ to spin $j$ about the state spin $j$ should be in. The message $m_{ij}(s_j)$ is a vector of length $q$ where each component represents the tendency of spin $j$ to be in each of the state as seen by spin $i$. In accordance, the probability that spin $i$ is in a state $\alpha$, also known as its belief, $b_i^\alpha$, is proportional to all its incoming messages and the local evidence;

$$b_i^\alpha \propto \phi_i(\alpha) \prod_{j \in \mathcal{N}(i)} m_{ji}(\alpha) \tag{9}$$

where $\mathcal{N}(i)$ is the set of spins that are connected to spin $i$. Using the messages one can also calculate the pairwise statistics $b_{ij}^{\alpha\beta}$ by

$$b_{ij}^{\alpha\beta} \propto \psi_{ij}(\alpha,\beta)\phi_i(\alpha)\phi_j(\beta) \prod_{k \in \mathcal{N}(i)\setminus j} m_{ki}(\alpha) \prod_{l \in \mathcal{N}(j)\setminus i} m_{lj}(\beta) \ . \tag{10}$$

Searching for a self-consistent set of messages is performed by the following update rule:

$$m_{ij}(\beta) \longleftarrow \sum_{\alpha=1}^{q} \phi_i(\alpha)\psi_{ij}(\alpha,\beta) \prod_{k \in \mathcal{N}(i)\setminus j} m_{ki}(\alpha) \ . \tag{11}$$

Yedidia *et. al.* [13] show that the fixed points of the BP algorithm[2] correspond to minima of an approximation to the free energy, known in the physics literature as the Bethe approximation[3].

In a more recent work, Yedidia *et. al.* [14] extend the BP algorithm and define a family of methods called Generalized Belief Propagation (GBP). GBP is based passing messages between regions (groups of spins), in a similar way to BP which passes messages between single spins. The joint probability distribution of each region is calculated exactly. As the BP algorithm is linked to the Bethe approximation of the free energy, the GBP fixed points correspond to stationary points of a better approximation of the free energy called the Kikuchi approximation. The Kikuchi method approximates the free energy as a sum of local free energies defined over small regions of spins[4]. In case one selects the regions as pairs of spins the Kikuchi approximation reduces to the Bethe approximation.

# 3 The effect of labelled points

Labelled points are introduced in the SPC framework in a straightforward way, by fixing the spins of the labelled points at values that represent their type. We begin with an informal description of the way SPC performs clustering, which will help explain the effect

[2]The convergence of BP is not guaranteed. The proof assumes it converged.

[3]Yedidia *et. al.* show that the fixed points correspond to zeros of the free energy gradient which may also be saddle-points.

[4]The regions may overlap therefore correction terms must be introduced in order to correct for over counting.

of the labelled points. Consider points drawn from an infinite uniform distribution in a $D$ dimensional space. Such a system exhibits two phases; a ferromagnetic phase and a paramagnetic phase. In the ferromagnetic phase distant spins are likely to share the same value, whereas in the paramagnetic phase even close spins are nearly independent. It is convenient to define a length scale, $\xi(T)$ which measures the distance between points for which $\langle \delta(s_i, s_j) \rangle_T = 0.5$; spins closer than $\xi(T)$ have $\langle \delta(s_i, s_j) \rangle_T > 0.5$ and are called aligned. In the ferromagnetic phase at low $T$ all the spins are aligned, thus $\xi(T) = \infty$, while in the paramagnetic phase $\xi(T)$ decreases from some finite value at the transition temperature, to zero as the temperature is increased For a finite sample a pseudo transition between the phases occurs at $T$ for which $\xi(T)$ is on the order of the linear size of the system.

Now consider a more complicated scenario in which two regions of high density are separated by a low density region. Such a system exhibits three phases; ferromagnetic, super-paramagnetic and paramagnetic. In the ferromagnetic phase (low $T$), most of the spins in the system, in both the high- and low-density regions, are aligned. In the paramagnetic phase (high $T$), close spins, even in the high density regions, are nearly independent. In the super-paramagnetic phase (intermediate $T$), only the spins in the high-density regions are aligned within each high density region. Neglecting surface effects, this can be explained using two length scales; $\xi^H(T)$ for the high-density regions and $\xi^L(T)$ for the low densities. The temperature range of the super-paramagnetic phase is bound above by, $T_{sp}$, the temperature in which $\xi^H(T)$ corresponds to the size of the high density region and below by the temperature, $T_{fs}$, in which $\xi^L(T)$ is of the order of the distance between the two high density regions (which is also the size of the low density region).

Adding labelled points affects both transition temperatures and can be explained in terms of $\xi^L(T)$ and $\xi^H(T)$. We assume that each high density region originates from a different class; denote by $l^H$ the typical distance between two labelled points from the same region. The high temperature transition occurs at some $\tilde{T}_{sp}$ and we expect $\tilde{T}_{sp} > T_{sp}$ since as long as $\xi^H(\tilde{T}_{sp}) \gtrsim l^H$, the labelled points induce order on the high density region. At $\tilde{T}_{sp}$ we have $\xi^H(\tilde{T}_{sp}) \lesssim l^H$ and since $l^H$ is shorter than the region's size $\xi^H(\tilde{T}_{sp}) < \xi^H(T_{sp})$. The increase in the temperature at which the dense regions disorder can also be understood as an effective increase of the density caused by the labelled points.

The effect of low temperatures is more intricate. At $T_{fs}$ the two high density regions "feel" each other and, therefore, if they have different labels, a "border" must be formed between them. The location of the border depends on the specific properties of the system. The single ferromagnetic phase, observed for the unlabelled case, is replaced now by two possible phases, a Bayesian phase $B$ and a ferromagnetic phase, $F$. In the higher temperature phase, $B$, each class "penetrates" into the low density region and a border passes between them. As the temperature is lowered, a second phase, $F$, may appear in which one of the classes "overtakes" most of the system, except for small regions surrounding the labelled points of the other class (this phase appears only in case the mincut solution corresponds to phase $F$). This phase corresponds to the ferromagnetic phase in the unlabelled case. [5]

[5]The "ferromagnetic" phase itself may be divided into two. At a very low temperature only one of the

The main contribution of the labelled points is the formation of phase $B$. Both phase $B$ and the super-paramagnetic phase (of the unlabelled case) yield the correct classification of the points. In cases when the density difference between the regions is small and thus the super-paramagnetic phase is not observed over a significant temperature range, adding labelled points may reveal the correct classification.

## 3.1 Estimating the observables in the presence of labelled points

Estimating the observables in the unlabelled case can be performed by generating an ensemble of configurations using the Swendsen-Wang algorithm (SW) as in SPC, or can be approximated using Belief Propagation (BP) (or Generalized Belief Propagation - GBP). Presence of labelled points may complicate the estimation since they introduce "frustration" in the system, *i.e.* the configuration of minimal energy has unsatisfied bonds (the ones along the mincut border).

Since phase B yields the correct Bayesian classification, one must use a good estimation for the observables in this phase. There are several methods to perform the estimation. These methods are known to yield dissimilar results in the vicinity of the transition between phases B and F. The properties of the methods within phase B (far from its transition temperatures), in our toy problem, are discussed in Section **??**.

### 3.1.1 Swendsen-Wang (SW) with external fields

In order to introduce external fields into SW one can separate the energy into two terms; a pairwise term which is similar to the energy in the unlabelled case and a field term acting on the neighbors of the labelled points. According to Kandel and Domany [15], one can perform the regular SW step to build SW-blocks using the pairwise term, and then apply a regular Metropolis update step governed by the fields, to choose the state of the SW-blocks. This procedure, however, may not work well in phase B and particularly in the transition between phases F and B, since moves from a ferromagnetic state (in which most spins are of the same class) to the correct partition and back are very unlikely[6]

### 3.1.2 Metropolis with external fields

Introducing labelled points into the Metropolis algorithm is straightforward. As in other problems the Metropolis method suffers from long mixing time. Unbiased sampling requires

classes "overtakes" the whole system, which corresponds to the min-cut solution. At a higher temperature it may also be possible for other class-types to "overtake" the system, thus, an additional phase may be created. These phases are unimportant from our perspective since they do not provide a correct classification.

[6]Moves from the a ferromagnetic state to the correct partition require to delete the bonds along the correct partition. The probability for such an event is the same as in the unlabelled problem which is very low at this temperature; recall that phase A occurs at temperatures for which the unlabelled problem is in the ferromagnetic phase.

to move a domain-wall across the system. Since Metropolis performs single spin flips this may take unpractical times.

### 3.1.3 Belief Propagation (BP) and Generalized Belief Propagation (GBP) with external fields

Belief Propagation is an approximation method based on local message passing between spins (or local groups of spins, in the case of GBP). Fixing a labelled point $i$ to its known class $c_i$ is performed by introducing an infinite external field for state $c_i$. Hence, in the graphical-models notation $\phi_i(s_i = c_i) = 1$ and $\phi_i(s_i \neq c_i) = 0$. Near the transition there is a similar probability that system is in a ferromagnetic state (typical of phase F) or partitioned according to the typical cut of phase B. Such a distribution cannot be described by a product of local probabilities as in BP (and GBP). Therefore, the BP/GBP approximation around the transition temperature is poor.

### 3.1.4 Weighted BP with external fields

In order to overcome the problems encountered when using BP with external fields we suggest to extend the BP method by a mixture of several BP solutions which are weighed in a non-trivial manner. The obtained BP solutions have nearly zero overlap and represent globally different configurations. In Appendix C we present *Weighted BP* (WBP) and derive the optimal weighting coefficients.

### 3.1.5 Multicanonical Monte-Carlo with external fields

Muticanonical Monte-Carlo belongs to a group of algorithms called extended Monte-Carlo methods (see [16]) developed to correctly sample spin-glass system in which there are many frustrated bonds. Typically in system with frustrated bonds there are many different configurations with similar, low energies. These configuration reside in valleys surrounded by high energy barriers which must be crossed in order to achieve unbiased sampling. Extended Monte-Carlo methods sample from an extended distribution from which the Boltzmann distribution of Equ. (3) can be obtained by marginalizing the auxiliary variables. The extended distribution allows the system to move to higher energy configurations, and cross the energy barriers that way. The details of the Multicanonical method appear in Appendix A.

### 3.1.6 Which sampling method to use?

As the system is more frustrated, biased sampling or poor approximations are more likely to happen. For such a system, one needs to apply more complicated (*e.g.* Multicanonical) sampling methods. However, we do not know, at this point, of any systematic way of identifying the cases for which the simple methods suffice.

# 4 The classification algorithm

The raw output of the algorithm are the temperature dependent observables, the pairwise and single-point probabilities, as defined in Equations (5) and (6). Classification is performed at each value of the temperature. In order to perform hard classification one can simply assign each point to its most probable class;

$$c_i = \arg\max_{\alpha} b_i^{\alpha} \ . \tag{12}$$

This, however, ignores the difference between the most probable class and the next probable class. This difference, $\tau_i$, can serve as a measure of confidence in the classification. Points for which $\tau_i < \tau$, a user defined parameter, may either be left unassigned or treated in a way described below. These unassigned points can be of two kinds; (a) peripheral points of a labelled cluster which are far from the labelled points and hence the effect of the label on its class probabilities is small, and (b) points that belong to a cluster which has no labelled points, and thus may represent a new class. This provides the user with additional information: although points of type (b) do not belong to any of the classes with given labels, they still form a cluster which may correspond to a new type.

We suggest the following algorithm to classify the unassigned points, which can distinguish between the two kinds. The algorithm is based on the pairwise correlations, defined as $C_{ij}(T) = \sum_{\alpha} b_{ij}^{\alpha\alpha}(T)$. This statistic is used in SPC to generate the clusters. Each unassigned point $i$ is disconnected from its neighbors $j$ for which

$$C_{ij}(T) < \frac{1}{2}\left(1 + \frac{1}{q}\right) \tag{13}$$

(this value is the midpoint between perfect correlation and the random level $1/q$). Next, the classification of the points is performed using the resulting connected components: unassigned points (for which the margin between the two most probable classifications is too small) which belong to a connected component that does not contain any classified point are marked as a new class. Similarly, unassigned points belonging to a connected component which contains classified points of a single class are assigned to that class. Finally, if a connected component contains points which were assigned to different classes, the unassigned points are marked as "confused" between these classes.

# 5 Results on a toy problem

This section describes the classification results of our toy problem using our classification method (see Sec. 4). We compare the performance of the various sampling/approximation methods presented in Sec. 3.1. Figures 3 to 5 depict the number of misclassified points, where the ground truth is set according to the Bayesian classification rule. We also compare the algorithm to the unsupervised case, *i.e.* super-paramagnetic clustering of the same dataset without the labels. In this case, we perform the classification according to the

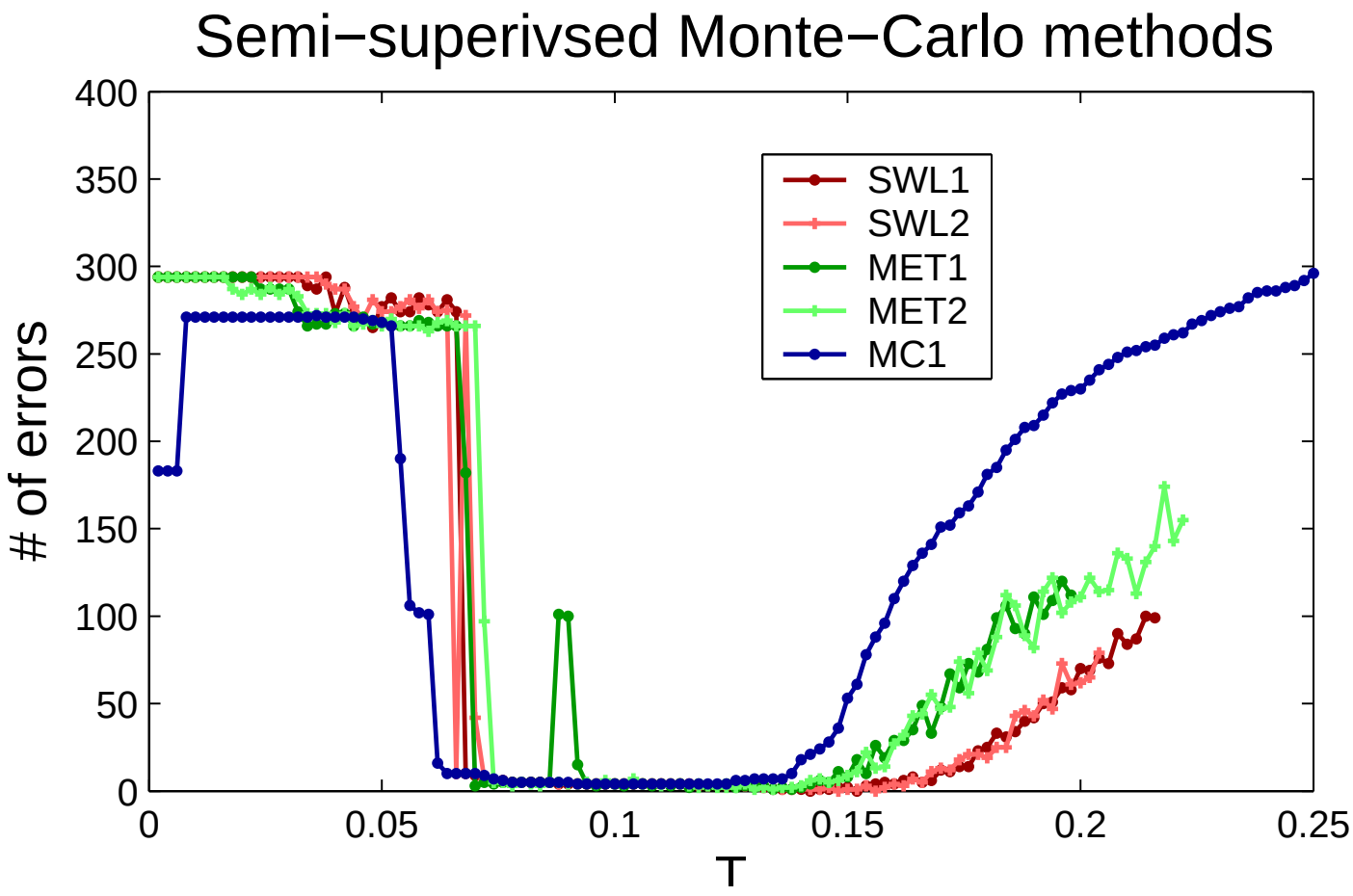



Figure 3: Error rates as a function of temperature for the Monte-Carlo methods: two runs of Swendsen-Wang with labelled points (SWL1, SWL2), two runs of Metropolis (MET1, MET2) and two runs of the Multicanonical Monte-Carlo method (MC1,MC2).

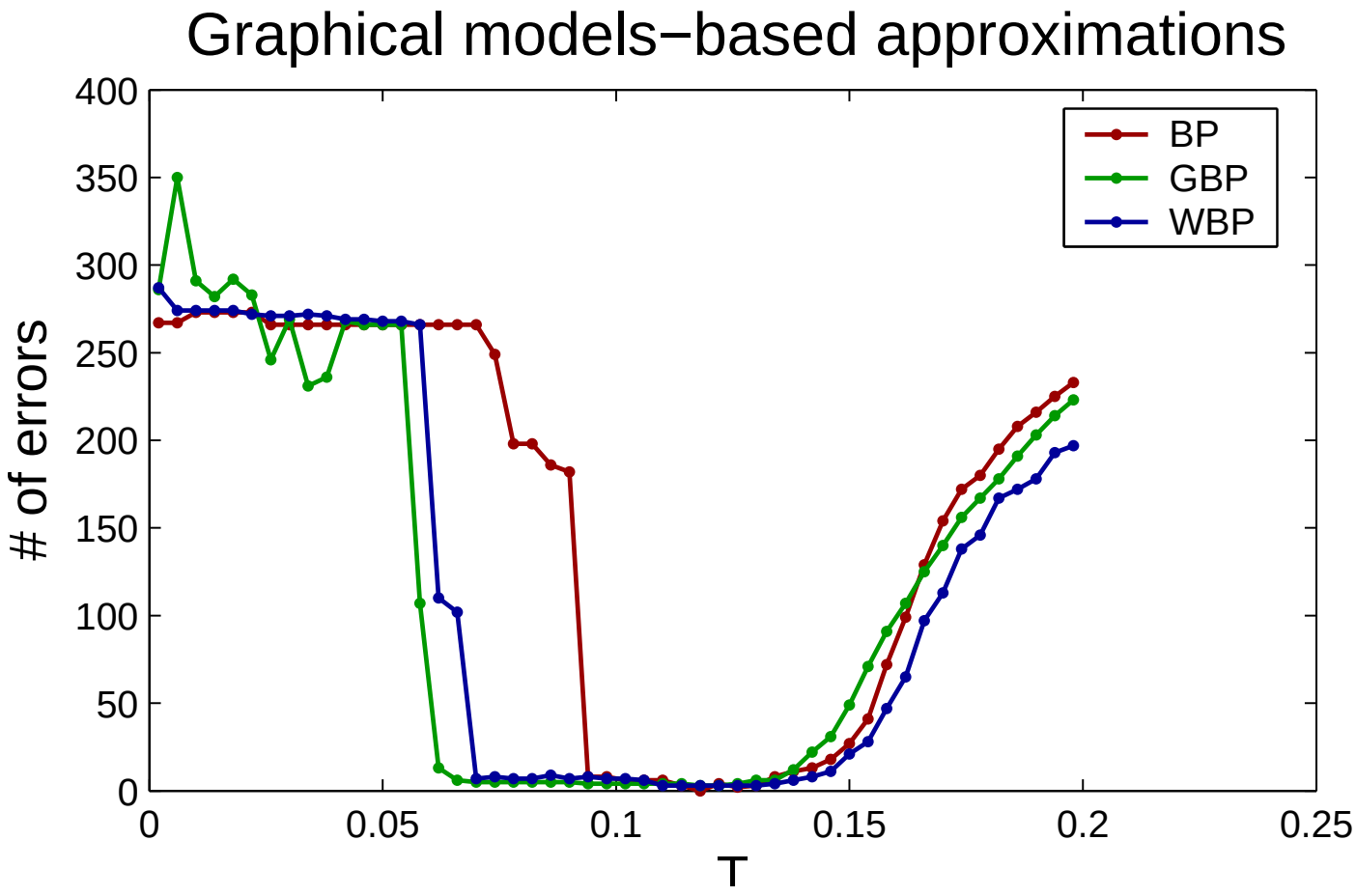



Figure 4: Error rates as a function of temperature for the graphical model-based methods: Belief Propagation (BP), Generalized Belief Propagation (GBP) and Weighted Belief Propagation (WBP).

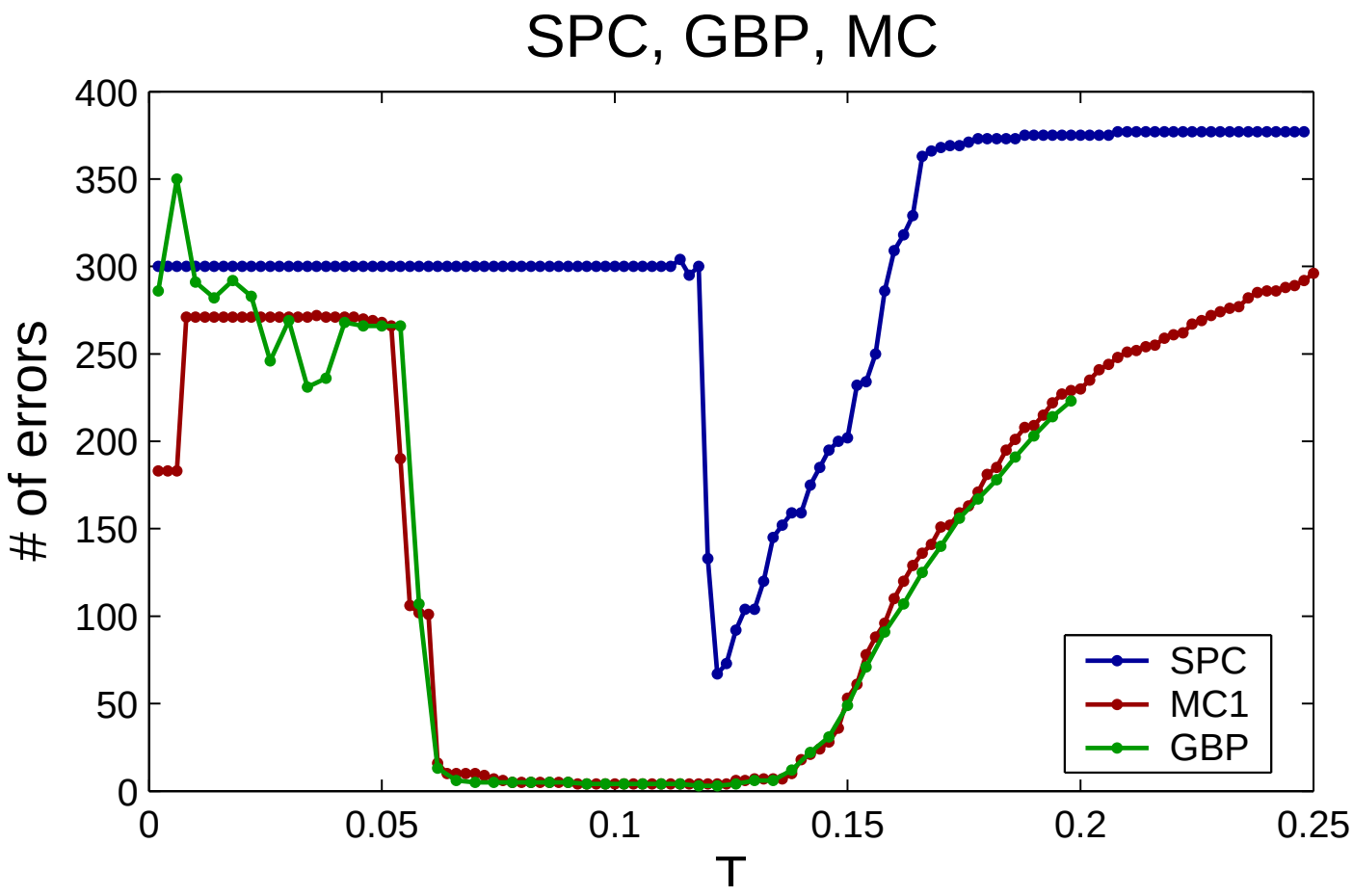



Figure 5: Error rates as a function of temperature for Super-Paramagnetic Clustering (SPC), Multicanonical Monte-Carlo method (MC) and Generalized Belief Propagation (GBP).

resulting clusters in the following way: we go over the class-types, $c = 1, \ldots, 4$, one-by-one, and identify the cluster, out of the yet unclassified ones, that contains most of the points of type $c$ (using the ground truth). All the points in that cluster are assigned to class $c$.

The most evident phenomena which demonstrate the effect of the labelled points, as seen in Figs. 3 to 5, is the lower error rate and the expansion of the temperature range at which these error rate are attained. Using only 7 labelled points, out of the 400, the error rate drops dramatically from a lowest value of about 70 points, found by SPC in a temperature interval of 0.002, to between 0 and 5 erroneous points across an interval of 0.15.

The figure also shows that, although the different sampling or approximation methods reach similar minimal error rates, their graphs are different. We discuss the differences among the Monte Carlo sampling methods and among the graphical-models based approximation methods, and then compare between the two.

## 5.1 Comparison among the Monte Carlo methods (SWL, METL, MCL)

Monte Carlo methods satisfy detailed balance which guarantees that for long enough running times they indeed converge to the desired distribution. Running for a finite time, however, does not guarantee this convergence and therefore we ran each method twice and compared the results[7]. Figures 3 to 5 show that in most temperatures both runs of all methods coincide. However, there is a temperature range, $[0.054, 0.076]$, in which the two runs of SWL and METL do not agree. In this temperature range both SWL and METL sample a confined subspace of configurations, hence producing the results are biased. The two runs of each method are restricted to different subspaces; In the first subspace the run is confined to a subspace where the typical configuration corresponds to phase B (the lower graphs), while in the second subspace the typical configuration corresponds to phase F (the upper graphs). In contrast to SWL and METL, MC succeeds to sample both subspaces and, therefore, produces a reproducible unbiased sample.

## 5.2 Comparison among the graphical-models methods (BP, GBP, WBP)

As seen in the figure, BP yields a good approximation for $T > 0.054$ but fails at lower temperatures. BP suffers from the same problem as METL and SWL as it is confined to the subspace that corresponds to phase F[8]. GBP and our WBP, on the other hand, produces similar results to MC. Below the transition, however, GBP fails to converge while WBP still coincides with the MC solution.

[7]The first run was performed by *increasing* the temperature and used the last configuration of the lower temperature as the initial configuration of the higher one. The second run *decreased* the temperature and used the last configuration of the higher temperature as the initial configuration of the lower one. This measures the hysteresis loop.

[8]The BP was initialized with uniform messages and user synchronous update scheme.

# 6 Discussion

In this paper we introduced an algorithm which extends super-paramagnetic clustering (SPC), a typical-cut clustering method, to the case when labels are provided for a subset of the data. This algorithm substantially improves over both unsupervised methods and recently introduced semi-supervised methods which seek for the minimal cut. By an example, we demonstrated that a small fraction of labelled points can dramatically reduce the number of errors, and increase the robustness of the classification. In order to calculate the typical cuts we use multicanonical Monte-Carlo which can correctly sample the non-trivial energy landscape of our problem. We also apply inference methods used in graphical models and introduce a novel extension to Belief Propagation.

The results of WBP, GBP and MC on our toy problem are similar. The differences between these methods may be expressed when analyzing larger data sets. The main advantage of using BP is that it gives a deterministic result which serves as good approximation. In our case, however, BP below the transition gives a poor approximation. Moreover, using a random asynchronous update scheme yields different results. WBP is a method to combine these different results and obtain much better approximations. However, WBP is no longer deterministic and requires long running times in order to capture all the different BP fixed points. The question – in which scenarios is WBP preferable over Monte-Carlo methods – is left for future work.

# References


[1] M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Phys. Rev. Lett.*, 76:3251–3254, 1996.

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of Computer Vision and Pattern Recognition*, 1997.

[3] Y. Gdalyahu, D. Weinshall, and M. Werman. Stochastic Image Segmentation by Typical Cuts, 1999.

[4] N. Shental, A. Zomet, T. Hertz, and Y. Weiss. Learning and inferring image segmentations with the gbp typical cut algorithm. In *Proceedings of International Conference on Computer Vision*, 2003.

[5] T. H. Cormen, C. L. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *International Conference in Computer Vision (ICCV)*, pages 377–384, 1999.

[7] X. Zhu, Z. Ghahramani, and J. Laferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, 2003.

[8] F. W. Wu. The Potts model. *Rev. Mod. Phys.*, 54:235, 1982.

[9] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[10] S. Wiseman, M. Blatt, and E. Domany. Super-paramagnetic clustering of data. *Phys. Rev. E*, 57:3767–3783, 1998.

[11] D. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, 2000.

[12] J.-S. Wang and R. H. Swendsen. Cluster Monte Carlo algorithms. *Physica A*, 167:565, 1990.

[13] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding Belief Propagation and its Generalizations. Technical report, Mitsubishi Electric Research Laboratories, 2002.

[14] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.

[15] D. Kandel and E. Domany. General cluster monte carlo dynamics. *Phys. Rev. B.*, 43:8539, 1991.

[16] Y. Iba. Extended ensemble monte carlo. *International Journal of Modern Physics C*, 12:623–656, 2001.

[17] B. A. Berg and T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267:249–253, 1991.

[18] B. A. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, 68:9–12, 1992.

[19] B. A. Berg and T. Celik. A New Approach to Spin Glass Simulations. *Phys. Rev. Lett.*, 69:2292–2295, 1992.

# A Appendix A: Typical cuts using Multicanonical Monte Carlo

Standard MCMC (Monte-Carlo Markov Chain) methods generate sample of configurations from the canonical distribution $\propto exp(-\mathcal{H}(\mathcal{S})/T)$, to estimate thermal averages of observables. Since the energy landscape in our case is ragged, at low temperatures such methods suffer from slow mixing time, *i.e.* they are confined to only a part of the configuration space which is a valley surrounded by high energy barriers. Therefore, other configurations which have the same energy but reside in other unexplored parts of the space would be missed.

In case the observable has a different value in the unexplored regions the estimation would be incorrect.

In order to overcome this problem, we apply an extended Monte Carlo method called *Multicanonical Monte Carlo* [17–19]. Instead of sampling from $\propto exp(-\mathcal{H}(\mathcal{S})/T)$ it generates a sample of configurations drawn from $\propto 1/D(E(\mathcal{S}))$ where $D(E)$ is the density of states defined by the number of different states within a given energy range

$$D(E)dE = \{\mathcal{S} : E < \mathcal{H}(\mathcal{S}) < E + dE\} \ . \tag{A.14}$$

Sampling from this distribution generates configurations so that each energy enters with equal probability since $P(E) \propto D(E)1/D(E) = 1$ – one would get approximately an equal number of configurations at each of the energies, including the lowest. Sampling from $1/D(E(\mathcal{S}))$ is performed using an MCMC (e.g. Metropolis method) with the transition probability

$$p(\mathcal{S}_1 \to \mathcal{S}_2) = \min\left[\frac{D(E(\mathcal{S}_1))}{D(E(\mathcal{S}_2))}, 1\right] \ . \tag{A.15}$$

Since the energies are sampled uniformly, the MCMC moves back and forth between low energy configurations and high energy ones. Passing through high energy configurations is most likely to erase any memory of the originating valley. Thus, when the MCMC moves to lower energy configurations other valleys are sampled.

Another advantage of the multicanonical method is that it is straight forward to calculate canonical averages by using the following technique:

$$\langle\mathcal{O}\rangle_T = \sum_{\{\mathcal{S}\}} \mathcal{O}(\mathcal{S})P(\mathcal{S},T) = \int_E \langle\mathcal{O}\rangle_E P(E;T)dE \tag{A.16}$$

where

$$\langle\mathcal{O}\rangle_E = \frac{1}{D(E)} \sum_{\{\mathcal{S}\}} \delta(\mathcal{H}(\mathcal{S}) - E)\mathcal{O}(\mathcal{S}) \tag{A.17}$$

and

$$P(E;T) = \frac{D(E)\exp(-E/T)}{\int_{E'} \exp(-E'/T)D(E')dE'} \ . \tag{A.18}$$

Assuming $D(E)$ is known, one only needs to estimate $\langle\mathcal{O}\rangle_E$ for each energy. Using Eq. (A.18), $\langle\mathcal{O}\rangle_T$ can be immediately calculated for all temperatures.

Until now we assumed that $D(E)$ is given. Since it is often unknown apriori, one needs to estimate it. Therefore the Multicanonical method has two stages: (i) Estimating the density of states, $D(E)$; (ii) sampling from $1/D(E(\mathcal{S}))$, and estimating $\langle\mathcal{O}\rangle_E$ which can then be used to calculate $\langle\mathcal{O}\rangle_T$.

## A.1 Estimating $D(E)$

Estimating $D(E)$ is done iteratively. $\hat{D}_{t=0}(E)$ is initialized to be uniform over the energy range $[0, \sum_{\langle i,j\rangle} J_{ij}]$, and then $\hat{D}_t(E)$ is updated in each iteration. Practically, the energy

range is divided into $N_E$ equally sized bins and the density of states is represented by a vector $\hat{D}_t(E_i)$ where $i = 1, \ldots, N_E$. In iteration $t$, one generates a sample of configurations drawn from a distribution $\propto 1/\hat{D}_t(E_i(\mathcal{S}))$ using an MCMC as described above (Eq. (A.15)). Denote by $h_t(E_i)$ the number of states in the sample whose energy falls in the $i$-th bin. Since $h_t(E_i)$ is approximately $D(E_i)/\hat{D}_t(E_i)$, the new estimate of the density of states is given by $\hat{D}_{t+1}(E_i) = h_t(E_i)\hat{D}_t(E_i)$. Note that sampling from the true $1/D(E(\mathcal{S}))$ distribution would yield a flat histogram $h_t(E_i)$. Therefore, the iterations are carried until a flat histogram is obtained.

# B Appendix B: Typical cuts using Weighted Belief Propagation (WBP)

This Appendix presents a novel graphical model-based approximation method termed Weighted Belief Propagation (WBP). This approximating aims to overcome the difficulties encountered using BP/GBP. The general idea is to combine several different BP solutions in an optimal manner.

We begin with a brief overview of the variational approach to obtain BP and GBP. Then, we present difficulties of applying BP and GBP to our problem. Next, the motivation for WBP is given, followed by its formal justification. Finally, we discuss some problematic issues of WBP.

## B.1 Variational approach and BP/GBP

A graphical model is a way to describe the joint distribution $P(S)$ which, in our case, is given by

$$P(S;T) = \frac{\exp(-E(S)/T)}{Z(T)} \ . \tag{B.19}$$

Variational methods define a family of trial distribution functions, $\mathcal{Q} = \{q(S)\}$, out of which the closest distribution to $P(S;T)$ is sought. The family of functions is chosen such that finding the optimal $q(S)$ is tractable. Following [13], one can use

$$-1/TF\left[q(S)\right] = \log Z(T) - KL\left[q(S)\|P(S;T)\right] \tag{B.20}$$

and the inequality

$$\log Z(T) \geq \log Z(T) - \min_{q(S)\in\mathcal{Q}} KL\left[q(S)\|P(S;T)\right] = -1/T \min_{q(S)\in\mathcal{Q}} F\left[q(S)\right] \tag{B.21}$$

where $F[q(S)]$, the variational free energy, is

$$F[q(S)] = \langle E(S)\rangle_{q(S)} - T\langle -\log(q(S))\rangle_{q(S)} \tag{B.22}$$

and $KL\left[q(S)\|P(S;T)\right]$ is the Kullback-Leibler divergence between $q(S)$ and $P(S;T)$. Consequently, the optimal approximation is obtained by finding the function $\hat{q} \in \mathcal{Q}$ which minimizes the $KL$-divergence, or equivalently minimizes $F[q(S)]$. In case $\mathcal{Q}$ contains the true

distribution, an equality holds and $F[q(S)] = -T \log Z(T)$. For example, BP is a procedure for finding a minimum of $F[q(S)]$, where $\mathcal{Q}$ is the family of distributions which assume that the graph is singly connected and parameterized by the single spin probabilities $b_i^\alpha$ and the pairwise probabilities $b_{ij}^{\alpha\beta}$.

As opposed to sampling methods which average over configurations, to provide $b_i^\alpha$ and $b_{ij}^{\alpha\beta}$, the BP's fixed point readily specifies approximated "averaged" values.

When applying BP one has to decide upon two issues: The initial values of the messages (e.g. uniformly distributed), and the order by which the messages are updated (either synchronously or asynchronously). A specific choice of message initializations and update method, is termed a BP "scheme". In our example, it happens that different BP schemes yield different values for $b_i^\alpha$ and $b_{ij}^{\alpha\beta}$. For example uniform initialization of the messages (either with synchronous or asynchronous update rule) results in a Bayesian solution both in phase B (as expected) and in phase F. Applying other schemes give rise to many different types of solutions in phase F. Each of these solutions corresponds to a low energy family of configurations which are globally different, but *none* of them coincides with the correct solution. For example several families represent solutions in which one of the types "overtakes" the whole system. The choice between these solutions depend on the values of the messages arriving from the different labelled spins, which in turn depend on the BP scheme but also on the system size, its topology and the energy difference between the states.

In order to overcome this dependence on the BP scheme, and also to provide the correct solution we suggest to approximate it by weighing the different solutions found. The rational is that each solution represents the "average" configuration in a specific subspace determined by the BP scheme, however, a weighted average of these solutions may yield a better approximation. The optimal weighting of different solutions is a function of their free energy, as shown below.

## B.2 Weighted Belief Propagation (WBP)

In this section we treat the above problem from a broader perspective, and derive the optimal averaging weights. This approach can be applied to combine solutions of either BP, GBP or any other method which approximates the joint distribution $P(S;T)$.

The input to our method is a set of $k$ trial probabilities, $\{q_\mu(S)\}_{\mu=1}^k$. For example, these may be the results of applying several BP schemes. Following the variational approach described above, we search for the closest function in the mixture space $\mathcal{Q}^{k\text{-mixture}}$ which is defined as all possible linear combination of the $k$ trial functions subject to the $\sum_\mu \lambda_\mu = 1$ constraint. One needs to find those parameters $\lambda_\mu$ which minimize $KL\left[\sum_\mu \lambda_\mu q_\mu(S) \| P(S;T)\right]$. The advantage of using a mixture model is that it is capable of introducing global correlations in the system which are *not* possible in local approximations such as the Bethe and the Kikuchi approximations. If one assumes that $q_\mu(S)$ have a zero overlap, *i.e.* $KL[q_\mu(S) \| q_\nu(S)] = \infty \; \forall \mu \neq \nu$, which is approximately correct since

we require the found solutions to be a $KL$-divergence above some large value, one can use

$$KL\left[\sum_{\mu}\lambda_{\mu}q_{\mu}(S)\|P(S;T)\right] = \sum_{\{S\}}\sum_{\mu}\lambda_{\mu}q_{\mu}(S)\log\frac{\sum_{\mu}\lambda_{\mu}q_{\mu}(S)}{P(S;T)} \quad \text{(B.23)}$$

$$\approx \sum_{\mu}\sum_{\{S\}_{\mu}}\lambda_{\mu}q_{\mu}(S)\log\frac{\lambda_{\mu}q_{\mu}(S)}{P(S;T)} \quad \text{(B.24)}$$

$$= -H[\lambda_{\mu}] + \sum_{\mu}\lambda_{\mu}KL\left[q_{\mu}(S)\|P(S;T)\right] \quad \text{(B.25)}$$

$$\equiv J(\lambda_{\mu}) \quad \text{(B.26)}$$

where $H[\lambda_{\mu}]$ is the entropy of the $\lambda_{\mu}$ distribution and $\{S\}_{\mu}$ is the part of space in which $q_{\mu}(S) > 0$. Note that the different $\{S\}_{\mu}$ regions are non-overlapping due to the assumption regarding their $KL$-divergence. The minimum of Equ. (B.26) is obtained, using a Lagrange multiplier, by solving the $k+1$ equations:

$$\frac{\partial}{\partial\lambda_i}\left(J(\lambda_{\mu}) + \gamma(\sum_{\mu}\lambda_{\mu} - 1)\right) = 0 \quad \forall i \quad \text{(B.27)}$$

$$\sum_{\mu}\lambda_{\mu} = 1 \quad \text{(B.28)}$$

which gives

$$\log\lambda_i + 1 + KL\left[q_i(S)\|P(S;T)\right] + \gamma = 0 \quad \forall i \quad \text{(B.29)}$$

$$\sum_{\mu}\lambda_{\mu} = 1\ . \quad \text{(B.30)}$$

The solution is given by

$$\lambda_i = \frac{\exp\left(-KL\left[q_i(S)\|P(S;T)\right]\right)}{\sum_{\mu}\exp\left(-KL\left[q_{\mu}(S)\|P(S;T)\right]\right)} \quad \text{(B.31)}$$

$$= \frac{\exp\left(-KL\left[q_i(S)\|P(S;T)\right]\right)Z(T)}{\sum_{\mu}\exp\left(-KL\left[q_{\mu}(S)\|P(S;T)\right]\right)Z(T)} \quad \text{(B.32)}$$

$$= \frac{\exp\left(-\beta F\left[q_i(S)\right]\right)}{\sum_{\mu}\exp\left(-\beta F\left[q_{\mu}(S)\right]\right)}\ . \quad \text{(B.33)}$$

These weight are intuitive since each solutions is weighed according to its entropy and its Boltzmann weight.

## B.3 Difficulties in applying WBP

The main problem in applying WBP is obtaining enough BP results in order approach the correct solution. The results will be inaccurate if one of the solutions is missed, especially,

if this solution has a low free energy. Aiming to find as many different solutions as possible we tried various BP schemes that differ by: (i) their messages initialization and (ii) the random order of updating the messages. Out of all the BP results we select a non-redundant representative set in which no pair of functions have $KL$-divergence less than some value.

# Publication 4:

## DNA microarrays identification of primary and secondary target genes regulated by P53

Authors: K. Karuppiah, A. Ninette, G. Rechavi, J. Jakob-Hirsch, I. Kela, N. Kaminski, G. Getz, E. Domany and D. Givol
Published in: *Oncogene*, **20**, 2225–2234 (2001).

Oncogene (2001) 20, 2225–2234

www.nature.com/onc

# DNA microarrays identification of primary and secondary target genes regulated by p53

Karuppiah Kannan[1], Ninette Amariglio[2], Gideon Rechavi[2], Jasmine Jakob-Hirsch[2], Itai Kela[3], Naftali Kaminski[4], Gad Getz[3], Eytan Domany[3] and David Givol*[,1]

[1]*Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel;* [2]*Department of Pediatric Hematology-Oncology, The Chaim Sheba Medical Center and Sackler School of Medicine, Tel Aviv University, Israel;* [3]*Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel;* [4]*Functional Genomics Unit, The Chaim Sheba Medical Center and Sackler School of Medicine, Tel Aviv University, Israel*

**The transcriptional program regulated by the tumor suppressor p53 was analysed using oligonucleotide microarrays. A human lung cancer cell line that expresses the temperature sensitive murine p53 was utilized to quantitate mRNA levels of various genes at different time points after shifting the temperature to 32°C. Inhibition of protein synthesis by cycloheximide (CHX) was used to distinguish between primary and secondary target genes regulated by p53. In the absence of CHX, 259 and 125 genes were up or down-regulated respectively; only 38 and 24 of these genes were up and down-regulated by p53 also in the presence of CHX and are considered primary targets in this cell line. Cluster analysis of these data using the super paramagnetic clustering (SPC) algorithm demonstrate that the primary genes can be distinguished as a single cluster among a large pool of p53 regulated genes. This procedure identified additional genes that co-cluster with the primary targets and can also be classified as such genes. In addition to cell cycle (e.g. *p21*, *TGF-β*, *Cyclin E*) and apoptosis (e.g. *Fas*, *Bak*, *IAP*) related genes, the primary targets of p53 include genes involved in many aspects of cell function, including cell adhesion (e.g. *Thymosin*, *Smoothelin*), signaling (e.g. *H-Ras*, *Diacylglycerol kinase*), transcription (e.g. *ATF3*, *LISCH7*), neuronal growth (e.g. *Ninjurin*, *NSCL2*) and DNA repair (e.g. *BTG2*, *DDB2*). The results suggest that p53 activates concerted opposing signals and exerts its effect through a diverse network of transcriptional changes that collectively alter the cell phenotype in response to stress.** *Oncogene* (2001) **20,** 2225–2234.

**Keywords:** DNA microarray; cycloheximide; ts-p53; target genes; expression profile; clustering

## Introduction

The function of p53 as tumor suppressor is mainly due to its activity as a transcription factor that activates many genes in response to various types of stress (El-Deiry *et al*., 1992). This may be the basis for p53 protection of cells against DNA damage and various stress conditions that lead usually to growth arrest or apoptosis (Levine, 1997). Mice with p53 which was rendered transcriptionally inactive by mutations at codon 25 and 26 are predisposed to tumors similar to p53 deficient mice (Jimenez *et al*., 2000), demonstrating the importance of transcriptional activity for p53 function. The heterogeneity of gene transcription profile in response to p53 was demonstrated recently using transcriptionally regulated p53 expressed in colon cancer cell line (Zhao *et al*., 2000; Yu *et al*., 1999). This heterogeneity could result partly from secondary effect and from dependence on other transcription factors. In order to overcome part of this heterogeneity and to distinguish primary targets of p53, we used the temperature sensitive p53 (denoted Val135) (Michalovitz *et al*., 1990) expressed in the human lung cancer cell line H1299 to analyse the transcriptional programs induced by p53. Shifting the temperature to 32°C causes Val135 to assume wild-type p53 conformation and induce target genes. This conformational change does not require protein synthesis and allows for the analysis of p53 induced genes in the presence of protein synthesis inhibitor to prevent secondary effects brought about by the activated genes.

Here we analysed the profile of gene expression regulated by p53 at 32°C in the presence and absence of cycloheximide (CHX) using DNA microarrays containing ~7000 probes for human genes (Affymetrix, Santa Clara, USA). Less than 20% of the genes regulated by p53 were also regulated in the presence of CHX and therefore are defined as primary targets of p53. We were able to distinguish these genes as a cluster of primary targets among a large pool of p53 regulated genes by subjecting the expression data to clustering analysis using Super Paramagnetic Clustering (SPC) algorithm (Blatt *et al*., 1996). The results indicate that in addition to genes involved in cell cycle

*Correspondence: D Givol
Received 7 November 2000; revised 21 January 2001; accepted 29 January 2001

control and apoptosis, p53 regulates directly a plethora of genes involved in many aspects of cellular functions such as DNA repair, cytoskeleton, extracellular matrix and signal transduction. This suggests that in addition to its effect on cell cycle, p53 may exert some of its tumor suppressive effects through a diverse network of transcriptional changes that alter cellular phenotype and behavior in response to stress.

## Results

### *Effect of cycloheximide on p53 target genes*

Analysis of mRNA from H1299Val135 cells and H1299 control cells by hybridization to DNA microarrays showed that after 12 h at 32°C, 212 and 77 genes were up and down-regulated respectively by more than 2.5-fold. It is likely that some of these genes may be regulated because of secondary effects induced by the primary targets of p53. To identify the primary targets, we analysed the effect of p53 in the presence and absence of cycloheximide (CHX) at 32°C in both H1299Val135 and H1299 control cells. The inhibition of protein synthesis by CHX presumably prevents most of the seconday gene regulation (O'Hagan *et al*., 2000; Coller *et al*., 2000) that is not transactivated directly by p53. The results showed that in the presence of CHX, p53 remained stable for at least 12 h (Figure 1A) and induce significantly the mRNA of *p21waf*, a major target for p53 (Figure 1B). Evidently protein synthesis was indeed shut down as no p21waf or hDM2 (human MDM2) proteins were detected, in contrast to their presence in the experiment without CHX (Figure 1C,D). It was shown previously that at 32°C p53Val135 is relocalized to the nucleus (Ginsberg *et al*., 1991) and that it may therefore be protected from degradation in the presence of CHX due to a lack of nuclear exclusion and the absence of hDM2 synthesis (Haupt *et al*., 1997; Kubbutat *et al*., 1997). Hence, direct target genes of p53 will be transcriptionally activated, but because of inhibition of protein synthesis, the p53-induced proteins will not be synthesized and will not induce secondary or indirect p53 target genes.

### *p53 primary target genes identified in the presence of cycloheximide*

RNA from various time points (between 2 to 24 h) was isolated, labeled and hybridized to oligonucleotide microarrays (Genechip Hugene FL array, Affymetrix, Santa Clara, USA) which contains probes for $\sim$7000 human genes. Those genes that are transcriptionally regulated by p53 under both conditions, i.e. with and without CHX, are likely to be primary targets for p53-mediated transcription. In order to eliminate noisy data in the analysis of the hybridization experiments, we applied a very stringent filter; we selected genes that showed more than 2.5-fold induction or repression (over their controls) at three or more time points in the presence or absence of CHX. Two hundred and fifty-nine upregulated genes passed this filter for the experiment in the absence of CHX whereas 131 genes passed this filter for the experiment with CHX; 38 genes were found to be common to these two groups (see the Venn diagram in Figure 2A and Table 1). Denote by *G(38)* the group of 38 genes that have been identified by our analysis as possible primary upregulated targets of p53. It is important to note that $n=38$, the number of these genes, exceeds significantly the number $n_r$, that would have been obtained had we applied the same filtering procedure to random data. To estimate $n_r$, assume that for every gene in any single experiment, the over-expression level can exceed 2.5 with probability $P$, and that such over-expressions are independent random events. All together, nine measurements were taken for 7070 genes; by counting the number of occurrences of 2.5-fold over-expression, we estimate $P\approx 0.053$. The probability that a particular gene will be over-expressed at least three times at a level above 2.5 in the experiments without CHX is $5.7\cdot10^{-4}$; the same figure for the experiments with CHX is $1.4\cdot10^{-3}$; and for passing the filter in both

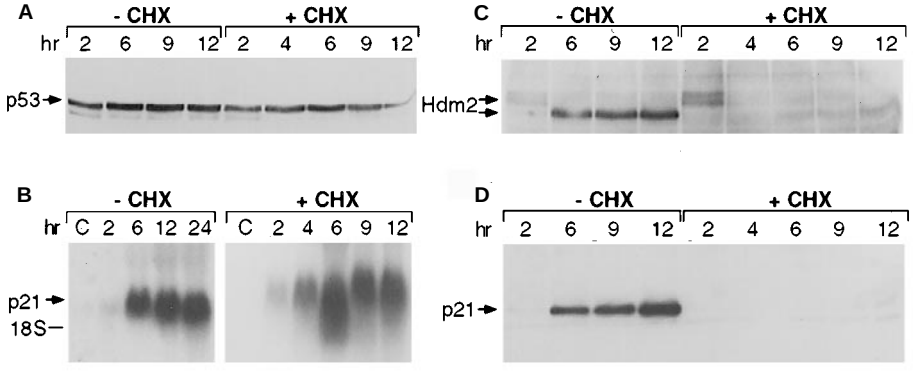



**Figure 1** Effect of cycloheximide (CHX) on H1299Val135 cells, demonstrating synthesis of mRNA and lack of protein synthesis. (**A**) Western-blot analysis of p53Val135 in H1299 cells in the presence and absence of CHX. (**B**) Northern-blot analysis of *p21Waf* in lysates from H1299Val135 cells grown in the presence and absence of CHX. (**C**) Western-blot analysis of Hdm2, (**D**) Western-blot analysis of p21waf in lysates from H1299Val135 cells grown in the presence and absence of CHX. **C**, control cells of H1299 devoid of p53. hr, indicates time after shifting the temperature to 32°C

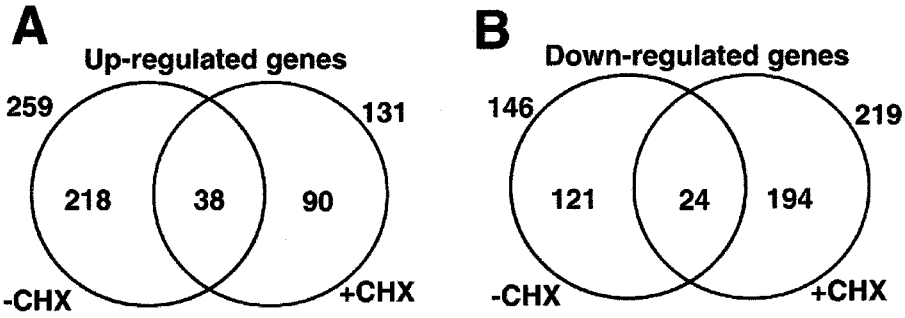



**Figure 2** Venn diagram of the number of genes that were regulated by p53 in the presence and absence of CHX. (**A**) Only genes that showed at least 2.5-fold up or (**B**) down-regulation in at least three of the time points in each experiment i.e. with or without CHX, were listed in this analysis. Note that only 38 of the up-regulated and 24 of the down-regulated genes were unaffected by CHX

experiments $8{\cdot}10^{-7}$. Multiplying these probabilities by the total number of genes tested, we get estimates for the numbers of genes that would have passed the filter and would have been assigned to the three groups. Next to these estimates, which represent the expected number of genes for random expression data, i.e. for a randomized expression matrix, we placed in parentheses the actual numbers, obtained by our analysis for the real data: with CHX: 10 (vs 131); without CHX: 4 (vs 259); and the number of genes expected to be in the overlap; $n_r = 5{\cdot}10^{-3}$ (vs 38). The large disparity between the numbers obtained under the assumption of a random process and the actual measured numbers proves beyond doubt the statistical significance of our findings.

The down-regulated genes that passed this filter were 146 and 219 respectively in the experiments without and with CHX and 24 genes were common to these groups (Figure 2B). The common genes were regulated by p53 irrespective of the presence or absence of CHX (Tables 1 and 2). The remaining 218 upregulated and 121 downregulated genes (Figure 2) may be indirect or secondary targets of p53 (see Table at our web site: http://www.weizmann.ac.il/home/ligivol/primary.html). In this study, we focus mainly on the analysis of the upregulated genes. Apparently CHX by itself also induced many genes probably due to removal of inhibitory signals for gene expression or changes in gene expression associated with clonal variability. Had we performed the analysis of p53 induced genes only in the presence of CHX, we would have detected 131 genes (Figure 2A) as primary targets, most of which may be unrelated to direct p53 regulation.

In order to analyse genes regulated by CHX in the absence of p53, H1299 cells were incubated at 32°C in the presence of CHX and RNA was collected at 2 and 6 h and analysed by hybridization to the DNA microarray. Three genes from the 38 primary upregulated genes (Table 1) were also upregulated by CHX alone. These are the *ATF3* (activated transcription factor 3), *Histone 2A* like protein and *Bak*. In a similar way the effect of temperature on H1299 cells was analysed by incubating H1299 cells at 32°C and collecting RNA for hybridization at 2 and 12 h. Three genes from the primary genes: *Lysyl oxidase*, *Diacylglycerol kinase* and *LISCH7* were found to be upregulated also after 12 h at 32°C. We keep these genes in the primary target list although they were upregulated by the control conditions (CHX or temperature). It is possible that these genes are transactivated by various stress conditions as well as by p53. For example, $p21^{waf1}$, a major primary target of p53 is also induced by a variety of external stimuli and stress conditions independent of P53 (Michieli *et al*., 1994).

### *Identifying primary p53 target genes by cluster analysis*

A major problem we must now face is that the identities of the genes of *G(38)*, that were designated above as possible primary upregulated targets of p53, are determined by our stringent filtering criteria. Had we set our threshold at, say, observing twofold increased expression (instead of 2.5) at two time points (instead of three), the number of genes that passed the filter would have been larger. The values of our filtering parameters (2.5-fold change at three or more times, in both experiments) were chosen in an arbitrary fashion; it is important to assess the extent to which relaxation of our filtering criteria will add primary p53 target genes beyond the 38 candidates that were already found. In order to reduce the dependence of our results on the precise values of the arbitrarily chosen filtering parameters, we performed cluster analysis on the data, using more relaxed filtering criteria to select the genes to be included in the analysis.

The first cluster analysis was done on the 259 genes that passed our filter for the experiment without CHX. That is, in our filtering we relaxed completely the restriction on the expression levels measured in the experiment with CHX (but did use the results of these experiments in the cluster analysis). Each of the 259 genes is represented by its expression levels taken at nine different time points (in the two experiments combined). The data were normalized as follows. Denote by $A_{ij}$ the $\log_2$ ratio of the expression of gene $i$ (where $i=1,2,...,259$), measured at experiment (and time) $j$ (with $j=1,2,...,9$), with respect to the control. For $j=1,...,4$ we divided $A_{ij}$ by $[\sum_{j=1}^{4} A_{ij}^2]^{1/2}$ and for

**Table 1** Primary target genes upregulated by p53

| *Accession no.* | | | | | *Ratio of gene expression at specific time points* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *−CHX* | | | | *+CHX* | | | | |
| | | | | *h* | *2* | *6* | *12* | *24* | *2* | *4* | *6* | *9* | *12* |
| | *Apoptosis* | G(47) | G(38) | G(9) | | | | | | | | | |
| X63717 | Fas/APO-1 cell surface antigen | + | + | | 2.2 | 5.6 | 7.9 | 8.3 | 3.2 | 6.8 | 3.3 | 5.7 | 9.8 |
| U82987 | Bcl-2 binding component 3 (bbc3) | + | + | | 6.2 | 16.5 | 16.8 | 8.4 | 7.6 | 13.4 | 28.5 | 32.5 | 30.5 |
| U00115 | Bcl-6 | + | + | | 2.5 | 3.3 | 5.8 | 6.5 | 3.7 | 4.7 | 6.1 | 6.0 | 6.8 |
| U16811 | Bak | + | + | | 4.0 | 7.0 | 4.4 | 6.8 | 1.0 | 2.3 | 4.5 | 4.7 | 7.4 |
| | *Cell cycle* | | | | | | | | | | | | |
| U09579 | p21 WAF1 | + | + | | 4.3 | 18.0 | 16.9 | 13.0 | 3.5 | 6.7 | 9.2 | 8.2 | 7.4 |
| D90070 | ATL derived PMA responsive peptide | + | + | | 4.0 | 4.5 | 2.8 | 4.1 | 1.8 | 3.5 | 3.4 | 4.3 | 6.1 |
| M60974 | GADD45 | + | | | 2.3 | 3.2 | 3.4 | 3.6 | 1.6 | 1.6 | 2.1 | 2.1 | 2.0 |
| | *DNA repair/replication* | | | | | | | | | | | | |
| U72649 | BTG2 | + | + | | 3.0 | 4.9 | 4.8 | 5.2 | 11.0 | 15.9 | 20.1 | 17.9 | 16.5 |
| U18300 | Damage-specific DNA binding protein | + | + | | 1.3 | 4.0 | 5.4 | 7.8 | 1.1 | 2.6 | 8.6 | 6.1 | 9.4 |
| U90551 | Histone 2A-like protein | + | + | | 2.6 | 2.6 | 2.5 | 3.5 | 1.7 | 3.1 | 4.5 | 8.8 | 16.7 |
| M15796 | PCNA | + | | | 2.3 | 3.9 | 5.3 | 5.3 | 1.7 | 1.4 | 1.7 | 1.9 | 3.1 |
| | *Receptors/ECM* | | | | | | | | | | | | |
| X72012 | Endoglin | | + | | 21.7 | 15.2 | 7.9 | 6.8 | 3.4 | 1.0 | 14.0 | 6.9 | 1.0 |
| U16306 | Versican | + | + | | 1.0 | 5.2 | 11.9 | 16.6 | 1.0 | 2.0 | 2.7 | 6.8 | 13.3 |
| M21904 | Heavy chain 4F2 | + | | | 7.7 | 6.5 | 8.2 | 6.1 | 0.9 | 1.0 | 1.6 | 1.3 | 1.4 |
| AF010193 | SMAD7 | | | + | 3.6 | 1.9 | 3.9 | 3.1 | 1.9 | 1.8 | 1.5 | 1.5 | 1.6 |
| | *Growth factors/inhibitors* | | | | | | | | | | | | |
| AB000584 | TGF-Beta Superfamily protein | + | + | | 3.5 | 38.6 | 67.2 | 50.1 | 11.5 | 16.0 | 38.4 | 33.6 | 32.4 |
| M62402 | IGFBP6 | + | + | | 17.5 | 17.7 | 18.5 | 17.6 | 22.3 | 13.9 | 27.8 | 25.4 | 28.9 |
| L42379 | Quiescin/QSCN6 | + | | | 2.7 | 2.4 | 2.5 | 2.8 | 1.3 | 2.2 | 2.3 | 2.3 | 1.7 |
| X97324 | Adipophilin | + | | | 5.2 | 6.3 | 7.1 | 10.0 | 1.5 | 1.9 | 1.7 | 1.9 | 3.8 |
| U72263 | Multiple exostoses type II protein | + | | | 2.9 | 2.7 | 3.1 | 6.2 | 1.5 | 2.0 | 1.4 | 1.8 | 3.2 |
| | *Cytoskeleton/cell adhesion* | | | | | | | | | | | | |
| X13839 | Vascular smooth muscle alpha-actin | + | + | | 6.5 | 14.6 | 44.3 | 98.3 | 6.4 | 9.0 | 18.9 | 18.4 | 26.9 |
| Z49989 | Smoothelin | + | + | | 3.7 | 4.3 | 3.9 | 3.5 | 1.7 | 2.2 | 4.6 | 4.1 | 3.4 |
| X05608 | Neurofilament subunit NF-L | + | + | | 3.0 | 7.7 | 16.6 | 32.0 | 5.0 | 8.8 | 9.6 | 16.8 | 32.2 |
| D82345 | NB Thymosin beta | + | + | | 2.4 | 3.4 | 4.2 | 5.4 | 1.6 | 2.5 | 3.8 | 2.8 | 3.9 |
| X93510 | LIM domain protein | | + | | 4.5 | 5.5 | 4.3 | 5.6 | 4.3 | 2.4 | 9.3 | 3.8 | 1.0 |
| | *Metabolism* | | | | | | | | | | | | |
| U24389 | Lysyl oxidase-like protein | + | + | | 16.7 | 17.5 | 23.0 | 22.9 | 5.5 | 4.2 | 9.7 | 8.3 | 6.9 |
| L38668 | UDP-Galactose 4 epimerase (GALE) | + | + | | 4.2 | 3.6 | 3.2 | 4.4 | 3.2 | 3.7 | 5.1 | 3.4 | 2.8 |
| Y12556 | cAMP activated Protein Kinase B | + | + | | 4.0 | 1.0 | 19.7 | 22.2 | 1.0 | 3.1 | 3.1 | 7.1 | 37.5 |
| U05572 | Lysosomal Mannosidase alpha B | | + | | 9.2 | 2.0 | 7.0 | 15.9 | 6.5 | 5.6 | 13.9 | 13.0 | 7.2 |
| Y09616 | Carboxylesterase (liver) | + | + | | 2.3 | 4.1 | 5.9 | 9.1 | 1.9 | 2.5 | 3.0 | 3.4 | 3.6 |
| U78735 | ABC3 | | | + | 4.5 | 3.0 | 2.8 | 3.6 | 1.9 | 2.0 | 1.9 | 1.4 | 1.1 |
| M20902 | Apolipoprotein C-I (VLDL) | + | | | 3.1 | 3.9 | 2.9 | 4.1 | 2.1 | 2.9 | 2.3 | 2.2 | 1.7 |
| U20325 | CART | + | | | 2.8 | 3.9 | 3.0 | 3.3 | 1.3 | 1.7 | 2.1 | 1.3 | 1.9 |
| M12625 | Lecithin-cholesterol acyltransferase | + | | | 1.4 | 2.9 | 3.0 | 3.0 | 0.8 | 1.4 | 3.6 | 2.6 | 2.4 |
| D87292 | Rhodanese | | | + | 7.7 | 4.9 | 5.4 | 10.5 | 1.6 | 1.6 | 1.7 | 1.3 | 1.2 |
| | *Neuronal growth* | | | | | | | | | | | | |
| U35139 | NECDIN related protein | | + | | 3.2 | 1.0 | 3.0 | 15.7 | 5.8 | 4.9 | 2.7 | 4.8 | 5.5 |
| M96740 | NSCL-2 gene | + | + | | 1.0 | 4.2 | 10.1 | 12.6 | 2.6 | 2.0 | 4.8 | 7.0 | 11.0 |
| U60062 | FEZI-T gene | + | + | | 1.0 | 2.7 | 2.7 | 4.0 | 1.6 | 2.3 | 2.6 | 5.6 | 8.7 |
| U72661 | Ninjurin 1 | + | + | | 1.4 | 3.7 | 7.6 | 10.4 | 1.9 | 3.1 | 6.0 | 6.3 | 7.8 |
| U48437 | Amyloid precursor-like protein | + | + | | 1.9 | 2.7 | 3.7 | 4.7 | 1.0 | 1.5 | 3.6 | 2.7 | 3.8 |
| | *Signal transduction* | | | | | | | | | | | | |
| J00277 | c-Ha-ras 1 | + | + | | 2.3 | 3.1 | 4.7 | 4.8 | 1.2 | 1.5 | 4.1 | 5.6 | 5.7 |
| X77777 | Intestinal VIPR related protein | | + | | 14.7 | 7.5 | 11.9 | 9.8 | 1.0 | 2.7 | 19.9 | 4.7 | 1.0 |
| X62535 | Diacylglycerol Kinase (alpha) | + | + | | 4.4 | 12.0 | 11.2 | 38.8 | 1.2 | 3.4 | 6.0 | 7.1 | 10.1 |
| U56998 | Putative ser/thr protein kinase | + | + | | 6.3 | 3.6 | 1.0 | 12.5 | 2.1 | 2.2 | 4.8 | 6.4 | 4.5 |
| L08835 | DM Kinase | + | | | 2.5 | 3.4 | 3.3 | 3.7 | 0.9 | 1.2 | 1.8 | 1.5 | 1.5 |
| L42176 | DRAL-FHL2 | | | + | 5.4 | 3.8 | 3.8 | 6.5 | 2.0 | 2.0 | 1.8 | 1.5 | 1.7 |
| | *Transcription* | | | | | | | | | | | | |
| L19871 | Activating transcription factor 3 | + | + | | 1.4 | 4.4 | 5.9 | 5.2 | 4.2 | 7.2 | 18.9 | 8.8 | 9.9 |
| U38315 | ZNF127-Xp | | + | + | 4.4 | 5.1 | 9.2 | 9.5 | 6.5 | 6.4 | 3.9 | 7.2 | 3.1 |
| AD000684 | LISCH7 | + | + | | 5.2 | 6.1 | 6.1 | 7.4 | 2.7 | 1.0 | 2.1 | 3.1 | 3.8 |
| M29580 | Zinc finger protein 7 | + | | | 1.2 | 2.6 | 3.6 | 4.7 | 1.2 | 1.2 | 1.2 | 1.3 | 1.2 |
| U90913 | Tip-1 | + | | | 3.4 | 2.7 | 3.1 | 3.7 | 1.5 | 2.1 | 2.3 | 1.9 | 2.3 |
| HG3494 | Nuclear factor NF-116 | | | + | 3.3 | 2.4 | 2.9 | 2.7 | 2.9 | 2.8 | 2.4 | 2.3 | 1.1 |
| | *Other* | | | | | | | | | | | | |
| U10099 | POM-ZP3 | + | + | | 3.2 | 4.2 | 4.7 | 4.8 | 1.8 | 2.0 | 4.0 | 3.5 | 3.3 |
| D87434 | KIAAA0247 | + | + | | 1.0 | 4.5 | 4.7 | 6.6 | 3.5 | 7.5 | 10.3 | 15.4 | 18.9 |
| U33147 | Mammaglobin 1 | | + | + | 3.5 | 1.0 | 4.5 | 4.1 | 4.5 | 4.2 | 2.9 | 4.6 | 1.0 |
| J05016 | Disulfide isomerase related protein | + | | | 2.6 | 3.8 | 6.0 | 8.7 | 2.1 | 1.4 | 1.8 | 1.8 | 2.3 |

*Continued*

**Table 1** (*Continued*)

| Accession no. | | | | Ratio of gene expression at specific time points | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | −CHX | | | | | +CHX | | | |
| | | | | *h* | *2* | *6* | *12* | *24* | *2* | *4* | *6* | *9* | *12* |
| U81556 | OS4 | | + | 2.7 | 2.6 | 2.4 | 3.0 | 1.6 | 2.2 | 1.8 | 1.4 | 1.3 |
| S58544 | Infertility-related sperm protein | + | | 3.0 | 3.9 | 7.6 | 3.1 | 1.6 | 2.0 | 1.7 | 3.0 | 4.3 |
| D63481 | KIAA0147 | + | | 1.8 | 2.6 | 3.1 | 2.6 | 1.2 | 1.2 | 1.5 | 2.1 | 1.7 |
| Z35093 | SURF-1 | | + | 2.2 | 2.6 | 2.6 | 3.4 | 1.9 | 2.0 | 2.0 | 1.6 | 1.4 |
| U94747 | WD repeat protein HAN11 | + | | 2.3 | 3.1 | 3.4 | 3.4 | 0.8 | 1.3 | 3.0 | 1.4 | 3.1 |

Primary targets are defined as genes that changed their expression over 2.5-fold in at least three time points in the presence and absence of cycloheximide (CHX). The ratio of gene expression was determined by the expression level at each time point divided by that of H1299 contol cells at 2 h at 32°C. The list denoted *G(38)* consists of the genes in the overlap in Figure 2. The list denoted *G(47)* constitutes the set of primary p53 targets generated from the cluster analysis in Figure 3 (see text). The list denoted *G(9)* contains the genes in cluster **a** in Figure 3A

**Table 2** Primary target genes downregulated by p53

| Accession no. | | Ratio of gene expression at specific time points | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | −CHX | | | | +CHX | | | | |
| | | *h* | *2* | *6* | *12* | *24* | *2* | *4* | *6* | *9* | *12* |
| | *Apoptosis* | | | | | | | | | | |
| U45878 | Inhibitor of Apoptosis protein 1 (MIHB) | −4.0 | −8.0 | −3.9 | −3.5 | −1.9 | −1.8 | −2.8 | −3.1 | −2.6 |
| U37546 | Inhibitor of Apoptosis protein 2(MIHC) | −4.2 | −10.7 | −14.9 | −15.8 | −2.8 | −2.8 | −3.0 | −3.7 | −3.2 |
| | *Angiogenesis* | | | | | | | | | |
| L22548 | Collagen type XVIII alpha 1 | −8.4 | −8.4 | −8.4 | −7.6 | −2.9 | −2.9 | −2.9 | −2.9 | −2.9 |
| | *Cell Cycle* | | | | | | | | | |
| L78833 | BRCA-1 | −2.9 | −3.1 | −3.1 | −3.1 | −4.1 | −4.5 | −4.5 | −2.3 | −4.5 |
| M74093 | Cyclin E | −0.8 | −2.9 | −4.0 | −3.2 | −1.4 | −2.6 | −6.9 | −4.5 | −3.7 |
| U77949 | Cdc6-related protein | −2.2 | −3.3 | −5.9 | −6.4 | −4.9 | −4.4 | −9.1 | −1.6 | −2.1 |
| M72885 | GOS2 | −33.4 | −137.0 | −172.0 | −143.0 | −9.3 | −15.1 | −11.5 | −21.0 | −107.0 |
| | *DNA repair/replication* | | | | | | | | | |
| M96684 | Purine rich element binding protein A | −4.2 | −9.3 | −3.0 | −2.5 | −1.5 | −3.5 | −5.3 | −5.3 | −2.4 |
| | *Receptors/ECM* | | | | | | | | | |
| U31201 | Laminin gamma 2 chain | −8.5 | −8.5 | −8.5 | −8.5 | −2.0 | −2.7 | −7.1 | −2.4 | −3.9 |
| U90716 | Caxsackie virus and adenovirus receptor | −5.4 | −3.2 | −4.3 | −2.5 | −2.8 | −3.0 | −4.2 | −2.3 | −1.1 |
| U60975 | Receptor gp250 precursor | −2.9 | −5.6 | −1.9 | −3.1 | −1.5 | −1.3 | −6.3 | −3.4 | −2.8 |
| J04970 | Carboxypeptidase M | −6.2 | −6.2 | −5.8 | −6.2 | −6.9 | −6.9 | −6.9 | −6.1 | −6.9 |
| U17566 | Folate transporter | −0.8 | −2.8 | −10.3 | −10.3 | −3.0 | −8.8 | −10.3 | −2.6 | −2.5 |
| | *Metabolism* | | | | | | | | | |
| U00238 | Glutamine PRPP amidotransferase | −1.6 | −3.0 | −3.5 | −3.9 | −1.6 | −2.0 | −7.2 | −9.2 | −4.5 |
| | *Neuronal growth* | | | | | | | | | |
| D11428 | Peripheral myelin protein 22 | −0.3 | −2.7 | −2.7 | −2.7 | −3.5 | −2.9 | −10.9 | −9.2 | −10.7 |
| S78296 | Neurofilament-66 | −5.8 | −6.9 | −5.1 | −6.1 | −2.6 | −2.6 | −6.1 | −2.8 | −2.2 |
| U79255 | Amyloid B-precursor binding protein | −0.6 | −8.3 | −8.3 | −8.3 | −5.0 | −5.0 | −2.5 | −5.0 | −5.0 |
| U73960 | ADP-ribosylation factor-like protein 4 | −4.7 | −3.4 | −6.4 | −9.8 | −1.4 | −2.2 | −5.9 | −5.9 | −4.2 |
| | *Transcription* | | | | | | | | | |
| U80017 | Basic transcription factor 2 p44 | −2.3 | −2.6 | −2.7 | −3.1 | −1.5 | −2.2 | −8.0 | −4.0 | −3.3 |
| X16706 | Fra-2 | −1.9 | −5.4 | −5.4 | −5.4 | −3.9 | −2.8 | −2.0 | −1.9 | −3.1 |
| | *Other* | | | | | | | | | |
| HG2855 | HSP 70 | −1.9 | −2.6 | −2.9 | −3.4 | −1.0 | −1.3 | −3.1 | −6.9 | −20.2 |
| L19183 | MAC30 | −3.2 | −2.9 | −3.3 | −4.0 | −2.0 | −2.7 | −3.8 | −2.7 | −3.4 |
| AB000467 | RES4-25 | −2.4 | −5.0 | −5.0 | −5.0 | −2.1 | −1.1 | −6.3 | −2.5 | −6.3 |
| U79273 | Clone 23933 | −4.9 | −16.3 | −2.5 | −3.6 | −4.2 | −4.2 | −4.2 | −4.2 | −2.7 |

The 24 genes that were downregulated in either the presence of absence of CHX

$j=5,...,9$ by $[\sum_{j=5}^{9} A_{ij}^2]^{1/2}$; the resulting 9-component vector represents gene *i*. The 259 genes were clustered by SPC (Blatt *et al*., 1996; Getz *et al*., 2000) (see below). Genes with similar expression profiles (over the time courses of both experiments) are represented by two nearby vectors and are placed in the same cluster. This cluster analysis answers directly two questions: (1) Do all, or a majority of the genes of *G(38)* cluster together?; and (2) What other genes cluster together with these possible primary targets? If the answer to the first question is positive, we can identify an expression profile which is characteristic of primaries, and identify those genes that share this profile and cluster together with the genes of *G(38)* as good candidates for being primaries (even though they did not pass the original stringent filtering process). Furthermore, if we find that some of the members of *G(38)* have significantly different expression kinetics than this characteristic profile, these genes should possibly be removed from the list of primaries.

Regarding the first question – the genes of *G(38)* are special in that their expression increased at least 2.5-fold at three or more time points of both

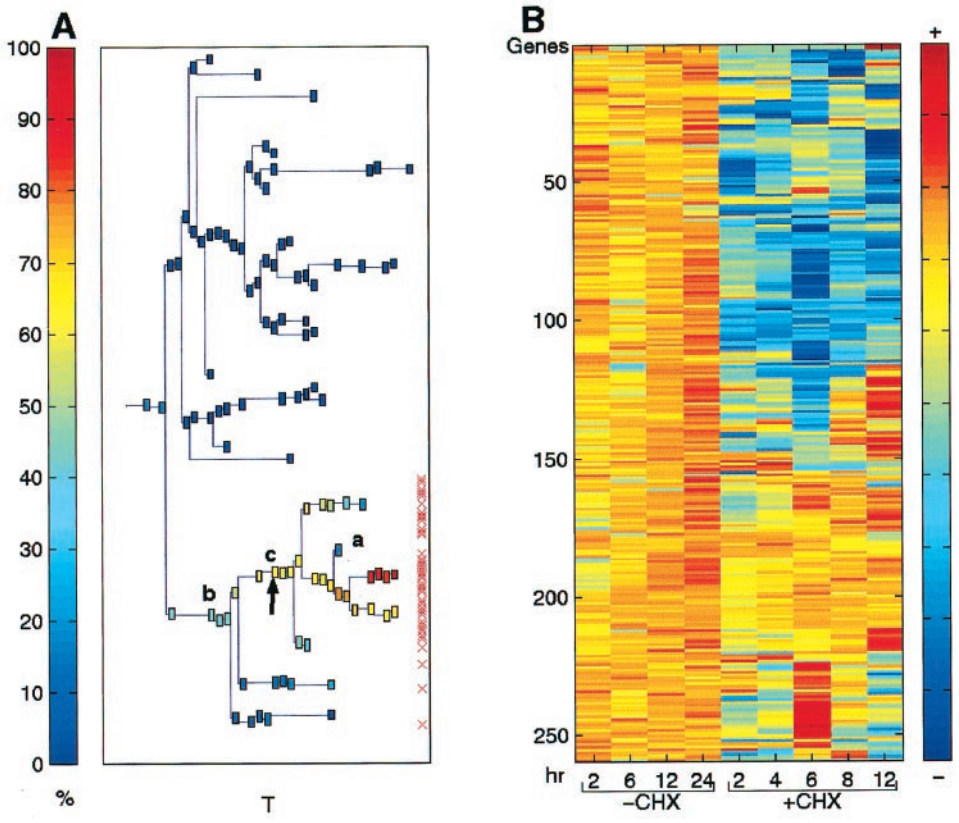



**Figure 3** Clustering results using super-paramagnetic clustering (SPC) for the 259 genes that were upregulated at three time points or more upon activation of p53 in the experiment without cycloheximide (CHX). (**A**) The dendrogram of the genes that include clusters of size 4 and larger. Each cluster is represented by a box colored according to the per cent of primary target genes (38 genes, see Table 1) contained in the cluster. The distribution of the 38 primary target genes is marked by red crosses at the right. (**B**) The normalized log ratio of the nine experiments (four without CHX and five with CHX) are plotted. The genes are ordered according to the dendrogram on the left. The color represents induction (red) or repression (blue). T, a parameter of the SPC algorithm that controls the resolution at which the cluster is found. %, per cent of primary target genes in the cluster. The cluster marked by an arrow (**c**) contains 87% of the 38 primary genes. The cluster marked by **b** contains all the 38 primary genes and the cluster marked by **a** contains the nine genes that show different kinetics (Table 1 and Figure 4)

experiments. Hence their representative 9-component vectors are likely to be close – but some may also be uncorrelated. For example, a gene whose expression decreases with time will have low correlation with one that increases. This situation may happen even if both genes passed the filtering criteria.

The results of our cluster analysis are summarized in the dendrogram of Figure 3A. The parameter $T$ on the horizontal axis controls the resolution at which the data are viewed. At $T=0$ all 259 genes are in a single cluster; as $T$ increases, large groups split into smaller ones. The boxes indicate clusters that contain more than four genes. Each box is colored according to its 'purity' – the percentage of members of $G(38)$ among the genes contained in the corresponding cluster.

When we reorder the genes according to their position in the dendrogram, i.e. rearrange the rows of the expression data matrix according to the order imposed by the clustering process, the color-coded-matrix of Figure 3B is obtained.

Next, we marked the positions of the members of $G(38)$ by red crosses. All 38 are in the low-level cluster

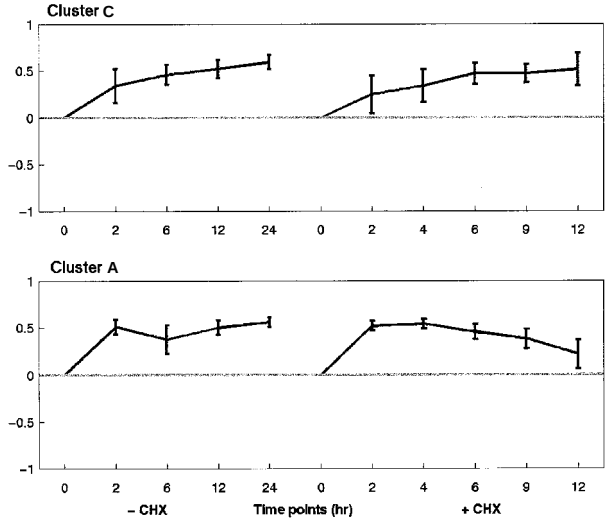



**Figure 4** Average expression profiles of genes in clusters c and a of Figure 3A. The expression profile of each gene in the cluster was normalized as described in the text

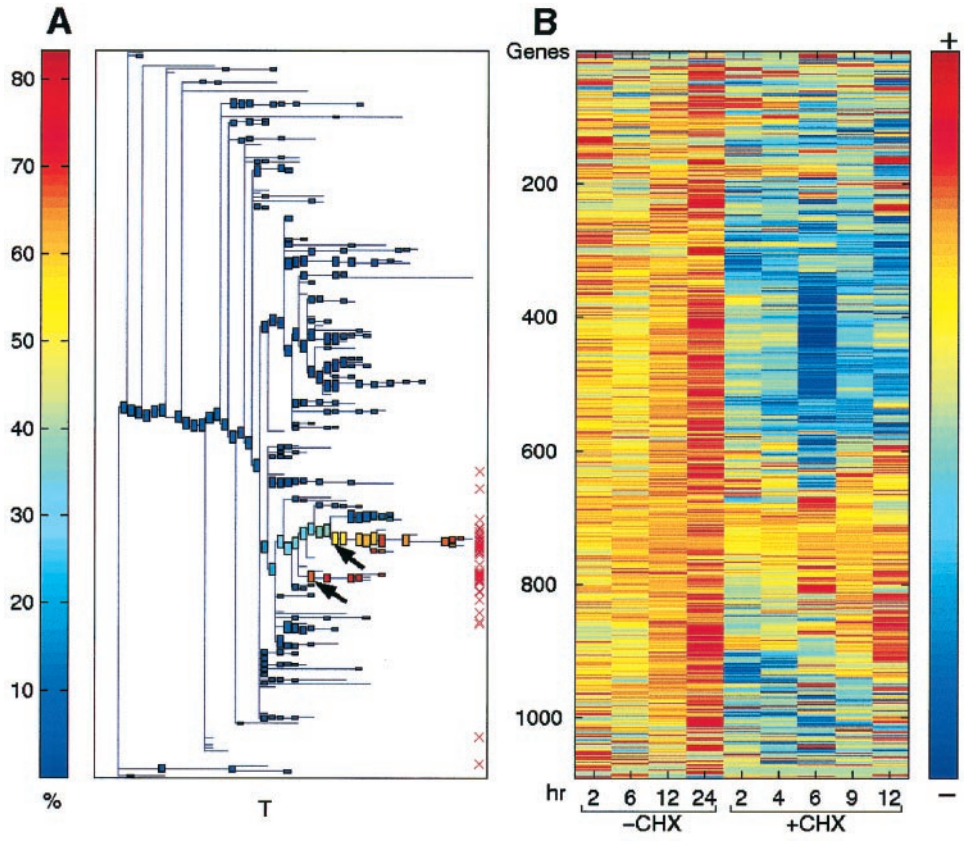



**Figure 5** Clustering results using superparamagnetic clustering (SPC) for 1090 genes that were upregulated in the experiment without cycloheximide. In this analysis a relaxed filtering condition was used and all the genes that were upregulated 2.5-fold at least once (1090) in the experiment without CHX were included. (**A**) Dendrogram of the genes including clusters of size 5 and larger. The distribution of the 47 primary genes is marked by red crosses at the right. (**B**) Normalized log ratio of the nine experiments are plotted. The color represents induction (red) or repression (blue). Other details as in Figure 3. Note the primary gene-containing cluster resolves into two distinct clusters (marked by arrows) by splitting away the non-primary gene containing clusters

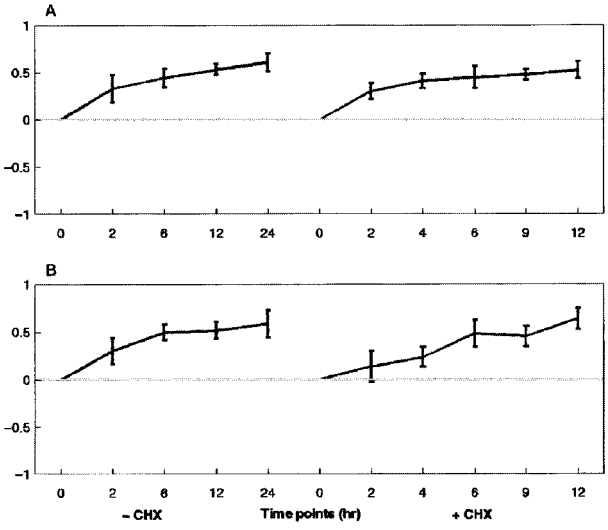



**Figure 6** Average expression profiles of genes in clusters of Figure 5. (**A**) genes from the upper arrow cluster. (**B**) genes from the lower arrow cluster (Figure 5A). The expression profile of each gene was normalized as described in the text

denoted by *b* which, however, contains also 58 additional genes. This cluster branches and breaks into sub-clusters which have higher percentages of genes from *G(38)*. In order to identify a characteristic primary expression profile we want to work with a cluster which has many members of *G(38)* (for better statistics) and also has a high percentage of them. These two requirements conflict; as we move up on the dendrogram, the clusters become purer, but also decrease in size. Hence we decided to start with the cluster *c* marked by the arrow as our working point. It contains 33 of the *G(38)* genes and, in addition, 23 genes that did not pass the original filter, but have expression profiles that are similar to those of *G(38)*.

The average expression kinetics of the genes of this cluster in the two experiments is shown in Figure 4. Note the fairly similar kinetics with and without CHX, with the expression level increasing monotonously with time. All but nine genes of *c* share these features of the expression kinetics. The nine which differ appear on the dendrogram in the vicinity of the cluster denoted by *a*; two of these belong to *G(38)*. The average expression kinetics of these nine genes is shown in

Figure 4 (bottom panel); it clearly differs from that of Figure 4 (top panel), and their kinetics with and without CHX are different. Hence we decided to discard these nine genes from our list of designated primaries. The average expression kinetics of the remaining 47 genes is very similar to Figure 4, with reduced scatter (error bars). The group of these 47 genes, denoted *G(47)*, constitutes our final set of proposed primary p53 targets (Table 1).

This analysis identified a characteristic kinetic profile of primary p53 targets in either the presence or absence of CHX; the group of genes that share this profile contains 31 out of our 38 original candidates and 16 additional ones, that happened to fail our stringent filtering criteria. The various genes are listed and the groups to which they belong are properly identified in Table 1.

Some of the known p53 target genes such as *Gadd45* (Zhan *et al*., 1998) and *PCNA* (Shivakumar *et al*., 1995), were now included as primary targets in addition to the original 38 primary target genes, indicating that this is a sensible way to 'fish' for further potential primary targets. Most of these added genes exhibited $\sim$twofold induction at several of the time points both in the presence or absence of CHX experiments and were previously not known to be p53 target genes.

Next, we have put the stability of our identification of primary p53 targets against changing the procedure and parameters of selection to an extremely demanding test. We performed a similar cluster analysis on a much larger set of genes, including now all those that were upregulated at least twofold, at (at least) a single time point of the experiment without CHX. This very relaxed criterion selected 1090 genes to cluster; that is, as compared to the previous clustering analysis, the number of genes was increased fourfold, including now extremely noisy expression data. The results obtained when these 1090 genes were clustered are presented in Figure 5. To our satisfaction, we found that 3/4 of our 47 proposed primary p53 targets (that were identified above), belong to two very stable gene clusters, denoted by arrows on Figure 5A. The left of the two contains 24 of the genes of *G(47)* and 20 new genes, whereas the right one contains 11 from *G(47)* and seven new associated genes. The average expression kinetics of the two clusters is shown in Figure 6.

It is important to understand that adding all these extra noisy genes to the set of 259 that were analysed before could well have resulted in a total loss of the signal that was identified above. The fact that the two clusters that are rich in previously identified primaries indeed contain 35 out of the 47 is gratifying and indicates the stability of our method.

Finally, we turn to discuss our choice of the clustering algorithm to be used (SPC) (Blatt *et al*., 1996; Getz *et al*., 2000). The optimal algorithm for analysis of gene expression data should have the following properties: the number of clusters should be determined by the algorithm itself and not externally prescribed (as is done for SOM and K-means) (Sherlock, 2000; Young, 2000); stability against noise; generating a hierarchy (dendrogram) and providing a mechanism to identify in it robust, stable clusters; ability to identify a dense set of points, that form a cloud of an irregular, non-spherical shape, as a cluster. SPC, a hierarchical clustering method recently introduced by Blatt *et al*. (1996) is the algorithm that best fits these requirements.

## Discussion

In this study we have used the temperature sensitive p53 expressed in the lung cancer cell line H1299 to analyse p53-regulated genes utilizing oligonucleotide microarray containing 7070 probes. We employed inhibition of protein synthesis by CHX in an effort to limit the p53-regulated genes to primary targets by exclusion of possible secondary effect by newly synthesized proteins. By measuring gene expression at four or five time points we clearly show that a group of genes show consistent p53-dependent regulation in either presence or absence of CHX. This group consists of less than 20% of the genes regulated by p53 in the absence of CHX, and defined here as primary targets. The criteria used to group these genes (2.5-fold change in at least three time points) is somewhat arbitrary and may discriminate against genes which show consistent changes but did not reach the 2.5-fold change in all the time points analysed. We therefore used a clustering method which allows the grouping of genes based on the kinetics of their expression and showed that most of the primary genes group together in one or two clusters. These clusters contain additional genes which show similar kinetic behaviors in their change of expression even though their level of expression did not search the stringent criteria. This group contains some of the established targets of p53 (e.g. *Gadd45* and *PCNA*) and demonstrate the usefulness of this clustering method, since it overcame small experimental variation in measurement of hybridization intensity and relies on the pattern of expression in the presence or absence of CHX.

In order to compare our results with known p53 targets we used the list compiled by El-Deiry (1998). Some of the genes in this list are not present on the chip (*Bax*, *Killer/DR5*, *PAG608*, *14-3-3σ* and *B99*). From the rest of the p53 targets on this list *p21waf*, *Fas/Apo*, *Gadd45* and *PCNA* were detected as primary target genes (Table 1). *MDM2* (a known p53 target) expression was increased more than 2.5-fold after 4 h in the presence of CHX but reached 2.67-fold expression only after 24 h without CHX. *CyclinG* expression level reached 2.4-fold only after 12 h in CHX and 2.1-fold after 12 h without CHX. *Cathepsin D*, *Thrombospondin*, *IGFBP5* and *PI3* kinase were not over pressed in this cell line in our experiment. This type of heterogeneity of expression indicates that not all p53-targets are activated in all conditions and a similar conclusion was pointed out also by other studies (Yu *et al*., 1999; Zhao *et al*., 2000). We also tried to analyse for the presence of p53 consensus

targets in several of the genes listed in Table 1. This was performed by searching for the sequence of the two consensus decamers RRRCWWGYYY, separated by a gap of 0–13 nucleotides, in the sequence which is 2 kbp upstream and downstream to the RNA start. In our list of primary targets, seven genes were shown previously to have consensus sequence and their promoters respond to p53; 15 genes were found to have consensus p53 targets with up to two mismatches and five genes did not have the genomic sequence available yet (data not shown). We feel that database search that is not supported by experimental analysis of individual genes is not very informative at present.

Our results from the analysis of p53 regulated primary target genes show that p53 directly upregulates proapoptotic and cell cycle inhibitors (e.g. *Bak* and *p21waf*) (see Table 1) and downregulates antiapoptotic and cell cycle genes (e.g. Inhibitor of Apoptosis and *Cyclin E*) (see Table 2). Thus p53 seems to activate concerted opposing signals to control cell proliferation and apoptosis. Although the identified primary genes cluster together, indicating the similarity in their expression kinetics, functional classification of these genes revealed that they are very different in their cellular activities (Table 1). The genes involved in cell cycle, growth arrest, apoptosis and perhaps DNA repair seem to be activated in most cell types analysed (e.g. *p21waf*, *Fas/ApoI*, *bcl2* binding protein) (Zhao *et al*., 2000; Komarova *et al*., 1998). On the other hand, genes involved in many other cellular functions like cytoskeleton, extracellular matrix (ECM), growth factors and their receptors and signal transduction may be more specific to the cell type analysed and this may be one of the causes for different lists of p53 regulated genes in various cells lines as was pointed out previously (Zhao *et al*., 2000; Yu *et al*., 1999). For example, in H1299 cells we found activation of several genes known to be involved in neuronal growth (Table 1) which may be unique to this cell type.

Several reasons may account for the heterogeneity of p53 target genes. It is possible that some of the targets require additional factors as coactivators. For example the transactivation of the mismatch repair gene *MSH* requires both p53 and c-*jun* or UV irradiation (Scherer *et al*., 2000). An additional factor may be the level of p53 activation by phosphorylation and other post-translation modification. Different transactivation properties may be related to a different level of p53 modification (Vousden, 2000; Oda *et al*., 2000).

We have noted before the induction of TGF$\beta$ by p53 (Kannan *et al*., 2000) and another study confirmed that *TGF-$\beta$* is a primary target of p53 and demonstrated its growth suppressor activity (Tan *et al*., 2000). This was also evident in our current experiments and may potentially suggest an additional paracrine mechanism by which p53 exerts its tumor suppressive effects on neighboring tumor cells. Other genes in our list of primary genes like endoglin (a coreceptor of TGF-$\beta$) and versican may also be part of the *TGF-$\beta$* activating system (Massague *et al*., 2000). Interestingly, the tumor suppressor activity of TGF-$\beta$ may be related mainly to its function as an inducer of p15 (*ink4b*) expression (Hannon and Beach, 1994). The p15 is an inhibitor of cdk4 and cdk6 and its overexpression may lead to G1 arrest (Massague *et al*., 2000), thus TGF-$\beta$ may behave as a tumor suppressor (Tang *et al*., 1998). The induction of *p21waf* by p53 and *p15* by TGF-$\beta$ which itself is a target of p53 may provide an efficient control on the cell cycle. In the H1299 cell line, TGF-$\beta$, IGF binding protein 6, thymosin beta, diacylglycerol kinase alpha and neuronal growth related genes were substantially induced and were among the primary targets suggesting that these genes may also be relevant to the tumor suppressive phenotype induced by p53 as are the classical cell cycle related p53 inducible genes. Further analysis and cataloguing of p53 primary target genes from various cell lines may explain how p53 orchestrates diverse cellular signals as a tumor suppressor and could possibly lead to the identification of potential therapeutic molecular targets for therapy in different cellular targets. The approach presented here of defining primary targets of p53 is an important additional step in sorting out p53 targets at various cellular contexts.

## Materials and methods

### *Cell lines, cycloheximide treatment and total RNA isolation*

The human lung cancer cell line H1299 (lacking endogenous p53) expressing the mouse temperature sensitive mutant p53Val135 (a gift of Dr M Oren) which on temperature shift to 32°C will assume wild-type p53 conformation was used in this study. The control cell line used was the parental H1299 cells without ts-p53Val135. The cells were maintained at 37°C in RPMI medium containing 10% fetal bovine serum (FBS). Where specified, the cells were exposed to cycloheximide (10 $\mu$g/$\mu$l) 30 min prior to temperature shift. The cells were temperature shifted to 32°C, and after specific time intervals, harvested in TRIZOL solution (Gibco BRL, USA) and total RNA was isolated as per manufacturer's instructions.

### *Northern and Western-blot analyses*

Equal amounts (10 $\mu$g) of total cellular RNA were used for Northern-blot analysis, using standard protocols (Ausubel *et al*., 1990). Autoradiography was done on Kodak X-ray films with intensifying screens at −70°C. For Western-blots, total cell lysates were prepared by lysing cells in NP40 lysis buffer (0.5% NP40, 150 mM NaCl, 50 mM Tris-Cl, pH 7.5, 1 mM DTT, 25 ng/ml aprotinin, 25 ng/ml leupeptin and 1 mM PMSF). Equal amounts (60 $\mu$g) of total protein were fractionated by SDS–PAGE on a 15% gel and transferred to nitrocellulose membranes (BA.85, Schleicher and Schuell). The primary antibodies used was a mixture of mAbs 421/248 for p53, Ab C19 (Santa Cruz) for p21 and a mixture of mAbs 4B2/2A9 for human MDM2 protein. The secondary antibodies used were HRP goat anti-mouse IgG (Sigma) and HRP goat anti-rabbit IgG (Jackson Laboratories).

### *Preparation of labeled cRNA and hybridization of microarrays*

Total RNA was isolated from H1299val135 cells incubated at 32°C for 2, 6, 12, 24 h and H1299 control cells at 32°C for 2, 12 h. H1299val135 cells at 32°C in the presence of

cycloheximide were harvested only up to 12 h at 2, 4, 6, 9 and 12 h time points and H1299 control cells with cycloheximide at 32°C were harvested at 2 h time point and total RNA was isolated. Biotin labeled cRNA was synthesized and hybridized as described (Kaminski *et al*., 2000) to the Genechip HuGene FL array (Affymetrix, Santa Clara, USA) which contains probes for ~7000 human genes. Scanned output files were visually inspected for hybridization artifacts. Arrays lacking significant artifacts were analysed using Genechip 3.3 software (Affymetrix). Arrays were scaled to an average intensity of 1200 per gene and analysed independently. The expression value for each gene was determined by calculating the average of differences (perfect match intensity minus mismatch intensity) of the probe pairs in use for this gene. Ratios were determined by dividing the average difference of H1299Val135 cells for each time point to that of H1299 control cells at 2 h time point with or without cycloheximide in the respective experiments. A value of 100 was assigned to all measurements lower than 100. The expression data for all the hybridization experiments can be obtained from the corresponding author upon request.

*Clustering analysis*

Clustering analysis was performed twice; first to the 259 genes that were upregulated over 2.5-fold at least three times in the absence of CHX (in at least one of the three experiments, the gene had to be 'present' in the Present/Absent call provided by Affymetrix software). The second clustering analysis was applied to the 1090 genes that were upregulated over twofold at least once, in the absence of CHX experiment (also in that experiment it should be called 'present'). In both cases, each gene was represented by nine (four experiments without CHX and five experiments with CHX) values representing the log ratio of the average difference in the experiments to the average difference in their corresponding control. The log ratios were then normalized for each gene by dividing separately for the two experiments. The clustering algorithm measured the distance between the genes using the regular Euclidean distance between their normalized values. Note that this distance measure is more suitable than the Pearson correlation between the log ratios, since the latter may place induced and repressed genes in the same cluster.

**Acknowledgments**

We are grateful for the Arison Dorsman family's donation to the Center of DNA Chips in Pediatric Oncology. This work was supported in part by the Yad Abraham Research Center for Cancer Diagnosis and Therapy, the Rich Foundation for Leukemia Research and the Germany-Israel Science Foundation (GIF). G Rechavi holds the Gregorio and Dora Shapiro Chair for Hematologic Malignancies, Sackler School of Medicine, Tel Aviv University. We are thankful to the Crown Human Genome Center of the Weizmann Institute for the help received.

**References**

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K. (1990). *Current Protocols in Molecular Biology*. Green/Wiley Interscience, New York.

Blatt M, Wiseman S and Domany E. (1996). *Physical Rev. Lett.*, **76,** 3251–3254.

Coller H, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN and Golub TD. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 3260–3265.

El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW and Vogelstein B. (1992). *Nat. Genet.*, **1,** 45–49.

El-Deiry WS. (1998). *Sem. Cancer Biol.*, **8,** 345–357.

Getz G, Levine E, Domany E and Zhang MQ. (2000). *Physica A.*, **279,** 457–464.

Ginsberg D, Michalovitz D, Ginsberg D and Oren M. (1991). *Mol. Cell Biol.*, **11,** 582–585.

Hannon GJ and Beach D. (1994). *Nature,* **371,** 257–261.

Haupt Y, Maya R, Kazaz A and Oren M. (1997). *Nature,* **387,** 296–299.

Jimenez GS, Nister M, Stommel JM. Beeche M, Barcarse EA, Zhang X-Q, O'Gorman S and Wahl GM. (2000). *Nat. Genet.*, **26,** 37–43.

Kaminski N, Allard JD, Pittet JF, Zuo F, Griffiths MJD, Morris D, Huang X, Sheppard D and Heller RA. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 1778–1783.

Kannan K, Amariglio N, Rechavi G and Givol D. (2000). *FEBS Lett.*, **470,** 77–82.

Komarova EA, Diatchenko L, Tokhlin OW, Hill JE, Wang ZJ, Krivokrysenko VI, Feinstein E and Gudkov AV. (1998). *Oncogene,* **17,** 1089–1096.

Kubbutat MH, Jones SN and Vousden KH. (1997). *Nature,* **387,** 299–303.

Levine AJ. (1997). *Cell,* **88,** 323–331.

Massague J, Blain SW and Lo SR. (2000). *Cell,* **103,** 295–309.

Michalovitz D, Halevy O and Oren M. (1990). *Cell,* **62,** 671–680.

Michieli P, Chedid M, Lin D, Pierce JH, Mercer WE and Givol D. (1994). *Cancer Res.*, **54,** 3391–3395.

Oda K, Arakawa H, Tanaka T, Matsuda K, Tanikawa C, Morit T, Nishimori H, Tamai K, Tokino T, Nakamura Y and Taya Y. (2000). *Cell,* **102,** 849–862.

O'Hagan RC, Schreiber-Agus N, Chen K, David G, Engelman JA, Schwab R, Alland L, Thomson C, Ronning DR, Sacchettini JC, Meltzer P and DePinho RA. (2000). *Nat. Genet.*, **24,** 113–119.

Scherer SJ, Maier SM, Seifert M, Hanselman RG, Zang KD, Muller-Hemerling HK, Angel P, Walter C and Schartl M. (2000). *J. Biol. Chem.*, **275,** 37469–37473.

Sherlock G. (2000). *Curr. Opin. Immunol.*, **12,** 201–205.

Shivakumar CV, Brown DR, Deb S and Deb SP. (1995). *Mol. Cell Biol.*, **15,** 6785–6793.

Tan M, Wang Y, Guan K and Sun Y. (2000). *Proc. Natl. Acad. Sci. USA,* **97,** 109–114.

Tang B, Bottinger EP, Jakowlew SB, Bagnall KM, Mariano J, Anver MR, Letterio JJ and Wakefield LM. (1998). *Nat. Med.*, **4,** 802–807.

Vousden KH. (2000). *Cell,* **103,** 691–694.

Young RA. (2000). *Cell,* **102,** 9–15.

Yu J, Zhang L, Hwang PM, Rago C, Kinzler KW and Vogelstein B. (1999). *Proc. Natl. Acad. Sci. USA,* **96,** 14517–14522.

Zhan Q, Chen IT, Antinore MJ and Fornace AJ. (1998). *Mol. Cell Biol.*, **18,** 2768–2778.

Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH and Levine AJ. (2000). *Genes Dev.*, **14,** 981–993.

# Chapter 3

# Gene Expression Applications

## 3.1 Introduction

This section serves as an introduction to Publications (5)-(10) that describe projects in which we applied the methods discussed in the previous Chapter to analyze gene expression data. First I give a brief overview of the types of questions asked in this field and in particular in our projects. Following that I list each of the projects separately.

Projects that involve gene expression analysis can be grouped according to the kind of questions being asked. First I introduce these groups and then give a more detailed description of typical studies as well as ours divided according to these groups. Since gene expression experiments are expensive many research groups try to use their data to answer questions from different categories.

### 3.1.1 Genes, gene-networks and pathways

The first category consists of questions regarding *genes* and *gene-networks* and fall in the realm of functional genomics. In this category, the aim is to gain understanding on how the cell functions, assign function to genes and decipher the gene regulatory networks. Clustering methods (see Section 2.2.2) are used to generate clusters of genes with correlated gene expression that are assumed to be co-regulated or at least belong to the same biological process or pathway. Such clusters can, therefore, hint on the function of the genes in the cluster by means of "guilt by association"; genes of unknown function are likely to have similar or related function to other genes in the cluster.

In addition, a group of co-regulated genes can help identify the transcription factors that control the process and even detect binding sites of yet unknown transcription factors. Such analyses involve exploring the promoter regions of these genes in the genome.

The experiments that are performed in this category study cells taken from different conditions or taken at different time points along some controlled process. Naturally, such studies began with simpler model organisms such as bacteria and yeast, and later advanced to human cell lines. Early studies in this category are Chu *et al.* [139] and Spellman *et*

*al.* [63] who studied yeast cell cycle. A major contribution to this field and especially to the development of analysis methods is the fact that these studies made their data public, allowing other researchers to test their methods on the same data. Publication (5) describes our work [140] in which we identified co-regulated gene clusters in yeast data and used them to find new regulatory sites in the genes' upstream regions. We addressed this problem by applying the super-paramagnetic clustering method (SPC) [74] to yeast expression data taken at several points along the cell cycle. Alter *et al.* used singular value decomposition (SVD) [55] to study the yeast cell cycle data and ordered the yeast genes according to their the phase of the cell-cycle in which they are activated. The group of N. Barkai collected a compendium of yeast expression data which was used to identify regulator motifs by applying the signature algorithm. Later Bergman *et al.* bergman02 showed that iterating the signature algorithm gives rise to a robust version of the SVD algorithm.

Publication (4) [141] describes our work on identification of genes which are primary and secondary targets of p53 (a key protein in most cancers). In this work we used cluster analysis, in a novel way, as a robust filter. It was used to eliminate the effect of arbitrary parameter values that were chosen in the preliminary step of the analysis.

### 3.1.2 Molecular differences between conditions

The next category focuses on comparing two or more *conditions* asking which genes are differentially expressed among them. The aim of these questions is to try to understand the different biology of the studied conditions on the molecular level; e.g. which pathways are active, which genes are the key players. Experiments of this type are most common in cancer research where one wants to learn about the differences between normal and malignant tissue or between tumors at different stages of development. Such studies can help identify new leads for therapeutic targets. These questions are addressed using statistical hypothesis testing in which one has several examples from the various conditions and looks for differences among them that are larger than expected by chance.

A related category of questions, that often appears together with the previous one, is *class prediction* and *class discovery*. A typical scenario for a type identification (classification) problem is the development of a prognostic tool. For example, if one wants to use the expression level of several genes to predict, at time of diagnosis, if a cancer patient will develop metastases at early or late stage (see Publication (10)) or to predict the sub-type of a tumor (see Publication (8)). Such predictions may have direct influence on prognosis and choice of therapy. These questions are addressed using supervised machine learning techniques (see Introduction and Chapter 2).

Tumors come in many variants and in many cases the current state of diagnosis and taxonomy groups together many different tumor types to a single one. An important step prior to development of a prognostic tool is the identification of the different types and sub-types of the disease (see Publication (8)). This type of class discovery questions are addressed by unsupervised or clustering methods (see Section 2.2.2). In some cases, even

when the tumor type is given, it is recommended to apply unsupervised techniques as a data cleansing step since tumor identification can be subjective and is prone to errors **??**.

The first large scale gene expression analysis of tumor samples was published by Alon *et al.* [71] in which 40 colon cancer samples and 22 normal colon samples were analyzed using oligonucleotide microarrays. The resulting expression matrix of $\sim 6000$ genes (rows) by 62 samples (columns) was used to ask questions from several categories. They used a global two-way approach in which two independent cluster analyses were performed; one of the genes and the other of the samples. Clustering of the samples succeeded to separate the tumors from the normal samples with few errors - a type discovery task, and some gene clusters grouped together genes which are known to belong the the same process - can be used for gene function assignment and biological insight. Alizadeh *et al.* [8] took a similar approach to analyze diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, and identified subtypes with distinct molecular profiles indicative of the differentiation stage of the B-cell. The subtypes had different survival distributions.

Such global analyses risk missing signals that involve small subsets of the "relevant" genes or samples, since these are masked by the contribution of the remaining (irrelevant but numerous) genes and samples. We proposed a method, called coupled two-way clustering (CTWC) [108] (Publication (1)), that iteratively finds groups of genes and samples and searches for structural signals in sub-matrices defined by these groups. CTWC was used successfully to study data from experiments on colon cancer, leukemia [108] and [113](Publication (9)), breast cancer [142] and [143] (Publications (5) and (10)), glioblastoma [144] (Publication (8)) and skin cancer [145]. The method is general and we and others also applied it to data from other domains: antibody reactivity analysis [102] (Publication (6)) and analysis of glycomolecules [101]. Note that for each new problem domain one should choose the appropriate preprocessing and distance measure between rows and columns of the matrix.

The outline of this chapter is as follows: First I describe gene expression technology and then I briefly describe the main results in Publications (5)-(10).

## 3.2 Microarray technology

The main goal of microarray technology is to measure, in a single experiment, the mRNA expression level of as many genes (or transcripts[1]) as possible in a cell population or tissue. The measurement is based on a *hybridization* reaction. Other technologies also play an important role in the process are reverse transcription and labelling. Below I briefly describe the technologies and the general experimental scheme and in the next sections the main two microarray technologies, synthetic oligonucleotide microarrays (Affymetrix GeneChips) and cDNA microarrays, are discussed. More details regarding these technologies can be found

[1] a single gene can have alternatively spliced mRNA variants

in [6, 146] and references therein.

### Hybridization

Hybridization is the process of attachment of complementary single stranded DNA or RNA molecules. Both double stranded DNA, RNA or DNA-RNA compounds are bound together by hydrogen bonds between matching base pairs, A-(T or U) and C-G. When a double stranded compound is heated beyond its melting temperature (65°C for double stranded DNA) or put in certain chemical environments, the two strands dissociate. As the temperature is reduced back below the melting temperature (or the original chemical environment is assumed), the single stranded molecules meet and bind back to their counterparts. Binding is based on base pairing; thus the affinity of hybridization of two perfectly matching strands is the highest. Non-specific hybridization can also occur but with significantly lower affinity.

### Reverse transcription

mRNA molecules extracted from a cell can be reverse transcribed[2] to produce a complementary DNA called cDNA. The cDNA is first single stranded but can later be doubled by *in vitro* DNA synthesis using a DNA polymerase. If one wants to work with RNA which has higher affinity to DNA, one can perform *in vitro* transcription (IVT) which generates a complementary RNA, called cRNA, based on the cDNA template.

### Labelling

Labelling of cDNA or cRNA is done during their synthesis by introducing labeled bases. In cDNA microarray experiments one labels the cDNA by using fluorescently labeled UTP (U-base) which replaces the unlabelled TTP. In the Affimetrix assay, cRNA is labelled by using biotin attached to UTP and CTP.

### A microarray experiment

A microarray experiment has two steps; microarray preparation and utilization. In general the microarray is prepared by attaching in each element of the array many identical single-strand DNA segments taken from a specific gene (or transcript). Each of the $10^3$ to $10^5$ elements serves as a detector for a different gene and is called a *probe*. In order to use the microarray, one first extracts the mRNA molecules from a sample cell population, reverse transcribe and label them to produce the *targets*. The targets are then poured onto the microarray and left to diffuse until they hybridize to their matching probes. After washing the unattached targets, the amount of labelled targets bound to each probe-element is read by measuring fluorescent (or radioactive) emission yielding an indirect measure of the

[2]Reverse transcription is performed by special enzymes originating from retro-viruses that transform their RNA to DNA and insert it to the host genome.

expression level of the specific gene in the examined cells. The scanned measurements are stored as an image file on a computer. Next, an image analysis software identifies the array elements in the image and separates them from the background. The final readings for each gene are estimated after taking into account the probe and background signals.

## 3.3 synthetic oligonucleotide microarrays

Affymetrix's GeneChip® arrays are high-density synthetic oligonucleotide microarrays which are manufactured using photolithographic methods adapted from the semiconductor industry. Short synthetic segments of cDNA (25 bases long) are fabricated, one nucleotide at a time, onto a silicon surface. As the technology improved new chips were produced capable of detecting an increased number of gene transcripts. Today, a single chip, the Human Genome U133, can detect as many as 33,000 different transcripts. The chip, $(1.28\text{cm})^2$ in size, is divided into square cells of $(18\mu\text{m})^2$ which are the probes.

On each probe there are millions of copies of the same specific 25-mer oligo[3]. Since short oligos suffer from relatively large amounts of cross-hybridization, the expression level of a gene is detected using a set of 11 to 16 probes, called probe-set, each representing a carefully selected sequence out of the gene's mRNA. The sequences are chosen by Affymetrix aiming to optimize their specificity, *i.e.* making sure the sequence is unique even among similar but unrelated sequences, and their sensitivity, *i.e.* their predicted hybridization properties are most sensitive at the working conditions [147]. Chalifa-Caspi et al. [92] showed that using updated gene sequences (from RefSeq [148], Ensembl [149] and GeneBank [148]) not all probe-sequences match their designated genes and, moreover, some probe-sequences match several genes. In order to estimate and later subtract the non-specific hybridization level and background signal, probes come in pairs; a perfect match probe (PM) and its corresponding mismatch probe (MM). The PM is a perfect complement of the target sequence and the MM is identical to the PM except for the base pare in the middle.

The target is prepared by extracting mRNA from the cells by either direct detection of the poly-A tale of the mRNA or by first extraction of total RNA and then purifying the mRNA out of it (which is only 3% of the total RNA). The amount of mRNA needed is at least $0.2\mu g$ poly-A mRNA or $5\mu g$ total RNA which can be extracted from about $10^6$ cells. The mRNA is reversed-transcribed into cDNA using a T7 promoter-tailed oligo-dT primer, which is then made double-stranded (by a DNA polymerase) and finally converted to labeled cRNA by *in vitro* transcription in the presence of biotinylated UTP and CTP. The cRNA is fragmented prior to hybridization to fragments of 35 to 200 bases [147]. A hybridization cocktail is prepared, including the fragmented target, probe array controls (used for controlling the quality of the target preparation, hybridization, washing and staining steps and calibration of the image analysis software), and few additional reagents. Hybridization takes place for 16 hours at 45°C. After the hybridization, the solution is re-

[3]abbreviation for oligonucleotide

moved and the microarray is washed and stained with Streptavidin Phycoerythrin (SAPE) that attaches to the biotin labeled cRNA molecules. The last step is scanning the microarray using a confocal laser scanner that shines at each probe a 570nm laser and captures the fluorescent emissions of the excited SAPE molecules.

From this point on all the analysis is performed on a computer (*in silico*). Affymetrix supplies a software (MAS) that performs the image analysis and combines the readings from all the probes in a probe-set and generates a single expression level estimate for each transcript.

## 3.4 cDNA microarrays

The cDNA microarray were first manufactured in Brown's lab in Stanford [150]. Before producing a cDNA microarray, one needs to select the probes to be spotted on the array and prepare sufficient quantities of them. Usually the cDNA library is bought or generated in-house and amplified using polymerase chain reaction[4] (PCR). A spotting robot dips an array of tips in a multi-well plate and drops circular spots on a glass slide; each spot corresponds to a specific cDNA.

In cDNA experiments the target is a mixture of two samples, usually, an experiment and a control, each labelled with a different dye. The target is prepared by first extracting the mRNA (as done for the synthetic oligonucleotide microarrays) and then labelling during reverse transcription (RT). One sample is fluorescently marked with Cy3 (green) and the other with Cy5 (red). The target mixture is then poured on the microarray and left to hybridized at 45° for 24 hours. During the hybridization the Cy3 and Cy5 marked cDNAs compete on binding to the probe. Since the hybridization properties of each probe are different (cDNA probes are of different length and base-composition yielding variation in melting temperatures) and there is a competition between the targets only ratio measurements are performed.

Finally, the emission is read by a scanning microscope in two sweeps, one for each wavelength. The resulting images are analyzed; the spots are separated from their background and a measurement for each spot and its background in both channels (Cy3 and Cy5) is supplied. Cy3 and Cy5 have different labelling efficiency hence one needs to scale the reading from one channel compared to the other. After subtracting the background signal and scaling, the ratio between the readings of a particular probe is an estimate of the true ratio of its mRNA expression levels in the two samples.

[4]Polymerase chain reaction (PCR) is an iterative process in which a DNA polymerase doubles the DNA quantity in each iteration.

## 3.5 Antibody reactivity measurements

Antibodies are proteins (immunoglobulins) of the immune-system that are present in the blood that identify and help target a response to an antigen (a small molecule, usually a peptide). Measuring antibody reactivities, *i.e.* the concentration of antibodies in the blood that attach to a specific antigen, is moving towards high throughput and large-scale experiments. In these experiments one wants to measure the antibody reactivity to as many as possible different antigens, thus providing a global view of antibody spectrum. As in gene expression, this global view, the vector of concentrations, is believed to represent the state of the immune-system and the entire organism [151]. In Publication (6), the reactivities of IgM and IgG (two type of antibodies) to 87 different antigens is measures by performing simultaneous (ELISA) measurements in a 96-well plate. The surface of each of the wells is coated by a different antigen. Antibodies extracted from a patient's serum are poured to the wells; those that bind to the antigen present in a well with a non-negligible reactivity will stick. These are referred to as *primary antibodies*. After washing the unbound primary antibodies, a fluorescently marked *secondary antibody* is added. This secondary antibody recognizes the non-active and non-varying "leg" of the primary antibodies. Measuring the concentration of the marked antibody in each well reflects the amount of bound antibodies to the specific antigen. This procedure is relatively accurate and is reproducible within 10% (see Publication (6)). Continuing towards larger-scale experiments, I. Cohen's lab is in the process of manufacturing spotted arrays of antigens and antibodies using the same technology as in cDNA microarrays [152].

## 3.6 Biological results

### 3.6.1 Identification of primary and secondary targets of p53

In this work, done in collaboration with David Givol's group, we used clustering to analyze microarray experiments and identify primary and secondary target genes regulated by p53 [141] (see Publication (4)). p53 is the most important tumor suppressor gene. Its function is mainly due to its activity as a transcription factor that activates many genes in response to various types of stress [153]. The genes that are affected by p53 (directly and indirectly) can be detected by replacing the wildtype p53 by a mutant that can be activated by an external signal (temperature decrease in this experiment). Measuring gene expression at 4 time points (2,6,12,24 hours) after p53 activation, and applying strict conditions to identify changing genes, revealed 259 genes that were upregulated. Some of these may be primary targets of p53, whereas others can be affected only indirectly by p53. To sort this out, a similar experiment, measured at 5 time points, was conducted in the presence of cycloheximid (CHX), which blocks the ribosome and prevents protein synthesis. In this experiment only *primary* (direct) targets of p53 can be affected. The 38 genes that were upregulated in both of the experiments are believed to be primary targets of p53.

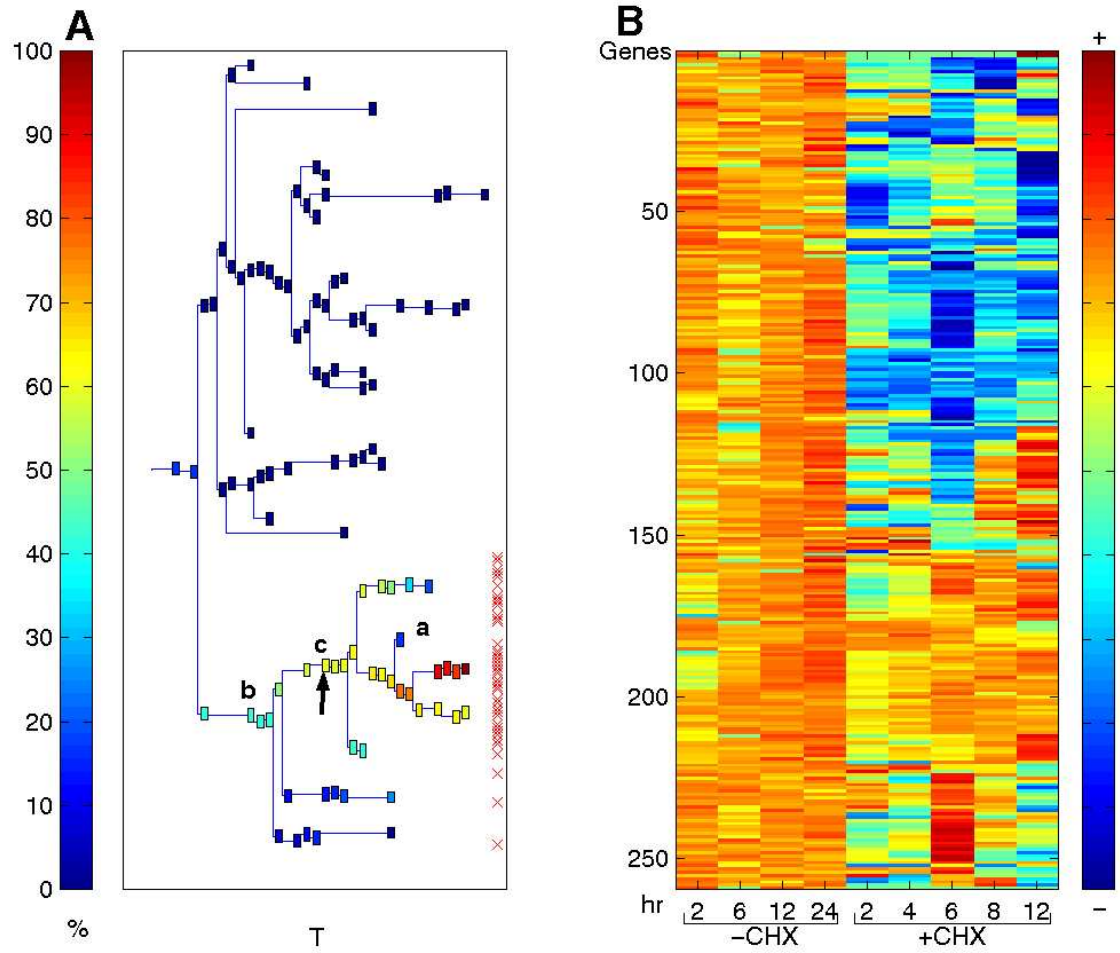



Figure 3.1: Clustering results using SPC for the 259 genes that were upregulated at three time points or more upon activation of p53 in the experiment without cycloheximide (CHX). (A) The dendrogram of the genes showing clusters of size 4 and larger. Each cluster is represented by a box colored according to the percent of the 38 genes found in the preliminary analysis. The distribution of the 38 genes appear as red crosses on the right. (B) The normalized log ratio of the nine experiments (with and without CHX). The genes are ordered according to the dendrogram on the left. The color represents induction (red) and repression (blue). The cluster marked by 'c' contains 87% of the 38 genes together with other genes that share a similar expression profile.

In this analysis we used the clustering results, in a novel manner, as a robust filter. Recall that the set of 38 putative primary targets was identified after strict, but rather arbitrary filters were applied on the gene expression levels. We clustered a much larger set of genes that met much more relaxed conditions and identified the gene cluster that included most of the genes found by the strict conditions (see Figure 3.1). The clustering helped to "fish out" other putative primary genes that do not pass the arbitrary strict conditions but have similar expression patterns to those in the initial set of primary genes. Among these additional p53 targets were known ones such as Gadd45 and PCNA genes.

The use of clustering to perform gene filtering constitutes a robust procedure, much less sensitive to changing the arbitrary filtering parameters used in the initial conditions and also to the experimental noise.

### 3.6.2 Identifying putative transcription binding sites

In this work we performed clustering on (publicly available [154]) yeast cell-cycle data [140] (Publication (5)). Eisen *et al.* [62] clustered the genes on the basis of data *combined* from 7 different experiments. In the same spirit of divide–and–rule that guided us in the development of CTWC, we suspected that mixing the results of different experiments may introduce noise into the data associated with a single one. Therefore we chose to use only the gene expression from a single process (cell division cycles following alpha-factor block-and-release [63]).

We clustered the expression profiles of the 2467 yeast genes of known function (for easier interpretation) over data taken at 18 time intervals (of 7 min) during two cell division cycles, synchronized by alpha arrest and release. We used the Super-Paramagnetic Clustering [74] (SPC) method to cluster the genes due to its robustness against noise and ability to identify stable clusters.

We focused on clusters whose temporal variation is on the scale of the cell cycle and the mean expression profile of the genes in the cluster is significantly non-constant with respect to the variability within the cluster. Three out of four clusters that satisfied both conditions contained genes with known activity in late G1, G2/M and S cell cycle phases. In addition, we identified three clusters of mainly ribosomal proteins that oscillated with an extraordinary cycle that exactly fit the time between two measurements. Later we found out that the odd and even experiments were performed on different dates, with one of the two sets frozen between harvesting and measurement, on different batches of spotted arrays, which gave rise to this artifact. Even though these "ribosomal oscillations" were merely an artifact, it is gratifying that we were able to discover them while previous analyses of the same data did not.

Zhang searched the upstream promoter region of genes in the stable clusters for conserved regulatory elements (using Gibbs DNA [155]). He found several DNA binding motifs; some were rediscoveries of already known ones and some were new. This type of analysis shows the power of combining results from different experiments (gene-array expression

measurements and DNA sequencing) to find new biologically significant results.

### 3.6.3 Separation of Gliomas into subgroups

Here I describe a work done in collaboration with M. Hegi's group from University Hospital in Lausanne, Switzerland (see Publication (8)). This work analyzes gene expression in different types of gliomas (family of brain tumors) and searches for genes that behave differently among them. Better understanding of the molecular mechanisms involved in gliomas and identification of different tumor classes may serve the development of targeted treatment strategies adapted to individual patients. Here we analyzed gene expression measured by performing cDNA-array experiments to 53 patient biopsies, comprising low-grade astrocytoma (LGA), secondary glioblastoma (ScGBM) which are respective recurrent high-grade tumors, and newly diagnosed primary glioblastoma (PrGBM). Even though PrGBM are indistinguishable from ScGBM by histology, they differ in their progression, genetic alterations and occur in different age groups. We demonstrated that human gliomas can be differentiated according to their gene expression. We found that low-grade astrocytoma have the most specific and similar expression profiles, whereas primary glioblastoma exhibit much larger variation between tumors. Secondary glioblastoma are in between and display features of both other groups. We identified several sets of genes with relatively highly correlated expression within the groups that (a) can be associated with specific biological functions and (b) effectively differentiate tumor class.

In this work we combined unsupervised and supervised analysis. At first the data was separated into a training set, which was used for the analyses, and a validation set which was later used to test our ability to predict the glioma class. The training set consisted of 36 tumor samples; 14 PrGBMs, 5 ScGBMs, 12 LGAs, 3 CLs (Cell lines), 1 RecGBM (recurrent glioblasotma) and 1 OAIII (Oligoastocytoma), for which the expression levels of 1185 genes were measured.

We started with the unsupervised analysis and followed our standard procedure; We first log-transformed the data and then filtered out genes with low variation, leaving 358 genes. Applying CTWC to the 358 by 36 expression matrix produced many gene clusters. Our main findings concentrate on two interesting ones: (i) A gene cluster that when used to cluster all the samples separates primary from low grade and secondary tumors. This cluster contains genes that are involved in angiogenesis, including VEGF and VEGFR1, but also IGFBP2, that has not yet been directly linked to angiogenesis. A dedicated experiment in which angiogenesis was externally induced, indeed showed upregulation of the IGFBP family, validating the clustering results. Figure 3.2 shows the dendrogram produced by clustering the tumors based on this angiogenesis cluster. (ii) A cluster of genes related to the immune system that, when used to clusters the samples, identified two clusters; one which contains mostly primary tumors and the other which contains nearly all the low grade and secondary ones together with 4 primary tumors. These 4 primary tumors were removed from patients with relatively long survival time ($> 87$ weeks, compared to 30 to

50 for others). This cluster suggests that the activity of the immune system is similar in the low grade, the secondary, and those primary tumors which come from patients with relatively long survival.

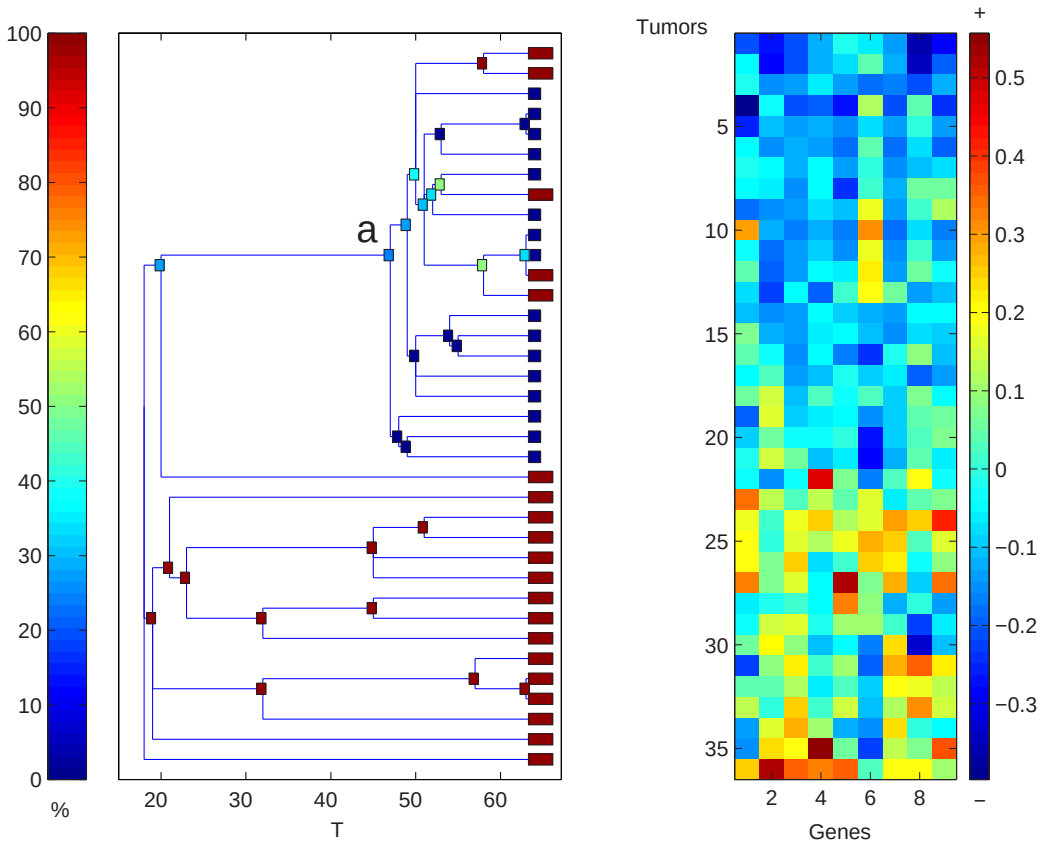



Figure 3.2: One of the clustering results obtained by CTWC for the glioblastoma data. Here all the samples are clustered based on a cluster of 9 genes (some of which related to angiogenesis and growth) which was found in the first iteration of CTWC. Long red boxes on the leaves in the dendrogram represent primary glioblastoma samples. The stable cluster marked by 'a' contains mainly non-primary samples which are characterized by low expression level of these 9 genes (see right matrix).

### 3.6.4 Cluster analysis of human antibody reactivities

In this work, performed in collaboration with Irun Cohen's lab (see Publication (6)), we applied the CTWC method to analyze a matrix containing reactivities of human antibodies to 87 different antigens measured on 40 persons; 20 healthy and 20 with type 1 diabetes mellitus. In this case instead of genes there are antigens and one measures the reactivity or a subject's serum with a set of antigens (i.e. the degree in which the subject's antibodies identify certain antigens). The two major finding of this work are: (i) the reactivity matrix

indeed has non-trivial structure which supports the idea that subjects with unusually similar reactivity patterns that may have a similar immuno-history and (ii) we found 6 significant clusters of antigens, that when used to cluster the subjects were able to distinguish (with few errors) healthy subjects from diabetics. Predicting the subject type by voting among these clusters yields only 3 classification errors.

Figure 2.1 shows two dendrograms of antigens obtained by clustering. The left dendrogram is obtained from the original matrix using sera from healthy and type 1 diabetes subjects, where as the right one is a typical dendrogram obtained by clustering a randomly permuted version of the same matrix. This demonstrates that stable clusters, ones that do not break for large $\Delta T$, are unlikely to be found by chance. The circled clusters are the ones used to distinguish healthy people from diabetics.

### 3.6.5 MLL translocations in Acute Lymphoblastic Leukemia

This work (Publication (9)) on leukemia data was done in collaboration with O. Ravid-Amir and H. Agrawal from our group and with Prof. E. Canaani's group. My role was to introduce the FDR method and to perform, with Osnat, the first phase of the supervised analysis.

The ALL-1 gene is directly involved in 5-10% of ALLs (acute lymphoblastic leukemia) and AMLs (acute myeloblastic leukemia). It is transformed by fusion to other genes or through internal rearrangements. The aim of this work was to study the molecular effects of ALL-1 modifications. The data consisted of 52 leukemia samples which included 25 ALLs, 12 of which carried ALL-1 modifications; 20 AML samples, 10 of which had ALL-1 modifications; 2 ALL and 5 AML cell lines with the ALL-1 modifications. Affymetrix chips were used to measure the expression level of 12600 genes. The main results of this work are: (i) These genes' expression profiles can distinguish ALL and AML tumors that carry the ALL-1 modifications from other ALLs and AMLs. (ii) The expression patterns of ALL-1-associated tumors, in particular ALLs, involve oncogenes, tumor suppressors, anti apoptotic genes, drug resistance genes etc., and correlate with the aggressive nature of the tumors. (iii) The genes whose expression differentiates between ALLs with and without ALL-1 rearrangement were further divided into several groups enabling separation of ALL-1- associated ALLs into two subclasses. (iv) AMLs with partial duplication of ALL-1 vary in their expression pattern from AMLs in which ALL-1 had undergone fusion to other genes.

### 3.6.6 Survival signature

The work presented in Publication (10), done in collaboration with I. Kela and L. Ein-Dor in our group, started by raising the question whether the results obtained and published by one experimental group are valid for data taken from the same type of disease by a different experimental group. Although such *transferability* of results should be at the heart of any scientific research, it is not that obvious in gene expression analysis. Factors that may

affect the results may enter at many stages, from selecting the patients, protocols used for surgery and freezing the samples, RNA extraction, chip technology and analysis methods. We focused on the analysis of breast cancer samples and, in particular, on analysis with respect to survival data. Survival prediction is of great importance in breast cancer since there is a large variation in the progression of the disease in different patients. In the current practice many patients are over-treated by adjuvant therapies, which are both toxic and expensive [156], since one does not know in advance the progression rate of the disease and how long it will take until metastases appear (which are usually the cause of death). Many research groups [43,52,64,80,81,157] are analyzing biopsy-samples taken immediately after initial diagnosis and are trying to identify the "survival signature", *i.e.* a pattern of gene expression that appears in patients with long survival times and hence can be used at an early stage to predict survival. The puzzling fact is that each research group comes up with its own list of genes, and the overlap between lists provided by different groups is nearly zero. This raises the question – what are the factors that cause these inconsistencies.

We started with a simpler version of this question, trying to keep all experimental factors constant, we concentrated on a single experiment, of Van't Veer *et al.* [52]. They analyzed 96 patients who were treated in the same manner. Their aim was to generate a prognostic tool based on gene expression, that can predict survival[5]. They separated the 96 samples to a training set of 77 samples and a test set of the remaining 19 samples, and measured the correlation of every gene with survival, over the training set. Next, a classifier was generated, based on the 70 genes most correlated with survival. This classifier succeeded to predict the outcome of the 19 test samples with only 2 errors. Prior to testing transferability of their classification to other data sets, we asked ourselves: (i) Is this classifier unique? or, can we generate classifiers based on other genes that can perform as well as their one? (ii) How robust is this process of generating a classifier? What classifier would have emerged had we selected a different training and test set?

Publication (10) deals with these questions and demonstrates that for any given partition to 77/19, one can identify several distinct sets of 70 genes (4 on average) whose predictive power is as good as that of the top 70 genes. Particularly, for the partition chosen by Van't Veer *et al.* , we identified 4 additional sets of 70 genes that attain equal or lower error rates both on the training set and the test set. Comparing the Kaplan-Meier curve for the patients (in the test set) predicted to have poor-prognosis with the curve for the ones classified as good-prognosis yielded $p$-values $< 0.01$ in all 4 classifiers[6]. In general, our picture is that many genes have a low but significant correlation with survival and thus different, large enough, sets of genes can overcome the noise and reach a high predictive power.

As for the robustness of the procedure, we show that different partitions to 77/19 give rise to different sets of top 70 survival-correlated genes (with an average overlap of

[5]Van't Veer *et al.* did not predict survival but rather *metastasis-free time interval* (MFTI) which is directly related to survival.

[6]The comparison is performed using the Mantel-Cox log-rank test. See Sec. 2.4.4 for technical details

12.2 genes using bootstrapping tests). One obvious explanation for this behavior is that estimating a gene's correlation with survival based only on 77 samples is still very noisy, and together with the fact that there are many genes with similar correlation, different genes reach the top ranks. Another deeper explanation may be that the group of 77 patients actually contains several sub-types of the disease, each of which has a different set of gene that are correlated with survival. The different composition of sub-types in each of the of 77 samples causes the large deviation in the list of top ranking genes. Some sub-types are already known; there are patients with ER+ (high levels of estrogen receptor protein) and ones with ER− [158] which have different survival-time statistics. Moreover, Sorlie *et al.* [43, 64] identified in an unsupervised manner five groups of patients that have distinct molecular profiles and also different survival curves. Recently, they validated these groups on other data sets including the one of Van't Veer *et al.* . These results may imply that there is no single "survival signature" and one needs to first separate the samples to families (which might be still unknown) and only then search for a survival signature in each family.

# Published Works

The following pages contain the published works related to the preceding chapter.

# Publication 5:

## Super-paramagnetic clustering of yeast gene expression profiles

Authors: G. Getz, E. Levine, E. Domany and M. Q. Zhang
Published in: *Physica A* **279**, 457–464 (2000).

ELSEVIER

Physica A 279 (2000) 457–464

PHYSICA A

www.elsevier.com/locate/physa

# Super-paramagnetic clustering of yeast gene expression profiles

G. Getz[a], E. Levine[a], E. Domany[a], M.Q. Zhang[b,*]

[a] *Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel*
[b] *Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, New York 11724, USA*

**Abstract**

High-density DNA arrays, used to monitor gene expression at a genomic scale, have produced vast amounts of information which require the development of efficient computational methods to analyze them. The important first step is to extract the fundamental patterns of gene expression inherent in the data. This paper describes the application of a novel clustering algorithm, super-paramagnetic clustering (SPC) to analysis of gene expression profiles that were generated recently during a study of the yeast cell cycle. SPC was used to organize genes into biologically relevant clusters that are suggestive for their co-regulation. Some of the advantages of SPC are its robustness against noise and initialization, a clear signature of cluster formation and splitting, and an unsupervised self-organized determination of the number of clusters at each resolution. Our analysis revealed interesting correlated behavior of several groups of genes which has not been previously identified. 

*PACS:* 02.50.−r; 87.16.−b; 87.17.−d; 87.80.−y; 89.70.+c

## 1. Introduction

DNA microarray technologies have made it straightforward to monitor simultaneously the expression levels of thousands of genes during various cellular processes [1,2]. The new challenge is to make sense of such massive expression data [3]. In most of the experiments, investigators compare the relative change of gene expression levels between two samples (one is called the target, such as a disease sample; the other is called the control, such as a normal sample). In a typical experiment simultaneous expression levels of thousands of genes are viewed over a few tens of time-points (or different tissues [4]). Hence one needs to analyse arrays that contain $10^5$–$10^6$ measurements.

* Corresponding author. Fax: +1-516 367-8461.
*E-mail address:* mzhang@cshl.org (M.Q. Zhang)

0378-4371/00/$ - see front matter 
PII: S0378-4371(99)00524-5

The aims of such analysis are typically to (a) group genes with correlated expression profiles; (b) focus on those groups which seem to participate in some biological process; (c) provide a biological interpretation of the clusters. Interpretations could be co-regulation of the mean cluster expression with a known process, a promoter common to most of the genes in the cluster, etc. (d) in experiments that compare data from different tissues (such as tumor and normal [4]) one also tries to differentiate them on the basis of their genetic expression profiles.

The sizes of the data sets and their complexity call for multi-variant clustering techniques which are essential for extracting correlated patterns in the swarm of data points in multidimensional space (for example, each relative gene expression profile with $k$ time-points may be regarded as a $k$-dimensional vector).

## 2. SPC

Currently, two clustering appoaches are very popular among biologists. One is average linkage, a hierarchical clustering method [5], with the Pearson correlation used as a similarity measure [6]. The other is self-organizing maps (SOMs) [7], whose most popular implementation for array data analysis is GENECLUSTER [8].

We present here clustering performed by SPC, a hierarchical clustering method recently introduced by Blatt et al. [9]. It is based on an analogy to the physics of inhomogeneous ferromagnets. Full details of the algorithm [10] and the underlying philosophy [11] are given elsewhere; here only a brief description is provided.

The input required for SPC is a distance matrix between the $N$ data points that are to be clustered. From such a distance matrix one constructs a graph, whose vertices are the data points and edges identify neighboring points. Two points $i$ and $j$ are called neighbors (and connected by an edge) if they satisfy the *K-mutual-neighbor* criterion, i.e., iff $j$ is one of the $K$ nearest points to $i$ and vice versus. With each edge we associate a weight $J_{ij} > 0$, which decreases as the distance between points $i$ and $j$ increases.

Assignement of the datapoints to clusters is equivalent to partitioning this weighted graph. Cluster indices play the role of the states of Potts spins assigned to each vertex (i.e., to each original data point). Two neighboring spins are interacting ferromagnetically with strength $J_{ij}$. This Potts ferromagnet is simulated at a sequence of temperatures $T$. The susceptibility and the correlation function for neighboring pairs of spins are measured. The pair correlation function serves to identify clusters: high correlation means that the two data points belong to the same cluster.

The temperature $T$ controls the resolution at which clustering is performed; the algorithm finds typical clusters at all resolutions. At very low temperatures all points belong to a single cluster and as $T$ is increased, clusters break into smaller ones until at high enough temperatures each point forms its own cluster. The clusters found at all temperatures form a dendrogram. Blatt et al. showed that the SPC algorithm is robust since the clusters are formed due to collective behavior of the system. The major

splits can be easily identified by a peak in the susceptibility. For more details see Refs. [9–11].

## 3. Yeast cell cycle and microarray data

We applied SPC on a recently published data set [14] to determine whether it could automatically expose known clusters without using prior knowledge. Eisen et al. [6] clustered the genes on the basis of data *combined* from seven different experiments. We suspected that mixing the results of different experiments may introduce noise into the data associated with a single one. Therefore we chose to use only a single time course, that of gene expression as measured in a single process (cell division cycles following alpha-factor block-and-release [12]). Furthermore, we focused on genes that have characterized functions (2467 genes) for easier interpretation.

Genetic controls and regulation play a central role in determination of cell fate during development. They are also important for the timing of cell cycle events such as DNA replication, mitosis and cytokinesis. Yeast is a single cellular organism, which has become a favorite model in molecular biology due to the easiness of genetic and biochemical manipulation and the availability of the complete genome. Like all living cells, the yeast cell cycle consists of four phases: $G1 \rightarrow S \rightarrow G2 \rightarrow M \rightarrow G1\ldots$, where S is the phase of DNA synthesis (replicating the genome); M stands for mitosis (division into two daughter cells), and the two gap phases are called G1 (preceding the S phase) and G2 (following the S phase). At least four different classes of cell cycle regulated genes exist in yeast [13]: G1 cyclins and DNA synthesis genes are expressed in late G1; histone genes in S; genes for transcription factors, cell cycle regulators and replication initiation proteins in G2; and genes needed for cell separation are expressed as cells enter G1. Early and late G1-specific transcription is mediated by the Swi5/Ace2 and Swi4/Swi6 classes of factors, respectively. Changes in the master cyclin/Cdc28 kinases are involved in all classes of regulation.

In the alpha-factor block-release experiments, MATa cells were first arrested in G1 by using alpha pheromone. Then the blocker was removed; from this point on the cell division cycle starts and the population progresses with significant cell cycle synchrony. Messenger RNA was extracted from the synchronized sample, as well as from a control sample (asynchronous cultures of the same cells growing exponentially at the same temperature in the same medium).

Fluorescently labeled cDNA was synthesized using Cy3 (“green”) for the control and Cy5 (“red”) for the target. Mixtures of equal amounts of the two samples were taken at every 7 min and competitively hybridized to individual microarrays containing essentially all yeast genes. The ratio of red to green light intensity (proportional to the ratio of RNA concentrations) was measured by scanning laser microscopy (see Ref. [12] for experimental details). The actual data provided at the Stanford website [14] is the log ratios.

In their analysis, Spellman et al. were focusing on identification of 800 cell cycle regulated genes (that may have periodic expression profiles). In our test of SPC, in addition to oscillatory genes we were also looking for any groups of genes with highly correlated expression patterns.

## 4. SPC analysis of yeast gene expression profiles

We clustered the expression profiles of the 2467 yeast genes of known function over data taken at 18 time intervals (of 7 min) during two cell division cycles, synchronised by alpha arrest and release. Denote by $E_{ij}$ the relative expression of gene $i$ at time interval $j$. Our data consist of 2467 points in an 18-dimensional space, normalised in the standard way:

$$G_{ij} = \frac{E_{ij} - \langle E_i \rangle}{\sigma_i}, \quad \langle E_i \rangle = \frac{1}{18}\sum_{j=1}^{18} E_{ij}, \quad \sigma_i^2 = \frac{1}{18}\sum_{j=1}^{18} E_{ij}^2 - \langle E_i \rangle^2 .$$

We looked for clusters of genes with correlated expression profiles over the two division cycles. The SPC algorithm was used with $q=20$ component Potts spins, each interacting with those neighbors that satisfy the $K$-mutual neighbor criterion [10] with $K = 10$. Euclidean distance between the normalized vectors was used as the distance between two genes. This distance is proportional to the Pearson correlation used by Eisen et al.

At $T = 0$ all datapoints form one cluster, which splits as the system is "heated". The resulting dendrogram of genes is presented in Fig. 1. Each node represents a cluster; only clusters of size larger than 6 genes are shown. The last such clusters of each branch, as well as non-terminal clusters that were selected for presentation and analysis (in a way described below) are shown as boxes. The circled boxes represent the clusters that are analysed below.

The position of every node along the horizontal axis is determined for the corresponding cluster according to a method introduced by Alon et al. [4]; proximity of two clusters along this axis indicates that the corresponding temporal expression profiles are not very different. The vertical axis represents the resolution, controlled by the "temperature" $T \geqslant 0$. The vertical position of a node or box is determined by the value of $T$ at which it splits. A high vertical position indicates that the cluster is stable, i.e., contains a fair number of closely spaced data points (genes with similar expression profiles).

In order to identify clusters of genes whose temporal variation is on the scale of the cell cycle, we calculated *for each cluster* a cycle score $S_1$, defined as follows. First, for each cluster $C$ (with $N_C$ genes) we calculate the average normalized expression level at all $j = 1, \ldots, 18$ time intervals and the corresponding standard deviations $\sigma^C(j)$

$$\bar{G}^C(j) = \frac{1}{N_C}\sum_{i\in C} G_{ij} \quad [\sigma^C(j)]^2 = \frac{1}{N_C}\sum_{i\in C}(G_{ij})^2 - [\bar{G}^C(j)]^2 .$$

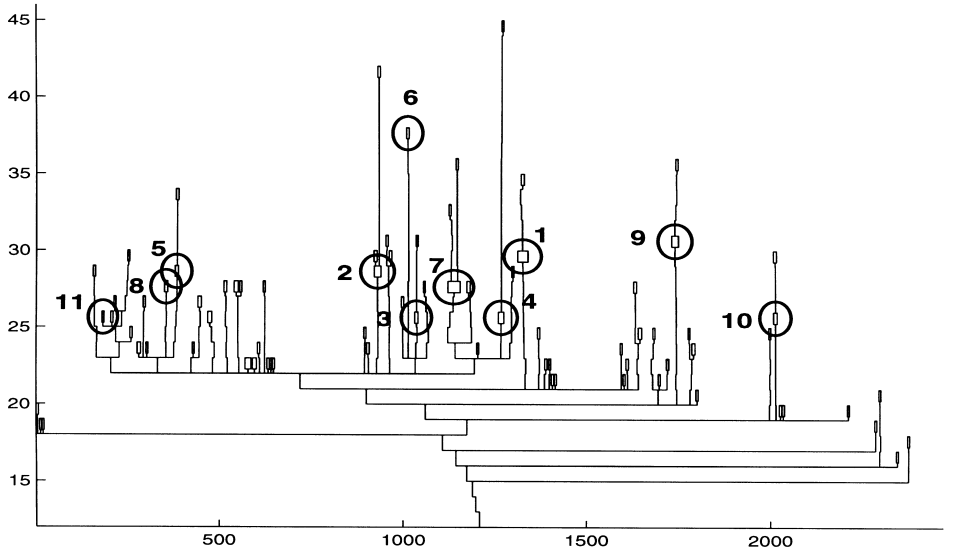



Fig. 1. Dendrogram of genes. Clusters 1–4 were selected according to our criteria, Eqs. (1) and (2). The other circled and numbered clusters are also interesting (see text).

Next, we evaluated the Fourier transform of the mean expression profiles $\bar{G}^C(j)$ for every gene cluster $C$. To suppress initial transients, the Fourier transform is performed only over $j = 4,\ldots,18$. Denote the absolute values of the Fourier coefficients by $A_k$; the ratio between low-frequency coefficients and the high-frequency ones was used as a figure of merit for the time scale of the variation. We observed that clusters that satisfy the condition

$$S_1^C = \frac{\sum_{k=2}^{4} A_k}{\sum_{k=6}^{8} A_k} > 2.15\,, \tag{1}$$

have the desired time dependence, and found 29 clusters (consisting of 167 genes) to have such scores. For many of these clusters, however, the temporal variation was very weak, i.e., of the same order as the standard deviations $\sigma^C(j)$ of the individual gene expressions of the cluster. We defined another score, $S_2^C$, for which we required

$$S_2^C = \frac{1}{18}\sum_{j=1}^{18}\left[\frac{\bar{G}^C(j)}{\sigma^C(j)}\right]^2 > 5.6\,. \tag{2}$$

For clusters $C$ that satisfy this condition the "signal" significantly exceeds the noise. We select a cluster if its score exceeds 5.6, while its parent's score does not. Only four clusters, containing 86 genes, satisfy both conditions (1) and (2); these are numbered 1–4 on Fig. 1. Seven additional relatively stable clusters which did *not* satisfy our two criteria, but are of interest, are also selected and circled in Fig. 1.

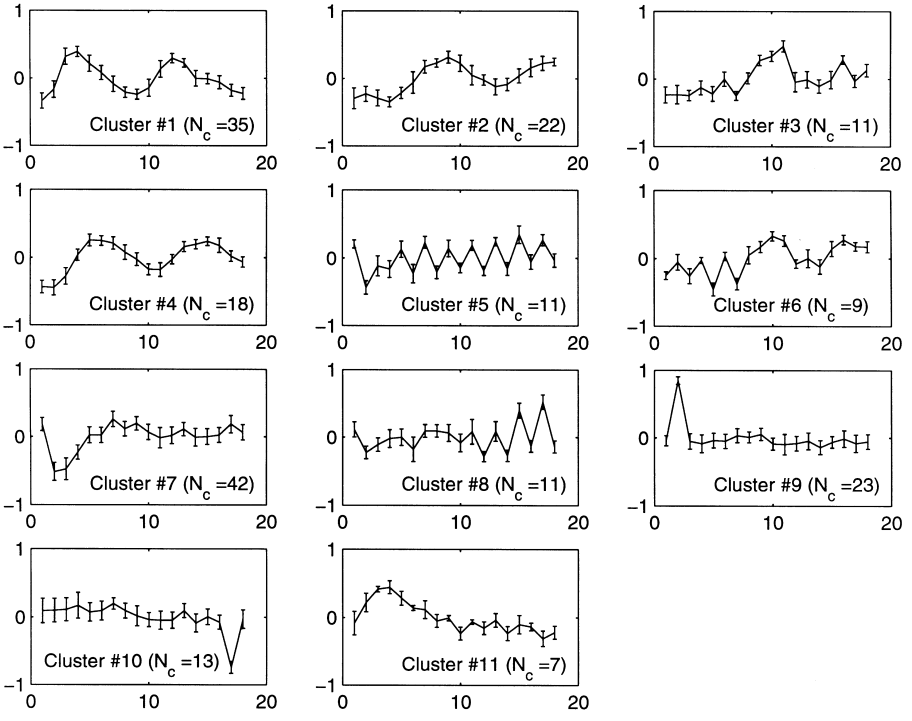



Fig. 2. Mean normalized expression of selected clusters, versus time, measured at intervals of 7 min. Error bars represent the standard deviations $\sigma^C(j)$. $N_c$ is the number of genes in each cluster. The clusters are numbered as in Fig. 1.

The corresponding time sequences are shown in Fig. 2: $\tilde{G}^C(j)$ is plotted for each cluster versus time $j$, with the error bars representing the standard deviations $\sigma^C(j)$. Clusters 1,2 and 4 clearly correspond to the cell cycle.

## 5. Details and interpretation of gene clustering

The full lists of genes that constitute the 11 selected clusters are given in our website [15]. We present here a short analysis of our clusters. We use standard notation for bases: R stands for A or G, W for A or T, K for G or T, N for any base.

*Cluster* # 1: These are mostly Late G1 phase-specific genes. They contain the major cell cycle regulators: Cln1,2, Clb5,6 and Swi4 as well as DNA replication and repair genes. One can easily detect MCB (ACGCGT) or SCB (CRCGAAA) sites in their promoters to which MBF (Swi6p + Mbp1p) and SBF (Swi6p + Swi4p) bind, respectively [13].

*Cluster* # 4: This cluster contains mostly S phase genes and is dominated by the histones. Histones are required for wrapping up nascent DNA into nucleosomes, their

promoters are regulated by CCA (GCGAARYTNGRGAACR), NEG (CATTGNGCG) as well as SCB (CGCGAAA) [3].

*Cluster* # 2: These are mostly G2/M phase genes. They contain the major cell cycle regulators: Clb1,2 and Swi5/Ace2. It is known that all genes co-regulated with Clb1,2 are mainly controlled by either Mcm1 at P-box (TTWCCYAAWNNGGWAA) or by Mcm1 + SFF through the composite site: (P-box)N2-4RTAAAYAA [12,3].

*Clusters* #5, #6 *and* #8: These are mostly ribosomal protein (RP) genes. The genome of Saccharomyces cerevisiae contains 137 genes coding for ribosomal proteins [16]. Since 59 genes are duplicated, the ribosomal gene family encodes 78 different proteins, 32 of the small and 46 of the large ribosomal subunit. They are co-regulated because they are sub-components of ribosome machinery for protein translation. All genes in cluster #6 reside on chromosomes 2, 4 and 5, except rpl11b which resides on chromosome 9. All genes in clusters #5 and #8 (which are very close in the dendrogram of Fig. 1) reside on chromosomes 8–16, except rps17b which resides on chromosome 4. It is likely that the expression of these ribosomal genes is correlated to their chromosomal locations. It is interesting that the expression profiles appear to have pronounced oscillations (throughout in #5, at early times in #6 and late times in #8). Like most of the RP genes, the ribosomal genes in the three clusters also contain multiple global Regulator Rap1p binding sites in their promoters within a preferred window of 15–26 bp [17]. The transcription of most RP genes is activated by two Rap1p binding sites, 250–400 bp upstream from the initiation of transcription. Since Rap1p can be both an activator and a silencer, it is not known whether Rap1p is responsible for the oscillation. This oscillation could be a result of interplay between cell cycle and Rap1p activity which determines the mean half life of the RP mRNAs (5–7 min, [18]). As fresh medium was added at 91 min during the alpha-factor experiments, the genes in #6 and in #8 may reflect different responses to the nutrient change.

*Cluster* #7: This cluster has 42 genes that are largely not cell cycle regulated. These genes have diverse functions in general metabolism. When searching promoter regions for regulatory elements using gibbsDNA [20], a highly conserved motif GCGATGAGNT is shared by 90 % of genes. This element seems to be novel, it has some similarity to Gcn4p site TGACTC and Yap1p site GCTGACTAATT [19]. When searching the yeast promoter database – SCPD [21], we found that the BUF site in the HO gene promoter and the UASPHR site in the Rad50 promoter appear to contain the core motif GATGAG. Although we do not know if this element is functional or what might be the trans-factor, it is still very likely that it may contribute the co-regulation of this cluster of genes.

*Cluster* #10: This cluster is characterized by a pronounced dip towards the end of the profile. They are not cell cycle regulated by and large, except Clb4 (a S/G2 cyclin) and Rad54 (a G1 DNA repair gene). By searching promoter elements, we found a conserved motif RNNGCWGCNNC that is shared by a subset of the genes (Clb4, YNL252C, Rad54, Rpb10, Atp11 and Pex13). It partially matches a PHO4 binding motif (TCGGGCCACGTGCAGCGAT) in the promoter of Pho8. However, the PHO4 consensus, CACGTK, does not appear in the conserved motif of our

cluster. Therefore, we suspect that it is a novel motif which should be tested by experiments.

## 6. Summary

We used the SPC algorithm to cluster gene expression data for the yeast genome. We were able to identify groups of genes with highly correlated temporal variation. Three of the groups found clearly correspond to well known phases of the cell cycle; some of our observations of other clusters reveal features that have not been identified previously and may serve as the basis of future experimental investigations.

## Acknowledgements

Research of E. Domany was partially supported by the Germany-Israel Science Foundation (GIF) and the Minerva foundation. Research of M.Q. Zhang was partially supported by NIH/NHGRI under the grant number HG01696.

## References

[1] D.J. Lockhart, H. Dong, M.C. Byrne et al., Nature Biotech. 14 (1996) 1675–1680.
[2] J. De Risi, V. Iyer, P.O. Brown, Science 278 (1997) 680–686.
[3] M.Q. Zhang, Genome Res. 9 (1999) 681–688.
[4] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Proc. Natl. Acad. Sci. 96 (1999) 6745–6750.
[5] J. Hartigan, Clustering Algorithms, Wiley, New York, 1975.
[6] M. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.
[7] T. Kohonen, Self-Organizing Maps, Springer, Berlin, 1997.
[8] P. Tamayo, D. Slonim, J. Mesirov et al., Proc. Natl. Acad. Sci. USA 96 (1999) 2907–2912.
[9] M. Blatt, S. Wiseman, E. Domany, Super-paramagnetic clustering of data, Phys. Rev. Lett. 76 (1996a) 3251–3255.
[10] M. Blatt, S. Wiseman, E. Domany, Neural Comput. 9 (1997) 1805.
[11] E. Domany, Physica A 263 (1999) 158.
[12] P.T. Spellman, G. Sherlock, M.Q. Zhang et al., Mol. Biol. Cell. 9 (1998) 3273–3297.
[13] C. Koch, K. Nasmyth, Curr. Biol. 6 (1994) 451–459.
[14] The data can be obtained from http://cellcycle-www.standford.edu.
[15] http://www.weizmann.ac.il/physics/complex/clustering/.
[16] W.H. Mager, R.J. Planta, J.G. Ballesta et al., Nucl. Acid. Res. 25 (1997) 4872–4875.
[17] R.F. Lascaris, W.H. Mager, R.J. Planta, Bioinformatics 15 (1999) 267–277.
[18] B. Li, C.R. Nierras, J.R. Warner, Mol. Cell. Biol. 19 (1999) 5393–5404.
[19] A.G. Hinnebusch, The Molecular and Cellular Biology of the Yeast Sacchromyces: Gene Expression, Vol. 2, Cold Spring Harbor Press, New York, 1992, p. 319.
[20] M.Q. Zhang, Comput. Chem. 23 (1999a) 233–250.
[21] J. Zhu, M.Q. Zhang, Bioinformatics 15 (1999) 607–611.

# Publication 6:

## Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity

Authors: F. Quintana, G. Getz, G. Hed, E. Domany and I. R. Cohen
Published in: *Journal of Autoimmunity* **21**, 65–75 (2003).

Journal of Autoimmunity 21 (2003) 65–75

journal of autoimmunity

www.elsevier.com/locate/issn/08968411

# Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity

Francisco J. Quintana[a], Gad Getz[b], Guy Hed[b], Eytan Domany[b], Irun R. Cohen[a*]

[a] *Department of Immunology, The Weizmann Institute of Science, Rehovot 76100, Israel*
[b] *Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot 76100, Israel*

Received 30 October 2002; revised 11 March 2003; accepted 26 March 2003

**Abstract**

Informatic methodologies are being applied successfully to analyze the complexity of the genome. But beyond the genome, the immune system reflects the state of the body in health and disease. Traditionally, immunologists have reduced the immune system, where possible, to one-to-one relationships between particular antigens and particular antibodies or T-cell clones. Autoimmune diseases, caused by an immune attack against a body component, are usually investigated by following the response to single self-antigens. In this study, we apply informatics to analyze *patterns* of autoantibodies rather than single species of autoantibodies. This study was designed not to replace traditional approaches to immune diagnosis, but to test whether meaningful patterns of autoantibodies might exist. Using an unbiased solid-phase ELISA antibody test, we detected serum IgG and IgM antibodies in the sera of 20 healthy persons and 20 persons with type 1 diabetes mellitus binding to an array of 87 different antigens, mostly self-antigens. The healthy subjects manifested autoantibodies to a variety of self-antigens, many known to be associated with autoimmune diseases. We investigated the patterns of these autoantibodies using a coupled two-way clustering algorithm developed for analyzing data from gene arrays. We now report that the reactivity patterns of autoantibodies to particular subsets of self-antigens exhibited non-trivial structure, which significantly discriminated between healthy persons and persons with type 1 diabetes. The results show that despite the wide prevalence of autoantibodies, the patterns of reactivity to defined subsets of self-antigens can provide information about the state of the body.


*Keywords:* Bio-informatics; Autoantibodies; Immunological homunculus

## 1. Introduction

The great advances in information about the molecules and cells comprising the immune system have frustrated immunologists; the immune system is patently more complex than originally thought. Autoimmunity is a notable example of the problem. Traditionally, investigators and clinicians have focused on selected autoantibodies to study or diagnose specific autoimmune diseases [1–4]. They sought to establish a one-to-one relationship between a particular autoantibody and a particular disease. In practice, however, the presence of autoantibodies in healthy persons [5] complicates the serological diagnosis of autoimmune disease and confounds our understanding as to how the immune system actually discriminates the self from the non-self [6]. Immunology is in need of informatics.

For their part, complexity science people have not given much thought to the immune system, and have focused mostly on genomics or the nervous system. The immune system, however, is a suitable subject for informatics: like the central nervous system, the immune system is self-organizing [6–8]; unlike the central nervous system, the immune system is functionally accessible as a system at the cellular level both in vivo and in vitro [9].

The present study applies informatics to autoimmunity: we characterize a set of molecules recognized by

* Corresponding author. Tel.: +972-8-934-2911; fax: +972-8-934-4103.
*E-mail address:* irun.cohen@weizmann.ac.il (I.R. Cohen).

0896-8411/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0896-8411(03)00064-7

autoantibodies in healthy persons and test whether the global patterns of autoantibodies might discriminate between a state of health and an autoimmune disease, such as type 1 diabetes mellitus. This concept has been already tested in the past for other human autoimmune conditions using complex mixtures of undefined self-antigens [10–13]. However, our aim was to use an ELISA system able to detect even small amounts of low-affinity autoantibodies binding to an array of 87 different antigens of known identity without preconceived bias. We did not use reactivity thresholds, as is usually done to define a 'negligible background', and we did not restrict the detection to the specific high-affinity antibodies usually associated with autoimmune disease. For the purpose of this analysis, we define autoantibodies as antibodies that bind to self-molecules in the ELISA conditions used in this study. We imply nothing about function. To analyze the patterns of the autoantibodies, we applied a clustering algorithm and tested the statistical significance of the results. Particular sets of self-antigens, most of which are not known to be associated with type 1 diabetes, were found to discriminate between the patterns of autoantibodies of the healthy subjects and those of the type 1 diabetes patients. These results demonstrate that even the low-affinity autoantibody repertoire is structured and can yield information about the state of the body [14] when analyzed with suitable informatic tools.

## 2. Materials and methods

### 2.1. Antigens

The 87 antigens used in these studies are enumerated in Table 1. These antigens include proteins, peptides, nucleotides and phospholipids reported to interact with antibodies. The antigens are classified according to their cellular localization, tissue distribution or function.

### 2.2. Antibodies

The second antibodies used in the ELISA assay were F(ab′)2 goat anti-human IgG+IgM linked to alkaline phosphatase and goat anti-human IgM linked to horseradish peroxidase. These antibodies were purchased from Jackson ImmunoResearch Laboratories Inc. (West Grove, PA, USA), and were used at a final dilution of 1:1500 in bovine serum albumin 0.3%.

### 2.3. Test samples

Serum samples were collected at the Hadassah Medical Center (Jerusalem, Israel), under the supervision of Dr Rivka Abulafia-Lapid and Professor Itamar Raz, from 20 healthy young-adult blood donors, with no family history of diabetes, and from 20 unselected type 1 diabetes patients. Most of the type 1 diabetes patients, too, were young adults, 21–34 years old (95% confidence interval; median, 23 years). The diagnosis was made on the basis of accepted clinical criteria: hyperglycemia, ketonuria, low body weight, the absence of a family history of type 2 diabetes and a standard (anti-glutamic acid decarboxylase (GAD)) antibody assay [3]. The HLA genotypes were not tested. The sera were collected within 4–7 weeks of diagnosis (95% confidence interval; median, 6 weeks). Informed consent was obtained. The samples were stored at −20 °C without any additive. The T-cell proliferative responses of these type 1 diabetes patients and healthy blood donors had been studied previously. No significant difference was found between the groups in their T-cell responses to the foreign antigen tetanus toxoid, although the diabetic subjects manifested heightened responses to the 60 kDa heat shock protein (HSP) self-antigen [15].

### 2.4. Solid-phase antibody assay

A standard ELISA assay was used. Antigens (10 μg/ml in phosphate-buffered saline (PBS)) were coated in 96-well ELISA plates (Maxisorp; Nunc, Roskilde, Denmark) by overnight incubation at 4 °C. The plates were washed with PBS 0.05% Tween, and blocked for 2 h with bovine serum albumin 3% (Sigma, Rehovot, Israel). The serum samples were diluted 1:100 in bovine serum albumin 0.3%, and 50 μl was added to each well. After 3 h of incubation at 37 °C, the sera were removed and the plates were washed with PBS 0.05% Tween. Bound antibodies were detected with an appropriate alkaline phosphatase or horseradish peroxidase-conjugated second antibody (Jackson ImmunoResearch Laboratories Inc.), 50 μl incubated for 1.5 h at 37 °C. The plates were washed with PBS, and *p*-nitrophenol phosphate or 2,2′-azino-bis (3-ethylbenzthiazoline-6 sulfonic acid; both from Sigma) were added, and the optical density (OD) in each well was read at 405 nm using a spectrophotometer.

We optimized the conditions of the assay: a direct correlation between the OD readings and dilutions of the sera were found between 1:50 and 1:200 dilutions. Accordingly, we chose 1:100 as the standard dilution of the test sera. The relationship between the OD and the incubation time was linear during 45 min of incubation (mean $r^2 \pm$ standard deviation (SD)=$0.98 \pm 0.02$). Therefore, we recorded the OD readings 30 min after the addition of the substrate. The assay was reproducible: the mean intra-assay coefficient of variation was 4.3%, and the mean inter-assay coefficient of variation was 9.5%. Correlation analysis of intra- and inter-assay variations yielded $r^2$ coefficient values 0.98 and 0.96, respectively, with $P$ value <0.0001 for both.

### 2.5. Cluster analysis

The OD readings corresponding to the antibody reactivities of a group of $N$ serum samples against a panel of $M=176$ different reactivities (87 antigens and one blank with two secondary antibodies) were placed in a matrix $A$, whose element $A_{js}$ represents the extent to which the serum of subject $s$ reacted with test antigen $j$ (the secondary antibody was absorbed in the index $j$). The 'immune state' of subject $s$ is represented by a vector $\mathbf{A}^{(s)}$ (of $M$ components). Similarly, antigen $j$ is represented by the ($N$ component) vector $\mathbf{A}^{(j)}$.

The following normalization was used and done only once, for $k=1,...,88$ (IgM) and $k=89,...,176$ (IgM+IgG):

$$B_{js}=\log A_{js}-\frac{2}{M}\sum_k \log A_{ks} \quad (1)$$

For $j=1,2,...,88$, the sum over $k$ is from 1 to 88, and for $j=89,90,...,176$, the sum is from 89 to 176. For every subject $s$, this operation produces a mean-centered set of values; if all readings $A_{js}$ are multiplied by a constant, it does not affect the $B$ variables. We take the log, since the noise on the readings is multiplicative, and we mean center the variables to eliminate dependence on concentration fluctuations from subject to subject.

Next, we renormalized the data by subtracting from each element the average value of the elements in the same matrix row (corresponding to a particular antigen) and dividing by the SD of the row. The elements of the resulting renormalized submatrix are denoted by $G_{js}$—for each antigen, the mean of $G_{js}$ vanishes and the sum of squares is 1. We analyzed the data using the method of Getz et al. [16] for analysis of gene expression. Thus, we identify subsets of $K$ serum samples and cluster them on the basis of their reactivities to a selected subset of antigens. In this way, the analysis uses various submatrices of the total data matrix $G$, described in detail in Ref. [16]. Briefly, we first cluster all antigens (using data from all sera) and identify stable antigen clusters. Next, we cluster the sera using, one at a time, the groups of antigens that emerged as stable clusters in the first step.

To assign related antigens to the same cluster, we singled out 'close' pairs of highly correlated antigens, as well as pairs that were highly anti-correlated. This 'closeness' is measured by the 'distance' $d_{j,l}$ between antigens $j$ and $l$, given by

$$(d_{j,l})^2=1-c_{j,l}^2 \quad (2)$$

where $c_{j,l}$ is the correlation coefficient of antigens $j$ and $l$, as measured over the $N$ samples, given by

$$c_{j,l}=\sum_{s=1}^{N} G_{js}G_{ls} \quad (3)$$

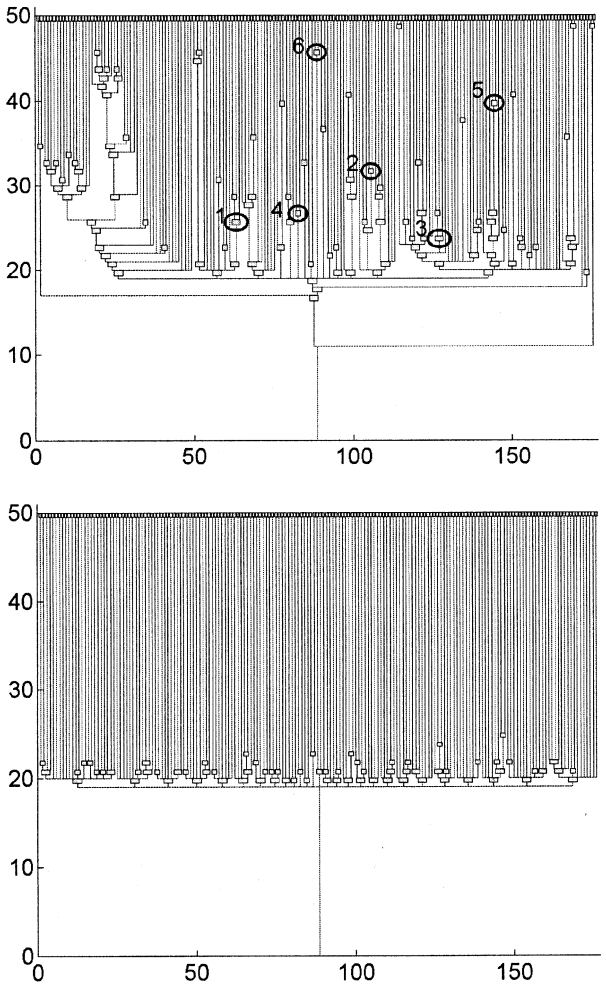



Fig. 1. Dendrograms of antigens obtained by clustering. (A) Dendrogram obtained from the original data matrix, using sera from healthy and type 1 diabetes subjects; the antigen clusters that are reported in Table 3 are circled and numbered. (B) Dendrogram of the antigens obtained by clustering a randomized matrix.

In contrast, the distance $D_{sp}$ between subjects $s$ and $p$ is the Euclidian distance

$$D_{sp}^2=\sum_j (G_{js}-G_{jp})^2 \quad (4)$$

These distance measures reflect similarity between pairs of subjects and pairs of antigens. Since the clustering method in this study relies heavily on proximity of pairs of points rather than wide separations, the results are only slightly sensitive to the precise measure used in Eq. (4).

We used an unsupervised clustering technique, the SPC clustering algorithm [17], which organizes the data in the form of a dendrogram, such as that shown in Fig. 1A, B. As a control parameter $T$ increases to a value $T_1(C)$, a cluster $C$ may be 'born' (when its 'parent' cluster breaks up into two or more subclusters, one of which is $C$). As $T$ increases further, to $T_2(C)>T_1(C)$, $C$

itself breaks up and 'dies'. A main advantage of SPC is that it provides a quantitative stability index, $R(C)=T_2(C)/T_1(C)$ for any cluster $C$. The larger the value of $R(C)$, the more statistically significant and stable (against noise in the data and fluctuations) is $C$. We used SPC within a coupled two-way clustering approach [16] to identify subsets of serum samples and of antigens, allowing meaningful partitions of the samples to emerge. The clinical labels were then used to *evaluate* the results (not to produce them). If a cluster of serum samples contained predominantly subjects with the same diagnosis, the cluster's predictive capacity was estimated. The effectiveness of the resulting classification was measured in terms of the *number of classification errors*, $N_e$ that were made by assigning the subjects of the cluster to one diagnosis and the rest of the subjects to the other diagnosis, healthy or type 1 diabetes. We evaluated *specificity* and *sensitivity*: *specificity* is the fraction of correctly diagnosed subjects present in a cluster (=1 for no false positives); *sensitivity* is the fraction of correctly diagnosed subjects that were included in the cluster, of the total number of subjects with the same diagnosis (=1 for no false negatives).

### 2.6. Combining classifiers

The sensitivity and specificity of 'final' classifier in this study were improved by combining several different sets of antigens. We classified a test sample as type 1 diabetes, for example, if it was so discriminated by a majority of the classifiers. Then we identified the samples, and evaluated the specificity and sensitivity of the combined classifiers.

### 2.7. Assessing statistical significance

Statistical analysis was carried out to test four questions: *first*, whether the pattern of antibody reactivity exhibited a non-trivial structure; *second*, how sensitive were the subject clusters to variations in the data—such as leaving out one subject; *third*, whether the clinical state of the subjects was reflected by their reactivity patterns; and *fourth*, how robust was the method, whether it is able to predict the clinical status of subjects with unknown clinical labels.

For the *first* question, we used the original data matrix and randomized all its entries, placing each in a random location. The randomized matrix was normalized and clustered; for every cluster $C$, the size (number of elements) $n(C)$ and its stability index $R(C)$ were recorded. For each $n$, we identified $R^*(n)$, the maximal value of $R$. This was repeated for *1000* random matrices and; for each size $n$, a histogram $P_n(R^*(n))$ was prepared, and the SD $\sigma_n$ and maximal value $R^*_{\max}(n)$ were determined. The maximal values of $R$, which were found for the stable clusters of size $n$ obtained from the real data, were compared with the extremal values found for the randomized matrices.

For the *second* question, we repeated the clustering of the subjects 40 times, leaving out one subject in each trial, and checked the effect on the resulting clusters.

For the *third* question, the $P$-values of the diagnostic labels were estimated by calculating the probability, $P$, that a previously selected 'discriminating' cluster of $C$ subjects would produce $e$ errors for randomly assigned clinical labels, with $e \leq N_e$, where $N_e$ is the number of 'errors' for the real labels (of $S$ diagnosed subjects). A high value of $P$ indicates that the discrimination produced by this cluster is not related to the diagnosis in a statistically significant way. This probability $P(e \leq N|C,S,N)$ was determined by the Fisher's exact two-sided test [18].

For the *fourth* question, we simulated a situation in which the clinical labels of 30 subjects (15 diseased, 15 healthy) were known and the diagnosis of the remaining 10 subjects was hidden. All 40 subjects were clustered as described previously, using, one at a time, the reactivities of each one of the six antigen clusters shown in Table 3. In the resulting dendrogram of subjects (obtained for each one of the six antigen clusters), we identified the stable clusters of subjects. These clusters were candidates to serve as classifiers; among them, we selected the one with the lowest number of errors based on the 30 '*known*' labels. This subject cluster was then chosen as a classifier. Next, the 10 'unknown' samples were diagnosed on the basis of their affiliation with the classifier cluster. This process was repeated for each one of the six antigen clusters enumerated in Table 3, using 100 random choices of 10 subjects with hidden labels. In this way, we could determine the ability of classifiers, that were constructed using 30 labeled subjects, to correctly identify 10 'unknowns'.

### 2.8. Principal component analysis

For comparison, the data were also analyzed by standard principal component analysis (PCA) [19].

## 3. Results

### 3.1. Serum autoantibodies in healthy blood donors

We tested the repertoire of autoantibodies in the sera of 20 healthy individuals using an array of 87 different antigens (listed in Table 1) and detection antibodies directed to IgM or IgG+IgM. Table 2 summarizes the self-antigens most frequently recognized by the autoantibodies of the 20 healthy blood donors. For comparison, Table 2 also contains the bacterial antigens lipopolysaccharide (LPS) and purified protein derivative (PPD) of *Mycobacterium tuberculosis*. Thus, healthy

persons may express a wide range of autoantibodies detectable by ELISA. To learn whether the patterns of such autoantibodies might be informative, we applied our clustering analysis to the healthy subjects and to a population of persons with the autoimmune disease type 1 diabetes mellitus.

### 3.2. Structure in the repertoire: self-antigen clusters

First, we clustered the antigens and then used the various antigen clusters as probes to cluster the subjects. To cluster the 87 antigens, we used the serum OD readings for all subjects (healthy or not). The serum reactivity data (for 2 isotypes × 87 antigens) exhibited a non-trivial structure, as is evident from comparing the dendrogram of Fig. 1A (obtained by clustering the original antigen matrix) with that of Fig. 1B (obtained by clustering a randomized matrix). The dendrograms obtained from the randomized matrices exhibited a sudden 'melt down' from a single cluster that contained all the points to small clusters with very low stability. The dendrogram of the actual data, in contrast, contained a cluster of five antigens, with stability index $R$=1.33. We tested 1000 different realizations of randomized matrices, and determined for each cluster size $n$ the corresponding extremal value $R^*_{\max}(n)$ and SD $\sigma_n$. The real data yielded stable clusters of size $n$, whose stability indices $R$ exceeded the corresponding random extremal value $R^*_{\max}(n)$ by at least $3\sigma_n$ (data not shown). Thus, the $P$-value for the presence of non-trivial structure in the antigen clusters was less than 0.001. Hence, groups of self-antigens do cluster together as collectives [7]; sera that react with one member of an antigen cluster will tend to react with other members of the antigen cluster.

### 3.3. Clusters as classifiers of type 1 diabetes

The identified subsets of antigens were then used to probe the sera of 20 healthy donors and 20 diabetes patients. Fig. 2 shows the dendrogram obtained using the IgM reactivities to insulin and to collagen I and the IgG+IgM reactivities to collagen I. This set of antigens generated a sensitivity of 85% and a specificity of 81% for diabetes. Other sets of antigens generated different dendrograms. Table 3 summarizes the findings and includes the $P$-values calculated for each cluster size $C$ and error number $N_e$, using Fisher's exact two-tailed test. This result shows the advantages of our methodology versus linear discrimination. If one distributes randomly two labeled groups, of 20 points in each, in 176 dimensional space, a separating linear manifold will be found with probability very close to 1 [20,21]. However, as we have shown, our cluster analysis definitely does not separate two randomly placed groups of points. Finally, we repeated the clustering process 40 times using the reactivities of antigen cluster 1 of Table 3, each time leaving out another subject. The resulting clusters were stable despite the omissions; 17 subjects appeared in all 40 trials in the cluster of the diabetes patients; 16 were indeed diseased and one was healthy.

The combined results produced an overall sensitivity of 95% and a specificity of 90%. Among the self-antigens that discriminated between type 1 diabetes and healthy sera were cardiolipin, collagen I, collagen X, cytochrome $c$ P450, cartilage extract (a commercial preparation rich in collagen I), aldolase, acetylcholine receptor (AchR), heparin and insulin. Among this list of molecules, only insulin has been noted previously to be a self-antigen in type 1 diabetes [22]. Neither GAD nor the 60 kDa HSP appeared among the discriminatory antigen clusters, although both self-antigens have been implicated in type 1 diabetes in other assays [3,15].

Using the antigen clusters shown in Table 3, we chose classifier subject clusters (using 30 known labels) and tested the 10 'unknown' subjects, obtaining the following results: for antigen cluster 1, all 100 trials selected the same 'original' subject cluster as in Table 3. For antigen cluster 2, the original subject cluster was selected in 78 trials, but two other clusters were also picked up, both 11 times. For antigen cluster 3, we found the original cluster 73 times, but two other clusters were also picked up, 22 and five times. For antigen cluster 4, the scores were 95 and 5. For antigen clusters number 5 and 6, the score was 100 selections of the respective original subject clusters.

In each one of the 100 simulations, we diagnosed the 10 'unknown' samples on the basis of their affiliation with the selected classifier clusters, and for each simulation, used the majority rule described previously to combine the results obtained for the six antigen clusters. We found the following error distribution: 29 occurrences with 0 errors (for 10 predictions), 45 with 1 error, 21 with 2 errors and 5 with 3 errors. The average number of errors for the 10 'unknown' subjects was 1.02; this is only slightly worse than the 3 errors obtained for the 40 subjects, using all their labels.

The PCA [19] has been regularly used to study patterns of autoantibodies to undefined antigens in tissue blots [23,24]. Applying PCA to our data, we identified the eigendirections associated with the leading eigenvalues; each eigendirection had significant projections onto more than 10 antigens. Next, we projected the data onto the plane spanned by the two leading directions. The best linear separator, found using non-normalized data, generated a comparable number of errors to that of antigen cluster 1 in Table 3 (data not shown). The PCA finds about 20 antigens that have sizeable contributions to the two leading principal directions. The present clustering method, in contrast, yields separation in several two- or three-dimensional spaces that are related to a few antigens.

Table 1
Antigens used (the catalogue number is given for those molecules purchased from Sigma)

| Group | Function/structure | Number | Antigen | Sequence (when applicable) | Catalogue |
|---|---|---|---|---|---|
| Cellular structure | Cytoskeleton | 1 | Actin | | A3653 |
| | | 2 | Tubulin | | T4925 |
| | | 3 | Myosin | | M6643 |
| | | 4 | Tropomyosin | | T4770 |
| | | 5 | Vimentin | | V4383 |
| | Extracellular matrix | 6 | Fibronectin | | F0895 |
| | | 7 | Collagen I | | C7774 |
| | | 8 | Collagen II | | C7806 |
| | | 9 | Collagen III | | C4407 |
| | | 10 | Collagen IV | | C7521 |
| | | 11 | Collagen V | | C3657 |
| | | 12 | Heparin | | H2149 |
| | | 13 | Laminin | | L6274 |
| | | 14 | Collagenase | | C9891 |
| Cellular membranes | Phospholipids | 15 | Cardiolipin | | C5646 |
| | | 16 | Glucocerebroside | | G9884 |
| | | 17 | Phosphatidylethanolamine | | P9137 |
| | | 18 | Cholesterol | | C1145 |
| Cellular metabolism | Glucose | 19 | Enolase | | E0379 |
| | | 20 | Aldolase | | A8811 |
| | | 21 | Acid phosphatase | | P1774 |
| | Apoptosis | 22 | Annexin 33 kDa | | A9460 |
| | | 23 | Annexin 67 kDa | | A2824 |
| | | 24 | Cytochrome *c* P450 | | C3131 |
| | Monooxigenases | 25 | Catalase | | C9322 |
| | | 26 | Peroxidase | | P6782 |
| | | 27 | Tyrosinase | | T7755 |
| | Others | 28 | Ribonuclease | | R4875 |
| Nucleus | Protein | 29 | Histone II A | | H9250 |
| | DNA | 30 | Double-stranded DNA | | D1501 |
| | | 31 | Single-stranded DNA | | D1501 |
| Plasma proteins | Carriers | 32 | Transferrin | | T4132 |
| | | 33 | Fetuin | | F2379 |
| | | 34 | Human serum albumin | | A8763 |
| | | 35 | Bovine serum albumin | | A9647 |
| | | 36 | Ovoalbumin | | A5378 |
| | Coagulation | 37 | Factor II | | F5132 |
| | | 38 | Factor VII | | F6509 |
| | | 39 | Fibrin | | F5386 |
| | | 40 | Fibrinogen | | F4883 |
| | Complement | 41 | C 1 | | C2660 |
| | | 42 | C 1 q | | C0660 |
| Immune System | Cytokines | 43 | Interleukin 2 | | I2644 |
| | | 44 | Interleukin 10 | | I9276 |
| | | 45 | Interleukin 4 | | I4269 |
| | Immunoglobulins | 46 | IgG | | I8640 |
| | | 47 | IgM | | I8260 |
| | | 48 | 1E10 Fab[a] | | |
| | TCR peptides | 49 | N4 | ASSLWTNQDTQY | NA |
| | | 50 | C9 | ASSLGGNQDTQY | NA |
| Tissue antigens | Heat shock protein | 51 | HSP60[b] | | |
| | | 52 | p277 | VLGGGVALLRVIPALDSLTPANED | NA |
| | Islet antigens | 53 | GAD | | G2126 |
| | | 54 | Insulin | | I0259 |
| | CNS | 55 | Human MOG[c] | | |
| | | 56 | Murine MOG[c] | | |
| | | 57 | Human MOG p94–116[c] | GGFTCFFRDHSYQEEAAMELKVE | |
| | | 58 | Rat MOG p35–55[c] | MEVGWYRSPFSRVVHLYRNGK | |
| | | 59 | MBP[d] | | |
| | | 60 | Brain extract | | B1877 |
| | Muscle and skeleton | 61 | AchR[e] | | |
| | | 62 | Myoglobulin | | M6036 |

Table 1 (*continued*)

| Group | Function/structure | Number | Antigen | Sequence (when applicable) | Catalogue |
|---|---|---|---|---|---|
| Tissue antigens | Joints | 63 | Cartilage extract | | C5210 |
| | Thyroid | 64 | Thyroglobulin | | T1001 |
| | Blood cells and platelets | 65 | Hemoglobin A | | H0267 |
| | | 66 | Spectrin | | S3644 |
| Foreign antigens | Proteins and peptides | 67 | TB PPD[f] | | |
| | | 68 | HSP65[g] | | |
| | | 69 | ecp27 | KKARVEDALHATRAAVEEGV | NA |
| | | 70 | mtp278 | EGDEATGANIVKVALEA | NA |
| | | 71 | GST[b] | | |
| | | 72 | KLH[h] | | |
| | | 73 | Pepstatin | | P5318 |
| | | 74 | R13 | EEEDDDMGFGLFD | NA |
| | Others | 75 | LPS | | L3755 |
| Synthetic polymers | Poly amino acids | 76 | Poly arginine | | P3892 |
| | | 77 | Poly lysine | | P4408 |
| | | 78 | Poly aspartic | | P6762 |
| | | 79 | Poly glutamate | | P4636 |
| | Oligonucloetides | 80 | PolyA | $A_{20}$ | NA |
| | | 81 | PolyT | $T_{20}$ | NA |
| | | 82 | PolyC | $C_{20}$ | NA |
| | | 83 | PolyG | $G_{20}$ | NA |
| | | 84 | PolyATA | $AT_{18}A$ | NA |
| | | 85 | PolyTAT | $TA_{18}T$ | NA |
| | | 86 | CpG | TCCATGACGTTCCTGACGTT | NA |
| | | 87 | GpC | TCCAGGACTTCTCTCAGGTT | NA |

[a] Fab fraction generated from a monoclonal antibody directed to peptide p277.
[b] Recombinant protein expressed in bacteria and purified using standard procedures.
[c] Kindly provided by Professor Avraham Ben Nun (The Weizmann Institute of Science).
[d] Kindly provided by Dr Felix Mor (The Weizmann Institute of Science).
[e] Kindly provided by Professor Sara Fuchs (The Weizmann Institute of Science).
[f] Produced at the Statens Seruminstitut, Copenhagen, Denmark.
[g] Kindly provided by Professor R. van deer Zee (Utrecht University, The Netherlands).
[h] Purchased from Pierce (Oud Beijerland, The Netherlands), catalogue number 77153.

## 4. Discussion

The results reported in this article relate to general issues in both biology and informatics. In general, this study confirms that unselected patterns of reactivity can be informative [11,12,23–32] not only in the nervous system, but also in immunology, where the emphasis has been traditionally on specific, high-affinity antibody molecules and single clones of cells. Complex systems use arrays of signals to generate information, and biologists too have to exploit arrays of information. The highly non-trivial task of mining a fairly large array of antigen reactivities for different subjects was accomplished by an unsupervised clustering methodology that identifies (relatively small) correlated groups of antigens, whose reactivities may be used to separate the subjects according to known clinical labels (which are used only *a posteriori*). This method starts from an unsupervised 'holistic' approach that looks at a large number of antigens, and proceeds on a reductionist path, identifying small subsets of antigens that may be relevant to some particular differentiation.

Immunologically, we found that the autoantibodies of the healthy subjects detectible by ELISA bound to many self-antigens implicated in autoimmune diseases (Table 2): histone II A, and single- and double-stranded DNA—targeted in systemic lupus erythematosus [2]; HSP60, insulin and GAD—associated with type 1 diabetes mellitus [22]; myelin basic protein (MBP) and myelin oligodendrocyte glycoprotein (MOG)—associated with multiple sclerosis [33]; the AchR—targeted in myasthenia gravis [1]; tyrosinase—associated with vitiligo [34]; myosin—associated with polymyositis [35]; and cytochrome *c* P450—associated with autoimmune liver disease [36]. Other frequent self-antigens included the serum proteins fibrinogen and clotting factor VII, heparin, enzymes and globin molecules. The present study did not deal with the precise specificity and affinity of the individual antibodies, as is usually done to investigate autoimmunity.

The documentation of autoantibodies binding to self-antigens in healthy people is compatible with the concept of the immunological homunculus [7,37]. The term immunological homunculus refers to the observation

Table 2
Frequencies of autoantibodies in healthy humans (to limit the number of self-antigens shown, the only those antigens are included to which at least 35% of the healthy subjects responded with an OD of greater than 0.3 nm)

| Antigen | Incidence of autoantibodies (%) | |
|---|---|---|
| | IgM+IgG | IgM |
| Tubulin | 50 | – |
| Myosin | 40 | – |
| Heparin | 75 | – |
| Acid phosphatase | 35 | – |
| Annexin 33 kDa | 55 | – |
| Cytochrome *c* P450 | 50 | 80 |
| Catalase | 65 | – |
| Tyrosinase | – | 45 |
| Histone II A | 65 | 45 |
| Double stranded DNA | 75 | 75 |
| Single stranded DNA | 100 | 95 |
| Factor VII | 70 | 100 |
| Fibrinogen | 90 | – |
| HSP60 | 40 | – |
| GAD | 100 | 70 |
| Insulin | 35 | 35 |
| MOG | – | 95 |
| MBP | 35 | – |
| AchR | 90 | 75 |
| Myoglobulin | 65 | 35 |
| Hemoglobin A | 50 | 45 |
| LPS | 85 | 45 |
| TB PPD | 90 | 50 |

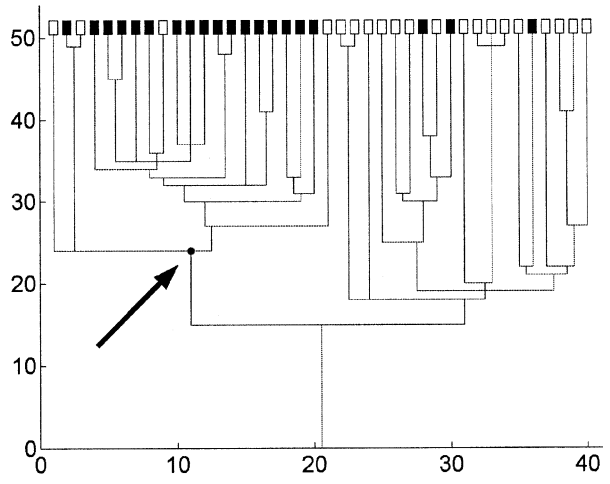



Fig. 2. Dendrogram of healthy subjects and type 1 diabetes subjects. Clustering was done using IgM reactivities to insulin, and IgM and IgM+IgG for collagen I. Healthy subjects are represented by white squares, and type 1 diabetes patients are represented by black squares. We used the cluster marked by the arrow to classify the subjects.

that T- and B-cell autoimmunity in healthy individuals is usually organized around particular sets of self-antigens [38,39]. Homunculus theory proposes that this natural autoimmunity is regulated by various mechanisms that prevent the transition of healthy autoimmunity to autoimmune disease [7,37,40–43]. Indeed, the development of autoimmune disease could be explained most simply by the failure of these control mechanisms [6,7]. The high prevalence of certain autoimmune reactivities shown in Table 2 might explain why the major autoimmune diseases are associated with the abnormal activation of just these autoimmune reactivities; they are already built into the healthy system [7,37]. The autoimmune disease process would seem to expand the quantity and select for high affinity of the particular autoantibodies involved in the disease. The natural autoimmune repertoire, which is quiescent in the healthy state, might serve as the ground for this pernicious autoimmunity. Clearly, we would like to know how natural autoimmunity to particular sets of self-antigens develops, what functions healthy autoimmunity might serve in body maintenance [44,45], how natural autoimmunity is controlled and how it deteriorates into autoimmune disease in certain persons [41,46]. The present study did not explore these biological questions, but rather focused on the informatic questions: (a) whether there are non-trivial structures and correlations in the reactivity patterns of autoantibodies and (b) whether the pattern of autoantibodies present in healthy persons might be distinguished from the pattern of autoantibodies present in an autoimmune disease, taking type 1 diabetes mellitus as our example.

Clinically, diagnostic tests for type 1 diabetes are constructed and standardized in a way that takes advantage of the greater amounts and higher affinities of the autoantibodies that are produced when an autoimmune disease process becomes activated; in clinical testing, the natural autoantibodies we detected, presented in Table 2, are buried in the background [47,48]. The association of type 1 diabetes with antibodies to GAD, an accepted assay, is observed, for example, only when the GAD antibody assay is done according to a standard protocol [3]. Our aim in this study, in contrast to standard procedure, was not to analyze different types of patients and the quantities or affinities of their particular autoantibodies, but to test for the presence of informative patterns in the global array of autoantibodies. This approach has been used in the past for the study of several autoimmune conditions in humans [10–13]. However, these previous studies did not use large panels of defined self-antigens and were not applied to human type 1 diabetes mellitus.

Coutinho and colleagues pioneered the study of patterns of autoantibodies binding to undefined antigens in blots of tissue extracts [23,24]; they used PCA, alone or in combination with hierarchical clustering [23], to study their results. The present work extends the study of patterns to defined self-antigens and is based on a novel cluster analysis [16] that identifies small groups of antigens whose reactivity patterns reflect a subject's clinical state. In this study, we show that healthy subjects and type 1 diabetes subjects can be distinguished, despite the

Table 3
Clustering of type 1 diabetes and healthy human serum samples (the antigen clusters shown in Fig. 1A were used to classify the subjects)

| Cluster number | Antigens | Sensitivity (%) | Specificity (%) | Cluster Size | Number of errors | *P*-value |
|---|---|---|---|---|---|---|
| 1 | IgM to collagen I<br>IgG+M to collagen I<br>IgG+M to insulin | 85 | 81 | 21 | 7 | $8.8 \times 10^{-5}$ |
| 2 | IgM to hAchR<br>IgM aldolase | 85 | 74 | 23 | 9 | 0.0011 |
| 3 | IgM to cartilage extract<br>IgG+M to cardiolipin<br>IgM to cardiolipin | 55 | 100 | 11 | 9 | 0.00015 |
| 4 | IgM to poly arginine<br>IgM to heparin | 80 | 70 | 23 | 11 | 0.0095 |
| 5 | IgM to collagen X<br>IgG+M collagen X | 95 | 63 | 30 | 12 | 0.0084 |
| 6 | IgM to cytochrome *c* P450<br>IgG+M cytochrome *c* P450 | 70 | 67 | 21 | 13 | 0.056 |
| | Combination (more than three classifiers) | 95 | 90 | | | |

presence of various autoantibodies in both groups, by the patterns of their autoantibodies. Furthermore, the patterns of reactivity appear to be disease-specific; type 1 diabetes subjects could be efficiently separated from persons with type II diabetes or Behçet's disease (manuscript in preparation). In other words, patterns of autoantibody reactivity may provide information beyond that seen in a simple one-to-one relationship between an antibody and an antigen [7,11,12,23–32,49]. The present work extends previous findings made in humans (reviewed in Refs. [39,50]) and introduces a novel two-step approach of first clustering the antigens and then using antigen clusters to cluster the subjects.

It is to be noted that we estimated the statistical significance of the clusters by empirically testing the frequency with which similar results could arise by chance using scrambled data. One thousand such computer experiments proved the significance of the antigen clusters derived from the real data. Furthermore, the statistical significance of the separation of diseased and healthy subjects, based on our clusters, was calculated analytically [18]. Finally, the clustering proved robust: the clinical labels of 10 'unknown' subjects could be clustered on the basis of the remaining 30, with a success rate of 90%.

We do not yet know the biological relevance of the particular autoantibodies or of their patterns to the pathophysiology of disease. It is to be noted that the discrimination between the healthy blood donors and the type 1 diabetes patients by clustering does not mean that the informative antibody reactivities are directly involved in the disease process. The differences in autoantibody patterns could have resulted from the disease itself, or from genetic or environmental factors associated directly or indirectly with susceptibility to the disease. Even factors such as age, gender and immunization history were not controlled. Nevertheless, the present findings fit the renewed appreciation of the importance of collective patterns in living systems. Collective interactions that form distinct reactivity patterns bear meaning in signal transduction, gene activation, neoplastic transformation, cell movement, organogenesis, brain function and almost any other subject presently of interest to biologists [51–57]. Biology has succeeded in reducing complex systems to component cells and molecules, but the emergent properties of living systems cannot easily be reduced to the one-to-one relationships of single components; informatic analysis of arrays of data is required. Like other complex systems, the immune system can be mined for information by studying arrays of data. Indeed, the repertoire of autoantibodies present in the individual is much closer to the individual's life experience than are the individual's genes. The immune system, like the brain, is an adaptive bio-informatic system in its own right [58].

## Acknowledgements

I.R.C. is the incumbent of the Mauerberger Chair in Immunology, and Director of the Center for the Study of Emerging Diseases. E.D. is the incumbent of the H. J. Leir Professorial Chair. His research was partially supported by GIF—the Germany–Israel Science Foundation. We thank Yeda Ltd, Dr Isaac Shariv CEO, for funding this study.

## References

[1] Al-Lozi M, Pestronk A. Organ-specific autoantibodies with muscle weakness. Curr Opin Rheumatol 1999;11:483–8.

[2] Pisetsky DS. Anti-DNA and autoantibodies. Curr Opin Rheumatol 2000;12:364–8.
[3] Verge CF, Stenger D, Bonifacio E, Colman PG, Pilcher C, Bingley PJ et al. Combined use of autoantibodies (IA-2 autoantibody, GAD autoantibody, insulin autoantibody, cytoplasmic islet cell antibodies) in type 1 diabetes: combinatorial islet autoantibody workshop. Diabetes 1998;47:1857–66.
[4] Zauli D, Cassani F, Bianchi FB. Auto-antibodies in hepatitis C. Biomed Pharmacother 1999;53:234–41.
[5] Avrameas S, Guilbert B, Dighiero G. Natural antibodies against tubulin, actin, myoglobin, thyroglobulin, fetuin, albumin and transferrin are present in normal human sera, and monoclonal immunoglobulins from multiple myeloma and Waldenstrom's macroglobulinemia may express similar antibody specificities. Ann Immunol 1981;132:231–6.
[6] Cohen IR. Discrimination and dialogue in the immune system. Semin Immunol 2000;12:215–9.
[7] Cohen IR. Tending Adam's garden: evolving the cognitive immune self. London: Academic Press, 2000.
[8] Coutinho A, Forni L, Holmberg D, Ivars F, Vaz N. From an antigen-centered, clonal perspective of immune responses to an organism-centered, network perspective of autonomous activity in a self-referential immune system. Immunol Rev 1984; 79:151–68.
[9] Abbas AK, Lichtman AH, Pober JS. Cellular and molecular immunology. 2nd ed. Philadelphia: W.B. Saunders, 1994.
[10] Sharshar T, Lacroix-Desmazes S, Mouthon L, Kaveri S, Gajdos P, Kazatchkine MD. Selective impairment of serum antibody repertoires toward muscle and thymus antigens in patients with seronegative and seropositive myasthenia gravis. Eur J Immunol 1998;28:2344–54.
[11] Ronda N, Haury M, Nobrega A, Kaveri SV, Coutinho A, Kazatchkine MD. Analysis of natural and disease-associated autoantibody repertoires: anti-endothelial cell IgG autoantibody activity in the serum of healthy individuals and patients with systemic lupus erythematosus. Int Immunol 1994;6:1651–60.
[12] Sundblad A, Ferreira C, Nobrega A, Haury M, Ferreira E, Padua F et al. Characteristic generated alterations of autoantibody patterns in idiopathic thrombocytopenic purpura. J Autoimmun 1997;10:193–201.
[13] Stahl D, Lacroix-Desmazes S, Heudes D, Mouthon L, Kaveri SV, Kazatchkine MD. Altered control of self-reactive IgG by autologous IgM in patients with warm autoimmune hemolytic anemia. Blood 2000;95:328–35.
[14] Coutinho A, Avrameas S. Speculations on immunosomatics: potential diagnostic and therapeutic value of immune homeostasis concepts. Scand J Immunol 1992;36:527–32.
[15] Abulafia-Lapid R, Elias D, Raz I, Keren-Zur Y, Atlan H, Cohen IR. T cell proliferative responses of type 1 diabetes patients and healthy individuals to human hsp60 and its peptides. J Autoimmun 1999;12:121–9.
[16] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A 2000; 97:12079–84.
[17] Blatt M, Wiseman S, Domany E. Super-paramagnetic clustering of data. Phys Rev Lett 1996;76:3251–5.
[18] Fisher RA. The logic of inductive inference. J R Stat Soc 1935; 98:39–82.
[19] Duda OR, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley, 2001.
[20] Cover TM. Geometrical and statistical properties with application in pattern recognition. IEEE Trans Electr Comput 1965; 14:326–34.
[21] Gardner E. Maximum storage capacity in neural networks. Europhys Lett 1987;4:471–85.
[22] Tisch R, McDevitt H. Insulin-dependent diabetes mellitus. Cell 1996;85:291–7.
[23] Nobrega A, Haury M, Grandien A, Malanchere E, Sundblad A, Coutinho A. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological 'homunculus' of antibodies in normal serum. Eur J Immunol 1993;23:2851–9.
[24] Haury M, Grandien A, Sundblad A, Coutinho A, Nobrega A. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. Scand J Immunol 1994;39:79–87.
[25] Haury M, Sundblad A, Grandien A, Barreau C, Coutinho A, Nobrega A. The repertoire of serum IgM in normal mice is largely independent of external antigenic contact. Eur J Immunol 1997; 27:1557–63.
[26] Malanchere E, Marcos MA, Nobrega A, Coutinho A. Studies on the T cell dependence of natural IgM and IgG antibody repertoires in adult mice. Eur J Immunol 1995;25:1358–65.
[27] Mouthon L, Nobrega A, Nicolas N, Kaveri SV, Barreau C, Coutinho A et al. Invariance and restriction toward a limited set of self-antigens characterize neonatal IgM antibody repertoires and prevail in autoreactive repertoires of healthy adults. Proc Natl Acad Sci U S A 1995;92:3839–43.
[28] Mouthon L, Lacroix-Desmazes S, Nobrega A, Barreau C, Coutinho A, Kazatchkine MD. The self-reactive antibody repertoire of normal human serum IgM is acquired in early childhood and remains conserved throughout life. Scand J Immunol 1996; 44:243–51.
[29] Nobrega A, Grandien A, Haury M, Hecker L, Malanchere E, Coutinho A. Functional diversity and clonal frequencies of reactivity in the available antibody repertoire. Eur J Immunol 1998; 28:1204–15.
[30] Vasconcellos R, Nobrega A, Haury M, Viale AC, Coutinho A. Genetic control of natural antibody repertoires. I. IgH, MHC and TCR beta loci. Eur J Immunol 1998;28:1104–15.
[31] Nobrega A, Haury M, Gueret R, Coutinho A, Weksler ME. The age-associated increase in autoreactive immunoglobulins reflects a quantitative increase in specificities detectable at lower concentrations in young mice. Scand J Immunol 1996;44:437–43.
[32] Nobrega A, Stransky B, Nicolas N, Coutinho A. Regeneration of natural antibody repertoire after massive ablation of lymphoid system: robust selection mechanisms preserve antigen binding specificities. J Immunol 2002;169:2971–8.
[33] Link H, Baig S, Jiang YP, Olsson O, Hojeberg B, Kostulas V et al. B cells and antibodies in MS. Res Immunol 1989;140:219–26 [discussion p. 45–8].
[34] Jimbow K. Biological role of tyrosinase-related protein and its relevance to pigmentary disorders (vitiligo vulgaris). J Dermatol 1999;26:734–7.
[35] Erlacher P, Lercher A, Falkensammer J, Nassonov EL, Samsonov MI, Shtutman VZ et al. Cardiac troponin and beta-type myosin heavy chain concentrations in patients with polymyositis or dermatomyositis. Clin Chim Acta 2001;306:27–33.
[36] Boitier E, Beaune P. Xenobiotic-metabolizing enzymes as autoantigens in human autoimmune disorders. An update. Clin Rev Allergy Immunol 2000;18:215–39.
[37] Cohen IR. The cognitive paradigm and the immunological homunculus. Immunol Today 1992;13:490–4.
[38] Goldrath AW, Bevan MJ. Selecting and maintaining a diverse T-cell repertoire. Nature 1999;402:255–62.
[39] Lacroix-Desmazes S, Kaveri SV, Mouthon L, Ayouba A, Malanchere E, Coutinho A et al. Self-reactive antibodies (natural autoantibodies) in healthy individuals. J Immunol Methods 1998; 216:117–37.
[40] Cohen IR. Regulation of autoimmune disease physiological and therapeutic. Immunol Rev 1986;94:5–21.
[41] Hurez V, Kaveri SV, Kazatchkine MD. Expression and control of the natural autoreactive IgG repertoire in normal human serum. Eur J Immunol 1993;23:783–9.

[42] Kumar V, Sercarz E. Distinct levels of regulation in organ-specific autoimmune diseases. Life Sci 1999;65:1523–30.
[43] Bach JF, Chatenoud L. Tolerance to islet autoantigens in type 1 diabetes. Annu Rev Immunol 2001;19:131–61.
[44] Moalem G, Leibowitz-Amit R, Yoles E, Mor F, Cohen IR, Schwartz M. Autoimmune T cells protect neurons from secondary degeneration after central nervous system axotomy. Nat Med 1999;5:49–55.
[45] Schwartz M, Cohen IR. Autoimmunity can benefit self-maintenance. Immunol Today 2000;21:265–8.
[46] Moudgil KD, Sercarz EE. The self-directed T cell repertoire: its creation and activation. Rev Immunogenet 2000;2:26–37.
[47] Kyriatsoulis A, Manns M, Gerken G, Lohse AW, Ballhausen W, Reske K et al. Distinction between natural and pathological autoantibodies by immunoblotting and densitometric subtraction: liver–kidney microsomal antibody (LKM) positive sera identify multiple antigens in human liver tissue. Clin Exp Immunol 1987;70:53–60.
[48] Bouanani M, Dietrich G, Hurez V, Kaveri SV, Del Rio M, Pau B et al. Age-related changes in specificity of human natural autoantibodies to thyroglobulin. J Autoimmun 1993;6:639–48.
[49] Ferreira C, Mouthon L, Nobrega A, Haury M, Kazatchkine MD, Ferreira E et al. Instability of natural antibody repertoires in systemic lupus erythematosus patients, revealed by multiparametric analysis of serum antibody reactivities. Scand J Immunol 1997;45:331–41.
[50] Stahl D, Lacroix-Desmazes S, Mouthon L, Kaveri SV, Kazatchkine MD. Analysis of human self-reactive antibody repertoires by quantitative immunoblotting. J Immunol Methods 2000;240:1–14.
[51] Augenlicht LH, Bordonaro M, Heerdt BG, Mariadason J, Velcich A. Cellular mechanisms of risk and transformation. Ann NY Acad Sci 1999;889:20–31.
[52] Berman DE, Dudai Y. Memory extinction, learning anew, and learning the new: dissociations in the molecular machinery of learning in cortex. Science 2001;291:2417–9.
[53] Downward J. The ins and outs of signalling. Nature 2001; 411:759–62.
[54] Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S et al. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 2001;292:2080–3.
[55] Moser B, Loetscher P. Lymphocyte traffic control by chemokines. Nat Immunol 2001;2:123–8.
[56] Rao CV, Arkin AP. Control motifs for intracellular regulatory networks. Annu Rev Biomed Eng 2001;3:391–419.
[57] Wilkie AO, Morriss-Kay GM. Genetics of craniofacial development and malformation. Nat Rev Genet 2001;2:458–68.
[58] Atlan H, Cohen IR. Immune information, self-organization and meaning. Int Immunol 1998;10:711–7.

# Publication 7:

## Coupled Two-Way Clustering Analysis of Breast Cancer and Colon Cancer Gene Expression Data

Authors: G. Getz, H. Gal, I. Kela, D. Notterman and E. Domany
Published in: *Bioinformatics* **19**, 1079–1089 (2003).

BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1079–1089
DOI: 10.1093/bioinformatics/btf876

# Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data

*Gad Getz [1], Hilah Gal [1], Itai Kela [1], Daniel A. Notterman [2, 3] and Eytan Domany [1,*]*

*[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel, [2]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA and [3]Department of Pediatrics, Robert Wood Johnson Medical School, New Brunswick, NJ 08903, USA*

Received on May 22, 2002; accepted on September 16, 2002

**ABSTRACT**

We present and review coupled two-way clustering, a method designed to mine gene expression data. The method identifies submatrices of the total expression matrix, whose clustering analysis reveals partitions of samples (and genes) into biologically relevant classes. We demonstrate, on data from colon and breast cancer, that we are able to identify partitions that elude standard clustering analysis.

**Availability:** Free, at http://ctwc.weizmann.ac.il.

**Contact:** eytan.domany@weizmann.ac.il

**Supplementary information:** http://www.weizmann.ac.il/physics/complex/compphys/bioinfo2/

## INTRODUCTION

Two nearly concurrent recent advances—the development of high density DNA chips and the deciphering of the human genome—hold great promise for significant progress in biomedical research. A large umber of studies have been published within the last years, attempting to classify, explain and perhaps help cure several human diseases, on the basis of gene expression levels measured for populations of diseased and healthy subjects. Different forms of cancer have been at the focus of such studies from early on, using all available chip technologies.

A DNA chip measures simultaneously the expression levels of thousands of genes for a particular sample. Since a typical experiment on human subjects provides the expression profiles of several tens of samples (say $N_s \approx 100$), over several thousand ($N_g$) genes whose expression levels passed some threshold, the outcome of such an experiment contains between $10^5$ and $10^6$ numbers. These are summarized in an $N_g \times N_s$ *expression table*; each row corresponds to one particular gene and each column to a sample, with the entry $E_{gs}$ representing the expression level of gene $g$ in sample $s$. Analysis of such massive amounts of data poses a serious challenge for the development and application of novel methodologies.

We present here *coupled two-way clustering* (CTWC), a recently introduced method (Getz *et al.*, 2000a), designed to 'mine' gene expression data, and demonstrate its strength by applying it to breast cancer and colon cancer data. The CTWC software is accessible at http://ctwc.weizmann.ac.il (Getz and Domany, 2003).

CTWC is based on *clustering*, and as such it is *unsupervised* and capable of discovering unanticipated partitions of the data, exploring its structure on the basis of correlations and similarities that are present in it. In the context of gene expression, such analysis has two obvious goals:

(1) Find groups of genes that have correlated expression profiles. The members of such a group may take part in the same biological process.

(2) Divide the tissues into groups with similar gene expression profiles. Tissues that belong to one group are expected to be in the same biological (e.g. clinical) state.

The straightforward way to carry out such analysis is to cluster the data in *two ways*. Denote the set of all genes that passed a threshold by $G1$ and the set of all samples by $S1$. Each gene is a point in an $|S1|$ dimensional space; the first clustering operation, $G1(S1)$, clusters all genes on the basis of their expression levels over all samples. The complementary operation, $S1(G1)$, clusters the samples on the basis of their expression levels over all $|G1|$ genes. A variety of clustering methods have been used to perform these operations. Clustering is based on some measure of similarity of pairs of samples $s, s'$ which, in turn, is governed by their 'distance' in the $|G1|$ dimensional space of expression levels.

As several groups noticed (Perou *et al.*, 2000; Cheng

*To whom correspondence should be addressed.

and Church, 2000; Califano *et al.*, 2000; Ihmels *et al.*, 2002; Tanay *et al.*, 2002), one runs into a severe difficulty with this simple 'all against all' clustering approach. The reason is that in general only a small subset of $N_r$ relevant genes is involved in one particular biological process of interest. Since usually $N_r \ll |G1|$, the 'signal' provided by this subset may be completely masked by the 'noise' generated by the much larger number of the other genes. Furthermore, it may well happen that in order to assign samples into two clinically meaningful classes (e.g. adenoma and carcinoma) on the basis of the $N_r$ relevant genes, one must first remove a previously identified group of samples (e.g. healthy tissue), and cluster only the remaining $N_s' < N_s$ tumors (using only the $N_r$ relevant genes). Thus one should look for special $N_r \times N_s'$ *submatrices* of the total expression matrix; such a search is problematic since an exhaustive enumeration of such submatrices is of exponential complexity. CTWC provides a heuristic method to search for such submatrices. It has been used successfully to mine data (Getz *et al.*, 2000a) from experiments on colon cancer (Alon *et al.*, 1999) and leukemia (Golub *et al.*, 1999), glioblastoma (Godard *et al.*, 2003), breast cancer (Kela, 2002) and antigen chips (Quintana *et al.*, 2003). We present here results obtained by a new, more interactive usage of CTWC on cDNA microarray data from breast cancer (Perou *et al.*, 2000, referred to as **PAL**; Sorlie *et al.*, 2001, referred to as **SAL**) and on oligonucleotide microarray data from colon cancer patients (Notterman *et al.*, 2001).

The analysis of Notterman *et al.* stopped at two way clustering, which is the first step of CTWC—here our aim is to demonstrate that by going beyond this step we uncover new partitions of the samples. The situation with the breast cancer data is more interesting. PAL noticed that simple two way clustering did not partition the samples in a meaningful way, and pruned their original set of $|G1| = 1753$ down to 496 'intrinsic genes', that were selected in a knowledge based way (which can be applied only if the data contains pairs of samples taken from the same patients). CTWC also identifies (much smaller) sets of genes that are used to cluster the samples, but it is done in an automated, objective, generally applicable way. It was not clear a priori that CTWC will reproduce the valuable observations of PAL and SAL, and even less that it will yield new results of possible biological or clinical significance.

## MATERIALS AND METHODS

*Expression data—breast cancer.* We studied two data sets on breast cancer. The first expression matrix was measured and analyzed by PAL and the second by SAL. The PAL study characterizes gene expression profiles of 84 samples (the set $S$), composed of 65 tumors (sample set $S1$) and 19 cell lines, using cDNA microarrays, representing 8,102 human genes. Twenty of the 65 tumors were sampled twice; 18 from patients who were treated with doxorubicin (chemotherapy) for an average of 16 weeks, with surgical biopsy done *before* and *after* the treatment, and two more tumors were paired with a lymph node metastasis from the same patient. The 25 remaining specimens included 22 tumors and three samples from normal breast tissues (nevertheless, we refer to these also as 'tumors'). The full expression matrix included 8,102 rows, each corresponding to a gene, and 84 columns, each corresponding to a sample. PAL first selected the subset of genes whose expression varied by at least four-fold from the median of the samples, in at least three of the samples tested. This filtering process left the set $G1$ of 1753 genes, each of which is represented by 84 expression values. In the final expression matrix PAL split the data into two submatrices; one of tissues and one of cell lines. The two submatrices were, separately, median polished (the rows and columns were iteratively adjusted to have median 0) before being rejoined into a single matrix. The expression matrix was two-way clustered; clustering the genes on the basis of the 84 samples [operation $G1(S)$], and clustering the 65 tumors using all 1753 genes [$S1(G1)$]. Since $S1(G1)$ did not yield any meaningful partition, PAL concluded that the 1753 genes were not an optimal set to classify the tumors, and they selected a subset $G^{(int)}$ of 496 'intrinsic' genes in the following way. They calculated for each gene an index that measures the variation of its expression between different tumors versus between paired samples from the same tumor. They ranked all 8102 genes according to this index, and chose the 496 top scorers. They argued that the expression levels of the top scorers on this list represent inherent properties of the tumors themselves rather than just differences between different samplings. From this point on they used the $496 \times 65$ expression level matrix to cluster the genes of $G^{(int)}$ and the tumors $S1$. This data is publicly available at the Stanford website (see PAL).

The second study of breast cancer, by SAL, characterized gene expression profiles of 85 tissue samples representing 84 individuals. 78 of these were breast carcinomas (71 ductal, five lobular, and two ductal carcinomas in situ, obtained from 77 different individuals; two tumors were from one individual, diagnosed at different times) three were fibroadenomas and four normal breast tissue samples were also included; three of these were pooled normal breast samples from multiple individuals (CLONTECH). These 85 samples included 40 tumors that were previously analyzed and described by PAL. Fifty-one of the patients were part of a prospective study on locally advanced breast cancer (T3/T4 and/or N2 tumors) treated with doxorubicin monotherapy before surgery followed by adjuvant tamoxifen in the case of positive ER and/or progesterone re-

ceptor (PgR) status (Geisler *et al.*, 2001). All but three patients were treated with tamoxifen. ER and PgR status was determined by using ligand-binding assays, and mutation analysis of the TP53 gene was performed as described in Geisler *et al.*. The cDNA microarrays used in this study were from several different print runs that all contained the same core set of 8,102 genes. In total, the 85 microarray experiments were carried out by using six different batches of microarrays and three different batches of common reference, each independently produced. SAL performed cluster analysis on two subsets of genes. One subset, of 456 cDNA clones (427 unique genes), was selected from the 496 'intrinsic' gene list, previously described by PAL. The second subset consisted of 264 cDNA clones, that exhibit high correlation with patient survival, selected from the set $G1$ of 1753 genes. Clustering analysis and patient classifications were based on the total set of 78 malignant breast tumors. Survival analysis was based on 49 patients with locally advanced tumors and no distant metastases (two of the 51 patients from this prospective study were retrospectively recorded to have a minor lung deposit and a liver metastasis, respectively) that were treated with neoadjuvant chemotherapy and adjuvant tamoxifen (Geisler *et al.*, 2001)

*Expression data—colon cancer* In addition, we studied a data set on colon cancer, previously published by Notterman *et al.*. The data set contains 22 tumor samples; 18 carcinoma and four adenoma, and their paired normal samples. The experiments with carcinoma and paired normal tissue were performed with the Human 6500 GeneChip Set (Affymetrix), and the experiments with the adenomas and their paired normal tissue were performed with the Human 6800 GeneChip Set (Affymetrix). First, following Notterman et.al, we created a composite database that included only accession numbers represented on both GeneChip versions. Values lower than 1 were adjusted to 1. Prior to application of CTWC, we filtered the data using a filtering operation very close to that used by Notterman *et al.*, remaining with 1592 genes. Data from the two different chips were brought to the same average expression level. The data was then log-transformed, centered about the mean and normalized. Second, we studied the 18 paired carcinoma samples separately. Of the 6600 cDNAs and ESTs represented on the array, only genes for which the standard deviation of their log-transformed expression values was greater than 1, were selected. After this filtering process we remain with 768 genes. These values were centered and normalized, prior to application of the CTWC algorithm. The samples were labeled according to additional information about the histological characteristics of the tumor samples, the estimated percentage of contamination with non-tumor cells, the presence of mutations in the p53 gene, the clinical disease stage and the mRNA extraction protocol that has been used.

## ALGORITHM

Since both SPC and CTWC have been described in detail elsewhere, we present here only brief, albeit self-contained reviews of the procedures.

### Superparamagnetic clustering—SPC

The idea behind this algorithm is rooted in the physics and phase transitions of disordered magnets ((Blatt *et al.*, 1996); for a detailed description see Blatt *et al.*, 1997). The four-step procedure presented here uses terminology of graph partitioning, which is more familiar to computer scientists.

*Step 1: Weighted graph.* $N$ data points are associated with 'positions' $\mathbf{X}_i$ in a $D$-dimensional space; they constitute $N$ nodes of a graph. Each node $i$ is connected by an edge $\langle ij \rangle$ to its neighbors $j$. We identify the neighbors $j$ of node $i$ on the basis of the distances $d_{ij} = |\mathbf{X}_i - \mathbf{X}_j|$[†]; the two points are neighbors if $j$ is one of the $K$ closest neighbors of $i$, and vice versa[‡]. To each edge $\langle ij \rangle$ we assign a weight $J_{ij} = f(d_{ij})$ where $f(x)$ is a decreasing function[§] of $x$.

*Step 2: Cost function for graph partitions.* To characterize a partition of the graph, we assign to every vertex $i$ an integer label (a Potts spin variable in Physics terminology), $S_i = 1, 2, \ldots q$[¶]. Any particular assignment of labels, $\{S_1, S_2, \ldots S_N\}$ corresponds to a partition of the graph, and is denoted by $\{S\}$ (in the physics terminology $\{S\}$ is referred to as a 'spin configuration'). $S_i = S_j$ indicates that in the partition $\{S\}$, nodes $i$ and $j$ belong to the same component, whereas $S_i \neq S_j$ means that they are in different components. We use the cost function

$$\mathcal{H}(\{S\}) = \sum_{\langle i,j \rangle} J_{ij} \left(1 - \delta_{S_i,S_j}\right). \qquad (1)$$

The sum runs over all the edges $\langle ij \rangle$ of the graph. No penalty is associated with $\langle ij \rangle$ if nodes $i$ and $j$ belong to the same component. If they belong to different components, edge $\langle ij \rangle$ picks up a penalty $J_{ij}$. Since for small $d_{ij}$ the value of $J_{ij}$ is high, this cost function places a high penalty for assigning two similar nodes to different components. The lowest cost, $\mathcal{H}(\{S\}) = 0$ is obtained

[†] Normally Euclidean distances are used.

[‡] $K$ is a parameter of the algorithm - for genes we use $10 \leq K \leq 20$. By superimposing the minimal spanning tree, we ensure that all vertices belong to a single connected component of the graph.

[§] We use $f(x) = (1/\sqrt{2\pi}a)\exp[-x^2/2a^2]$ ($a$ is the average of $d_{ij}$).

[¶] In many of the applications we tried, Potts spins with $q = 20$ states were used. $q$ has nothing to do with the number of clusters determined by the algorithm - see below.

when all data points are in the same group; the highest cost is reached if no point is in the same group as any of its neighbors. Hence the value of $\mathcal{H}(\{S\})$ reflects the *resolution* at which the partition $\{S\}$ views the data.

*Step 3: Ensemble of partitions.* Rather than choosing any particular partition (say by minimizing the cost function), we consider all configurations $\{S\}$ that have (nearly) the same value of $\mathcal{H}(\{S\}) = E$; to each of these we give the *same statistical weight*, whereas all $\{S'\}$ that correspond to different resolutions (and hence $\mathcal{H}(\{S'\}) \neq E$) get vanishing probability. This assignment of equal probabilities $P(\{S\})$ is the result of maximizing the entropy in order to generate an ensemble of partitions $\{S\}$, for which the only available information is that they have a particular fixed value of the cost $E$. The resulting ensemble of partitions is the microcanonical ensemble of Statistical Mechanics. For each value of $E$ one can sample this ensemble and measure average values of any quantity of interest (see below). It is, however, *technically more convenient* to use for such measurements the canonical ensemble. In this ensemble the weights $P(\{S\})$ are again assigned by maximizing the entropy. However, rather than allowing only partitions with a fixed resolution or cost $\mathcal{H} = E$, one requires that the ensemble average of $\mathcal{H}$ takes the value $E$:

$$\langle \mathcal{H} \rangle = \sum_{\{S\}} P(\{S\}) \mathcal{H}(\{S\}) = E. \tag{2}$$

This requirement is imposed as a constraint under which entropy is maximized, by means of a Lagrange multiplier, denoted $1/T$. In physics terminology $T$ is called the *temperature*. Rather than working at fixed $E$ one works (generates samples and takes averages—see below) at fixed $T$. By fixing the value of $T$ one controls, in effect, the resolution $E$; the two ensembles are completely equivalent in the limit of large number of data points (or spins). In the resulting canonical statistical ensemble of partitions each $\{S\}$) appears with the statistical (Boltzmann) weight

$$P(\{S\}) = e^{-\mathcal{H}(\{S\})/T} \Big/ \sum_{\{S'\}} e^{-\mathcal{H}(\{S'\})} . \tag{3}$$

At $T = 0$ only groupings with $E = 0$ have non-vanishing weight; at $T = \infty$ all partitions have equal weight. For a sequence of values of $T$ we calculate, by Monte Carlo simulation, the equilibrium average $\langle A \rangle$ of several quantities $A$ of interest, such as the magnetization, susceptibility and correlation of neighbor spins. The latter is the most important quantity we measure—the corresponding 'operator' is $A = \delta_{S_i,S_j}$, i.e. an indicator which takes the value 1 if points $i$ and $j$ are in the same component in partition $\{S\}$. The ensemble average of this object is the correlation function:

$$G_{ij} = \langle \delta_{S_i,S_j} \rangle, \tag{4}$$

$G_{ij}$ is the probability to find, at the resolution set by $T$, the data points $i, j$ assigned to the the same component. By the relation to granular ferromagnets we expect that the distribution of $G_{ij}$ is bimodal; if both spins belong to the same *ordered* grain (cluster), their correlation is close to 1; if they belong to two clusters that are not relatively ordered, the correlation is close to $1/q$.

*Step 4: Identifying clusters.* To produce 'hard' clusters on the basis of the $G_{ij}$, we construct a new graph, in a three-step procedure.

(1) Build the clusters' 'core' by thresholding $G_{ij}$. For every pair of neighbors $i$ and $j$, check whether $G_{ij} > \theta = 0.5$; if true, set a 'link' between $i, j$. Because of the bimodality of the distribution of $G_{ij}$ the decision to link $i, j$ depends very weakly on the value of $\theta$.

(2) Capture points lying on the periphery of the clusters by linking each point $i$ to its neighbor $j$ of maximal correlation $G_{ij}$.

(3) Data clusters are identified as the linked components of the graphs obtained in steps 1,2.

At $T = 0$ this procedure generates a single cluster of all $N$ points. At $T = \infty$ we have $N$ independent spins, and the procedure yields $N$ clusters, with a single point in each. Hence as $T$ increases, we generate a dendrogram of clusters of decreasing sizes.

This algorithm has several attractive features (Blatt *et al.*, 1997). One of these is the ability to identify stable (and statistically significant) clusters, which makes SPC most suitable to be used within the framework of CTWC. Furthermore, it allows a quantitative estimation of the $P$-value of a clustering operation, by clustering repeatedly randomized data and checking the fraction of instances in which stable clusters (i.e. as stable as those obtained for non random data) appeared. We identify stable clusters as follows. As we heat the system up, we record for every cluster two temperatures: $T_1$, at which it is 'born' (splits from its parent cluster) and $T_2$, at which it 'dies' (splits into siblings). The ratio $R = T_2/T_1$ is a measure of a cluster's stability. For example, in (Getz *et al.*, 2000a) we set the threshold $R_c$, beyond which a cluster is considered stable, at a value for which not even one of 500 experiments on randomized data gave a cluster with $R \geq R_c$.

SPC was used in a variety of contexts, ranging from computer vision (Domany *et al.*, 1999) to speech recognition (Blatt *et al.*, 1997). Its first direct application to gene expression data has been (Getz *et al.*, 2000b) for analysis of the temporal dependence of the expression levels in a

synchronized yeast culture (Eisen *et al.*, 1998), identifying gene clusters whose variation reflects the cell cycle.

Subsequently, SPC was used to identify primary targets of p53 (Kannan *et al.*, 2001) and p73 (Fontemaggi *et al.*, 2002).

### Coupled two way clustering—CTWC

The main motivation for introducing CTWC (Getz *et al.*, 2000a) was to *increase the signal to noise ratio* of the expression data. The method is designed to overcome two different kinds of 'noise'. The first was mentioned above; say only a small subset of $N_r$ genes participate in a biological process of interest, associated with a particular disease $A$. In this case we expect these $N_r$ genes to have correlated expressions over subjects with disease $A$. This correlation could, in principle, identify the diseased subjects as 'close' in expression space—but, in fact, for $N_r \ll |G1|$ the non-participating $|G1| - N_r$ genes completely mask the effect of the relevant ones on the distance between two diseased subjects. Hence as far as the process of interest is concerned, the non-participating $|G1| - N_r$ genes contribute nothing but noise, that masks the signal of the $N_r$ relevant ones. CTWC eliminates this noise by discarding the irrelevant genes.

The second noise-reducing feature of CTWC is that it uses the expression levels of a a set of genes, rather than one gene at a time. Thereby intrinsic noise in the expression averages out.

CTWC is an iterative process, whose starting point is the standard two way clustering mentioned above, i.e. the clustering operations $S1(G1)$ and $G1(S1)$. We keep two registers—one for stable gene clusters and one for stable sample clusters. Initially we place $G1$ in the first and $S1$ in the second. From $S1(G1)$ and $G1(S1)$ we identify stable clusters of samples and genes, respectively, i.e. those for which the SPC stability index $R$ exceeds a critical value and whose size is not too small. Stable gene clusters are denoted as $GI$ with $I = 2, 3, \ldots$ and stable sample clusters as $SJ$, $J = 2, 3, \ldots$ In the next iteration we use every gene cluster $GI$ (including $I = 1$) as the feature set, to characterize and cluster every sample set $SJ$. These operations are denoted by $SJ(GI)$; (note that $S1(G1)$ was already performed). In effect, we use every stable gene cluster as a possible 'relevant gene set'; the submatrices defined by $SJ$ and $GI$ are the ones we study. Similarly, all the clustering operations of the form $GI(SJ)$ are also carried out. In all clustering operations we check for the emergence of partitions into stable clusters, of genes and samples. If we obtain a new stable cluster, we add it to our registers and record its members, as well as the clustering operation that gave rise to it. If a certain clustering operation did not give rise to new significant partitions, we move down the list of gene and sample clusters to the next pair.

This heuristic identification of relevant gene sets and submatrices is nothing but an exhaustive search among the stable clusters that were generated. The number of these, emerging from $G1(S1)$, is a few tens, whereas $S1(G1)$ usually generates only a few stable sample clusters. Hence the next stage typically involves less than a hundred clustering operations. These iterative steps stop when no new stable clusters beyond a preset minimal size are generated, which usually happens after the first or second level of the process.

Since the $N_r$ relevant genes are expected to have correlated expression levels over at least a significant subset of the samples, we can expect at least a subset of them to form a stable cluster. Then when the members of such a cluster are used to recluster the samples, the noise generated by the very many irrelevant genes will be filtered out and we will get a clear separation of the samples to the desired classes. When CTWC was first introduced (Getz *et al.*, 2000a), we also studied several cases of artificially generated expression data, into which various correlations, partitions and sub-partitions were incorporated and then masked. CTWC successfully unraveled all this hidden structure from these toy problems (see link in Supplementary Information).

In a typical analysis we generate between 10 and 100 interesting partitions, which are searched[‖] for biologically or clinically interesting findings, on the basis of the genes that gave rise to the partition and on the basis of available clinical labels of the samples. It is important to note that these labels are used *a posteriori*, after the clustering has taken place, to interpret and evaluate the results.

## RESULTS

Lists of the genes that constitute each of the clusters $GI$ mentioned below are given in the supplementary information. One should note that in the experiments analyzed here no replicates of the measurements were made.

### Breast cancer—PAL

We posed the following questions:

(1) Do our methods of analysis reproduce the results obtained by PAL?

(2) Can we make observations that seem to be of interest and were not reported by PAL?

As to the first question—CTWC reproduced all the main findings of PAL directly, starting from the entire set $G1$ of 1753 genes, without filtering them to the intrinsic set.

[‖] This search is done in an automated manner, calculating various figures of merit for each stable cluster, defined on the basis of clinical or genetic information.

Second, we found new tumor classifications that were not mentioned by PAL.

*Reproducing the results of PAL.* PAL used lower case letters to identify gene clusters, and colors for samples (see their Figures 1 and 3). We use below their notation when comparisons are made.

$G1(S)$: Following PAL, we used the same feature set, $S$, of all samples and cell lines, to cluster $G1$, the full set of 1753 genes. Since we also used the same normalization, this operation provides a direct comparison of Average Linkage (the clustering method used by PAL) and SPC. All the gene clusters that were marked as interesting by PAL, were also found by our clustering operation (Kela, 2002).

$S(G1)$: Next, we clustered (separately) the cell lines and the tumors, using all 1753 genes. Since our normalization here differs from that of PAL, we cannot compare directly our results. However, in agreement with PAL, we also did not find any meaningful partitions of the tumors, $S1$, from this operation, leading to the same conclusion as reached by PAL: namely, that $G1$ is not suitable to classify the tumors and we should characterize them using different subsets of genes. From here on CTWC deviates from the procedure of PAL, who selected their 'intrinsic set' of 496 genes in a way that (a) necessitates having paired samples from the same patients (*before* and *after* chemotherapy), and (b) assumes that only genes that meet their criteria (similarity of matched samples) are to be used. CTWC, on the other hand, is an automated process, performing operations $S1(GI)$, i.e. clustering the tumors $S1$ using different stable gene clusters $GI$, one at a time. Clustering the 65 samples on the basis of these small subsets of genes, one at a time, enabled us to identify the subclasses of tumors that PAL found using their intrinsic set.

$S1(G4)$: Cluster $G4$ (that was obtained by the G1(S) clustering process) has 10 genes—it is our homologue of cluster j of PAL (see their Figure 1). The operation $S1(G4)$ generates a stable sample cluster which is quite similar to the ER+/luminal-like (blue) cluster of PAL (see their Figure 3); its members have high expression levels of $G4$. $S1(G4)$ identifies also PAL's basal-like (yellow) group, characterized by low expression levels of the $G4$ genes.

$S1(G46)$, $S1(G9)$: $G46$ is a cluster of 33 genes that are part of the proliferation cluster found by PAL. The operation $S1(G46)$ produces a good homologue of their normal-like (green) cluster. Members of this group show low expression levels of $G46$ genes. The normal-like samples are also identified in the operation $S1(G9)$: the 13 genes of $G9$ are a subgroup of cluster g of PAL. Normal-like tissues have high expression levels of the $G9$ genes.

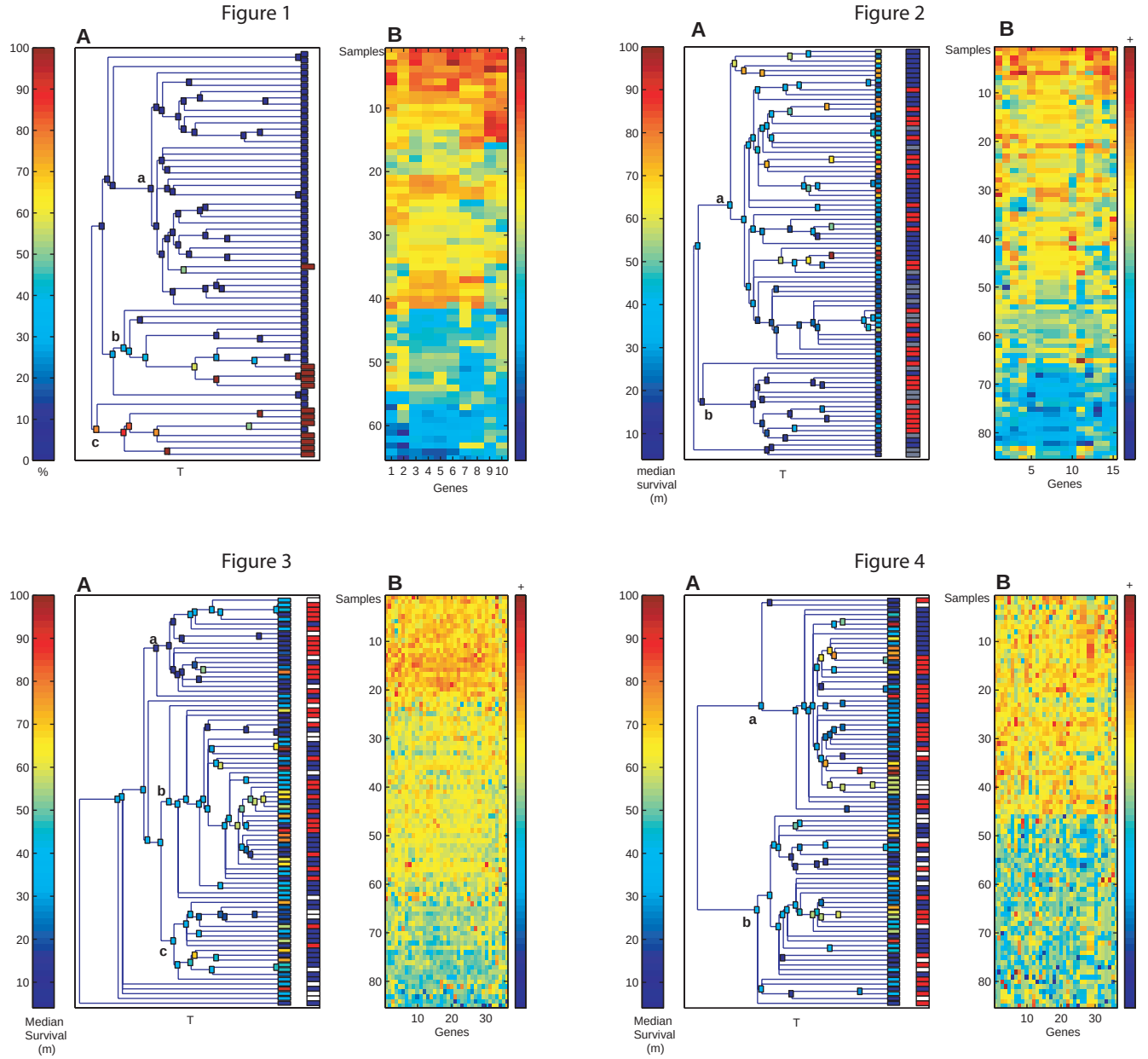
$S1(G21)$: This operation separates the Erb-B2+ (red) cluster from the other samples. $G21$ is homologous to gene cluster d from Figure 3 of PAL; its expression is high in the Erb-B2+ tumors.

*New observations (beyond PAL).* Of several new findings (Kela, 2002) we chose to highlight here one that bears on an issue that has been considered important by PAL: that of separating the ER+ and ER- tumors on the basis of their expression levels. We present two such classifiers, which demonstrate two different advantages of CTWC. The first classifier *could have been* discovered by PAL, since it is based on genes that *do belong to PAL's intrinsic set*, but their effect is masked by the large number of the 496 'intrinsic' genes; to see it, one has to zero in on a small subset, as is done by CTWC. The second classifier *could not have been discovered by PAL's analysis* since it is based on genes that are *not included in their intrinsic set*.

$S1(G4)$: The cluster $G4$ (10 genes) was described above—it is practically identical to cluster j from Figure 1 of PAL and to cluster c of their Figure 3. It contains the estrogen receptor and three other transcription factors (see supplementary information of PAL) related to the estrogen receptor pathway. The operation $S1(G4)$ generated the dendrogram presented in Figure 1A. The variation in the expression levels of the $G4$ genes correlates well with the direct clinical measurements of the ER protein levels in the tumors (supplementary information of PAL).

In the dendrogram Figure 1A the boxes representing sample clusters were colored according to the percentage of ER- samples, ranging from red (100%) to blue (0%). In Figure 1B the samples were ordered according to the dendrogram, and the colors represent the expression levels of the 10 genes. SPC generated three main branches (clusters); the upper **a** with highest expression values, **b** intermediate and the lowest **c**. Cluster **a**, the biggest (41 samples), contains all but two of the tumors of the luminal-like (blue) cluster of PAL (see their Figure 3). More interestingly, clusters **a** and **b**, contain 45 out of 48 of the ER+ tumors (see blue leaves). Cluster **c** is rich (seven out of 11) in ER- tumors. Designating as $ER+$ the samples in $NOT(\mathbf{c})$ (i.e. that *do not* belong to **c**), we get our best classifier, with efficiency (defined as the fraction of ER+ 'caught' in $NOT(\mathbf{c})$) $E = 45/48 = 0.94$ and purity (defined as the fraction of ER+ among members of $NOT(\mathbf{c})$) $P = 0.83$. The corresponding numbers obtained by PAL (for their 'luminal-like' cluster) were $E = 0.66$ and $P = 0.89$.

$S1(G30)$: $G30$ is a cluster of 15 genes, related to cell cycle proliferation. Only one of the 15 were included in PAL's intrinsic set. Clustering the 65 tumors using the expression levels of these genes generated the dendrogram presented in Figure 7A (see supplementary information). The boxes that represent sample clusters are colored according to their relative content of ER- samples. The




**Figs 1–4, Breast cancer**. **Fig. 1.** *S*1(*G*4): clustering 65 tumors using the expression levels of gene cluster *G*4. (**A**) The boxes in the dendrogram represent clusters; they are colored according to their percentage of ER-tumors (see color bar on left). (**B**) Clusters **a,b,c** are characterized, respectively, by high, intermediate and low expression levels (see color bar on right). **Fig. 2.** *S*1(*G*10): clustering 84 breast cancer samples according to the expression levels of gene cluster G10. The boxes in the dendrogram (**A**) represent sample clusters. They are colored according to the median value of the survival of the patients contained in each cluster, ranging from dark red (median survival of 100 months) to blue (median of 4 months)—see left color bar. (**B**) Clusters **a** and **b** exhibit high and low expression levels (see color bar at right), respectively. The central color bar represents p53 status: red—mutant, blue—wt and grey—unknown. Members of **b** are characterized by low expression, low survival and mutant p53. **Fig. 3.** *S*1(*G*33): the boxes in the dendrogram (**A**) represent sample clusters that are colored according to the median value of the survival of the patients contained in each cluster, ranging from dark red (median survival of 100 months) to blue (median—4 months)—see left color bar. (**B**) The clusters **a**, **b** and **c** exhibit high, intermediate and low expression levels (see color bar at right). The central color bar represents p53 status: red—mutant, blue—wt and white—unknown. Members of **a** are characterized by high expression, low survival and mutant p53. **Fig. 4.** *S*1(*G*36): (**A**) The genes of *G*36 gave rise to a very clear partition of the breast cancer samples to high (cluster **a**) and low expression levels. (**B**) No clinical interpretation of this partition has been found yet.

dendrogram exhibits a clear partition of the tumors into clusters **a** with high expression levels of the $G30$ genes and **c** with intermediate expression levels, as seen in Figure 7B. Cluster **c** contains 44 tumors, 38 of which were classified as ER+, three as ER- and three unknown. Hence this cluster captured the ER+ group with efficiency of $E = 38/48 = 0.79$ and purity $P = 38/44 = 0.86$. Cluster **a** contains a high proportion of ER- tumors; its sub-cluster **b** consists of five special ER+ tumors that have relatively high expression levels of the $G30$ genes.

### Breast cancer—SAL

Again we have two kinds of observations; those made using genes that were not included by SAL in their intrinsic set, and hence could not have been found by them, and observations made using genes that were included in the previous analysis.

Since there is considerable overlap between the samples of PAL and SAL, we did not repeat our attempt to reproduce all their findings. We did, however, study some aspects related to the clinical labels, that were the main additional feature of the SAL data. We emphasize here our findings concerning survival and p53 status. We found correlations between expression levels of several gene clusters and survival, and that the expression levels of these genes is also a predictor of p53 mutation status. We also present a very clear partition of the patients into two groups, for which we do not yet have any clinical interpretation.

$S1(G10)$: Cluster $G10$ contains 15 genes that are related to the ER pathway, including five of the 10 members of $G4$ mentioned in our analysis of PAL, (such as GATA-binding protein three). Clustering the 85 samples ($S1$) using $G10$, generates the dendrogram presented in Figure 2A. The boxes that represent sample clusters are colored according to the median value of the survival of the patients contained in each cluster, ranging from red (median survival of 100 months) to dark blue (4 months). Similarly to the results shown in Figure 1, the variation in the expression levels of the $G10$ genes correlates well with the direct clinical measurements of the ER protein levels in the tumors. The dendrogram of Figure 2A exhibits two main clusters; **a** contains most of the ER+ tumors, that exhibit higher expression levels of the $G10$ genes, as seen in Figure 2B, and **b**, which contains mainly ER-tumors that exhibit low expression levels of the $G10$ genes.

Analyzing the correlation with the p53 status, wild type (wt) vs mutant, and with the survival parameter we get similar results as were obtained by SAL. They showed that the basal-like samples, corresponding to our cluster **b**, come from patients with the shortest survival times and a high frequency of p53 mutations. Two of the 17 members of cluster **b** survived for 41 months and all the others—for less than 26 months. The correlation coefficient between survival and the average expression levels of the $G10$ genes is **0.47**. The Wilcoxon rank-sum test (WRST) indicated that the distributions of survival times of patients in cluster **b** and of the rest of the patients are significantly different ($P$-value $= 3.7 \cdot 10^{-4}$); patients that exhibit low expression levels of the $G10$ genes have short survival.

To indicate the p53 status, we placed a color bar next to the leaves of the dendrogram, on which the patients with mutant p53 are labeled red and the p53 wt—blue. Patients with unknown p53 status were labeled white. Note that the 17 patients of cluster **b** exhibit low expression levels of the $G10$ genes. Ten of these 17 are p53 mutant, five have unknown labels and only two are wt. Hence low expression levels of the $G10$ genes seem to go along with a mutated p53. The correlation coefficient of the average expression levels of $G10$ with p53 status is **0.4**; in particular, low expression is a good predictor of mutant p53. To substantiate the last statement, we compared the distributions (using WRST) of the median expression levels of patients with mutant p53 to wt. We found that the two distributions are significantly different ($P$-value$= 1.2 \cdot 10^{-4}$); the wt p53 patients exhibit high expression levels and the mutant p53 exhibit lower expression levels of the $G10$ genes.

$S1(G33)$: Cluster $G33$ contains 36 genes, related to cell proliferation, which include 10 out of the 15 members of cluster $G30$ found by CTWC in our analysis of the PAL data. Clustering the 85 samples using the expression levels of these genes generated the dendrogram presented in Figure 3A. The boxes are colored similarly to Figure 1; according to the median survival (in months), of the patients that belong to each cluster. The $G30$ genes partition the samples into three main clusters, **a, b** and **c**, as shown in the dendrogram. The corresponding $G33$ expression levels, as seen in Figure 3B, are of high, intermediate and low levels, respectively. The average expression level of the $G30$ genes is inversely correlated with survival (correlation coefficient $-$**0.24**). Cluster **a** contains patients with high expression and short survival; only one of its 21 members survived beyond 43 months, whereas clusters **b** and **c** contain long (up to 100 months) as well as short survival. Comparison of the distributions of the survival times of the patients in cluster **a** to those in clusters **b** and **c** indicates that there is a significant difference ($P$-value $= 0.0016$).

As to p53 status, we note that among the 21 patients in cluster **a**, 13 were mutant p53 and four had unknown status. Cluster **c**, of low expression levels, contains only two mutant p53 patients (out of 16 members of the cluster). The correlation coefficient between the average expression levels of $G33$ genes and p53 status is $-$**0.4**. Hence high expression levels of these genes is a good predictor for mutant p53, whereas low expression predicts

wt p53. Comparison of the distributions of the median expression levels between the p53–mutant and the p53–wt patients yields significantly different distributions ($P$-value $= 4.5 \cdot 10^{-5}$).

$S1(G36)$: Cluster $G36$ contains genes that are related to apoptosis suppression (e.g. bcl-2) and cell growth inhibition (e.g. INK4C cyclin-dependent kinase inhibitor 2c). Using the expression levels of this set of genes to cluster the 85 samples, we generate the dendrogram presented in Figure 4A. The boxes are colored similarly to Figure 3A, according to the median survival of the patients in each cluster. The dendrogram exhibits partition of the samples into two very distinct clusters; **a** contains patients with high expression levels and **b**—patients with low. We found no correlation between membership in either of these clusters and any of the clinical labels that were reported by SAL. However, the clarity of the partition calls for further investigation of the two groups of patients, which may reveal some so far unknown role played by the genes of $G36$ in breast cancer.

### Colon cancer

We applied CTWC to the colon data set of Notterman *et al.*, containing 18 paired carcinoma and four paired adenoma samples. We refer to the set of all 44 samples as S and to the 36 paired carcinoma samples as S1. We present gene clusters which differentiate the samples according to the known normal/tumor classification, previously shown by Notterman *et al.*. Furthermore, we show the advantage of CTWC in mining new partitions which have not been found using other clustering methods and may contain relevant biological information.

*Tumor—Normal separation.* $S(G8)$: $G8$ contains 55 genes, which show high expression levels in the normal samples compared to the adenoma and carcinoma. Several genes within this cluster are known to be repressed in colorectal neoplasms; for example, guanilyn and DRA (down-regulated in adenoma). Some of these genes were previously mentioned by Notterman *et al.*. Clustering the 44 samples, using the expression levels of $G8$, generated the dendrogram shown in Figure 5A.

The dendrogram exhibits a clear separation into two large clusters (**a** and **b**) and two small ones (**c** and **d**). Clusters **c** and **d** contain all the normal samples (both carcinoma and adenoma), **a**—the tumor carcinoma samples and **b**—the tumor adenoma samples. The colors (see bar on the right-hand side of the expression matrix—see reordered data) represent the expression levels of the genes in $G8$, with red (blue) denoting high (low) values.

$S1(G25)$: The data set we analyzed next contains the 18 carcinoma and their paired normal samples, $S1$. The group $G25$ contains 51 genes, some of which are known to be over expressed in carcinoma and are found to be related to colon cancer or other forms of neoplasma e.g. myc, matrilysin, GRO-$\gamma$ (see Notterman *et al.*, 2001), and additional genes which may very well be related to colon cancer. Clustering the 36 samples of $S1$, using the expression levels of the gene cluster $G25$, gave rise to a clear partition of the samples into two clusters; one of normal samples (**a**), and the other of tumor samples (**b**), with relatively high expression levels of the $G25$ genes in the tumor cluster (see Figure 8, supplementary information).

*New observations (protocols A,B).* $S1(G3)$: Two experimental protocols that were used; 16 RNA samples (paired samples 3–6,8–10,11) were extracted using a method that isolates mRNA prior to reverse transcription ('protocol A'), and the other 20 samples (paired samples 12,27,28–29,32–35,39–40) were prepared by extracting total RNA from the cells ('protocol B'). Clustering the 36 carcinoma samples, using the expression levels of the 27 genes of cluster $G3$, exhibits a clear partition of the samples into two clusters (see Figure 6A). Cluster **b** contains 20 tissues of protocol *B*, and cluster **a** contains 14 tissues of protocol *A*. This separation has two mistakes; both samples of patient 9 were labeled *A* and appear in the cluster of protocol *B*.

*New observations (unknown interpretation).* $S10(G24)$, $S10(G7)$, $S10(G12)$: Clustering only the 18 carcinoma samples ($S10$, obtained in a previous CTWC iteration) on the basis of their expression over different sets of genes, revealed the following partitions:

The clustering operation $S10(G24)$ generated a clear separation of the tumor samples into two clusters. Samples 33,34,35,40 are clustered together in **b**, and show high expression levels of the $G24$ genes (Fig. 9, supplementary information).

The operation $S10(G7)$ separated tumor samples 27,32,33,40 from the other 14; the small group has low expression levels of the $G7$ genes (Figure 10, supplementary information).

$S10(G12)$ clustered tumor samples 33,34,35,12,40 together (cluster **b** in Figure 11, supplementary information); the expression levels of the $G12$ genes are high in these 5 samples. Hence we discovered that tumor samples 33,40 and 35 were repeatedly separated from the remaining tumors, which implies that these patients may share some common characteristics, perhaps representing a true biological meaning. However, due to lack of additional information about the patients we were unable to determine the biological origin of this separation.

## DISCUSSION AND CONCLUSION

We described the *Coupled Two Way Clustering* method and demonstrated its ability to extract useful information

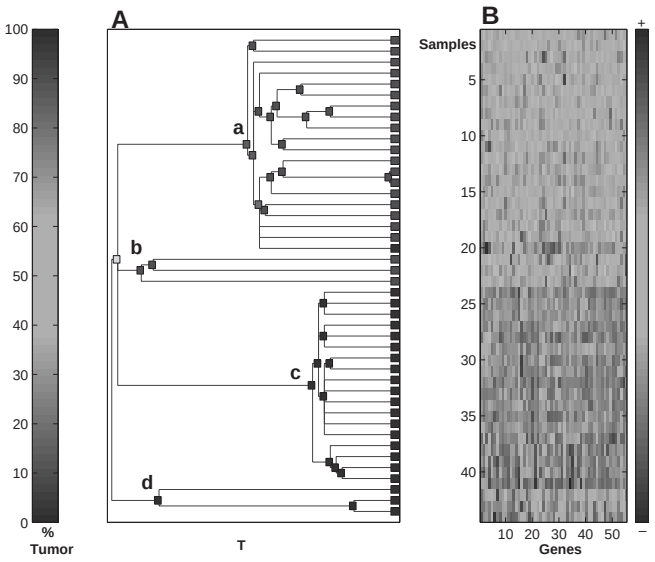



**Fig. 5.** *S*(*G*8): A clear separation of the tumor carcinoma and adenoma samples from the normal samples, using the *G*8 group of genes. (**A**) The boxes are colored (see supplementary information) according to the percentage of the tumor samples. (**B**) The expression level matrix of *S*1(*G*8). Rows correspond to all the samples and the columns correspond to the genes of cluster *G*8. The matrix shows relatively high expression levels of the *G*8 genes in the normal samples compared to the tumor samples.

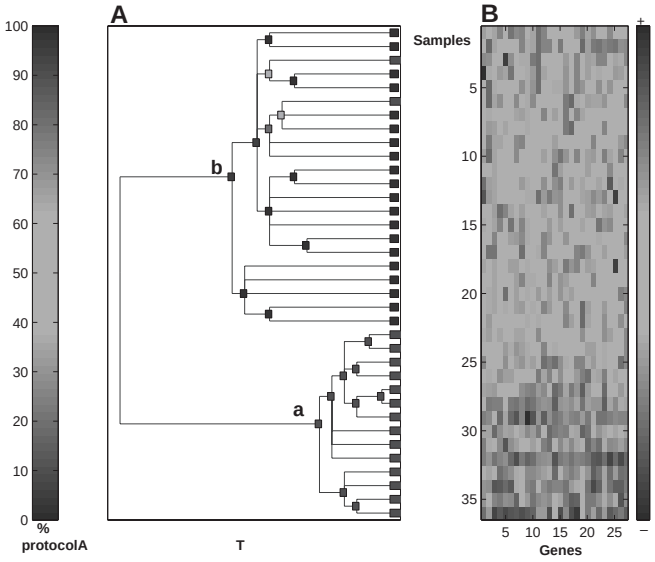



**Fig. 6.** *S*1(*G*3): Separation of the colon cancer samples according to protocols *A* and *B*. (**A**) The boxes are colored (see supplementary information) according to the percentage of protocol *A* samples (indicated by red). (**B**) The expression level matrix of S1(G3). Rows correspond to all the samples and the columns correspond to the genes of cluster G3.

from breast cancer and colon cancer data. For both data sets we reproduced the findings of previous analyses and discovered new structure of biological significance, demonstrating the advantages of CTWC compared to standard clustering techniques.

The central strategy of CTWC is to cluster the samples on the basis of their expression levels over small, correlated sets of genes, and vice versa. The relevant sets of genes and samples are found by using, one at a time, stable clusters of genes (or samples), that were identified in preceding iterations of the algorithm. Whenever such a clustering operation generates new, statistically significant partitions of the clustered objects, the result is recorded, to be used in further iterations and to be scanned for possible biological or clinical interpretation.

Perou *et al.* also reached the conclusion that performing an 'all against all' analysis does not reveal the effects of relatively small groups of relevant genes. They were able to produce significant findings only after reduction of the genes used to a smaller number. The smaller 'intrinsic set' was identified using a particular guiding principle, one that can be used only when there are at least two samples from each of several patients. Furthermore, the selection criteria used exclude genes that, according to our findings, do contain important information.

CTWC does not only generate the important partitions of the samples; it also identifies small groups of genes that are responsible for the separation of different classes. For both breast and colon cancer we found partitions that have no clear interpretation at the moment, a fact that demonstrates the strength of unsupervised approaches such as clustering; unsuspected structure buried in the data can be revealed.

## ACKNOWLEDGEMENTS

This research was partially supported by grants from the Germany—Israel Science Foundation (GIF), the Israel Academy of Sciences (ISF), the NIH under grant no. #5 P01 CA 65930-06 and the Ridgefield Foundation. We thank D. Botstein for directing us to the two papers of the Stanford group on breast cancer (PAL and SAL).

## REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Blatt,M., Wiseman,S. and Domany,E. (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.*, **76**, 3251–3254.

Blatt,M., Wiseman,S. and Domany,E. (1997) Data Clustering using a model granular magnet. *Neural Comp.*, **9**, 1805–1842.

Califano,A., Stolovitsky,G. and Tu,Y. (2000) Analysis of gene expression microarrays for phenotype classification. *Proc. Int.*

*Conf. Intell. Syst. Mol. Biol.*, **8**, 75–85.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Domany,E., Blatt,M., Gdalyahu,Y. and Weinshall,D. (1999) Superparamagnetic clustering of data: application to computer vision. *Comp. Phys. Comm.*, **121–122**, 5–12.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fontemaggi,G., Kela,I., Amariglio,N., Rechavi,G., Krishnamurthy,J., Strano,S., Sacchi,A., Givol,D. and Blandino,G. (2002) Identification of direct p73 target genes combining DNA microarray and chromatin immunoprecipitation analyses. *J. Biol. Chem.*, **277**, 43359–43368.

Geisler,S., Lonning,P.E., Aas,T., Johnsen,H., Fluge,O., Haugen,D.F., Lillehaug,J.R., Akslen,L.A. and Borresen-Dale,A.L. (2001) Influence of TP53 gene alterations and c-erbB-2 expression on the response to treatment with doxorubicin in locally advanced breast cancer. *Cancer Res.*, **6**, 2505–2512.

Getz,G., Levine,E. and Domany,E. (2000a) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Getz,G., Levine,E., Domany,E. and Zhang,M.Q. (2000b) Superparamagnetic clustering of yeast gene expression profiles. *Physica A*, **279**, 457–464.

Getz,G. and Domany,E. (2003) Coupled two-way clustering server. *Bioinformatics*, **19**, 1153–1154.

Godard,S., Getz,G., Kobayashi,H., Farmer,P., Delorenzi,M., Nozaki,M., Diserens,A.-C., Hamou,M.-F., Dietrich,P.-Y., Villemure,J.-G. *et al.* Taxonomy and Classification of Human Astrocytic Gliomas on the basisof gene expression, submitted.

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. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **5439**, 531–537.

Ihmels,J., Friedlander,G., Bergmann,S., Sarig,O., Ziv,Y. and Barkai,N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370.

Kannan,K., Amariglio,N., Rechavi,G., Jakob-Hirsch,J., Kela,I., Kaminski,N., Getz,G., Domany,E. and Givol,D. (2001) DNA microarrays identification of primary and secondary target genes regulated by p53. *Oncogene*, **20**, 2225–2234.

Kela,I. (2002) Clustering of gene expression data, M.Sc. Thesis, Weizmann Institute.

Notterman,D.A., Alon,U., Sierk,A.J. and Levine,A.J. (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.*, **7**, 3124–3130.

Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Quintana,F., Getz,G., Hed,G., Domany,E. and Cohen,I.R. (2003) Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity, in press.

Sorlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **19**, 10869–10874.

Tanay,A., Sharan,R. and Shamir,R. (2002) Biclustering gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (in print).

# Publication 8:

## Classification of human astrocytic gliomas on the basis of gene expression: A correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes

Authors: S. Godard, G. Getz, M. Delorenzi, P. Farmer, H. Kobayashi, I. Desbaillets, M. Nozaki, A-C. Diserens, M-F. Hamou, P-Y. Dietrich, L. Regli, R.C. Janzer, P. Bucher, R. Stupp, N. de Tribolet, E. Domany and M.E. Hegi
Published in: *Cancer Research* **63**, 6613–6625 (2003).

[CANCER RESEARCH 63, 6613–6625, October 15, 2003]

# Classification of Human Astrocytic Gliomas on the Basis of Gene Expression: A Correlated Group of Genes with Angiogenic Activity Emerges As a Strong Predictor of Subtypes[1,2]

**Sophie Godard, Gad Getz, Mauro Delorenzi, Pierre Farmer, Hiroyuki Kobayashi, Isabelle Desbaillets, Michimasa Nozaki, Annie-Claire Diserens, Marie-France Hamou, Pierre-Yves Dietrich, Luca Regli, Robert C. Janzer, Philipp Bucher, Roger Stupp, Nicolas de Tribolet, Eytan Domany, and Monika E. Hegi[3]**

*Laboratory of Tumor Biology and Genetics [S. G., H. K., I. D., M. N., A-C. D., M-F. H., N. d. T., M. E. H.] of the Department of Neurosurgery [L. R., N. d. T.], Multidisciplinary Oncology Center [R. S.], and Division of Neuropathology [R. C. J.], University Hospital (CHUV), 1011 Lausanne, Switzerland; Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel [G. G., E. D.]; Swiss Institute of Bioinformatics [M. D., P. F., P. B.]; and NCCR Molecular Oncology [M. D., P. F., M. E. H.], Swiss Institute for Experimental Cancer Research, 1066 Epalinges, Switzerland; and Department of Oncology, Hôpital Universitaire Genève, Switzerland [P-Y. D.]*

## ABSTRACT

**The development of targeted treatment strategies adapted to individual patients requires identification of the different tumor classes according to their biology and prognosis. We focus here on the molecular aspects underlying these differences, in terms of sets of genes that control pathogenesis of the different subtypes of astrocytic glioma. By performing cDNA-array analysis of 53 patient biopsies, comprising low-grade astrocytoma, secondary glioblastoma (respective recurrent high-grade tumors), and newly diagnosed primary glioblastoma, we demonstrate that human gliomas can be differentiated according to their gene expression. We found that low-grade astrocytoma have the most specific and similar expression profiles, whereas primary glioblastoma exhibit much larger variation between tumors. Secondary glioblastoma display features of both other groups. We identified several sets of genes with relatively highly correlated expression within groups that: (*a*) can be associated with specific biological functions; and (*b*) effectively differentiate tumor class. One prominent gene cluster discriminating primary *versus* nonprimary glioblastoma comprises mostly genes involved in angiogenesis, including *VEGF fms-related tyrosine kinase* 1 but also *IGFBP2*, that has not yet been directly linked to angiogenesis. *In situ* hybridization demonstrating coexpression of *IGFBP2* and *VEGF* in pseudopalisading cells surrounding tumor necrosis provided further evidence for a possible involvement of *IGFBP2* in angiogenesis. The separating groups of genes were found by the unsupervised coupled two-way clustering method, and their classification power was validated by a supervised construction of a nearly perfect glioma classifier.**

## INTRODUCTION

Because of their diffusely infiltrating behavior, LGA[4] (WHO grade II) cannot be resected completely and will usually recur. At relapse, progression to anaplastic astrocytoma (WHO grade III) or glioblastoma multiforme, the most malignant form of gliomas (WHO grade IV), is common. Most glioblastoma arise *de novo* without evidence of a less malignant precursor lesion and are termed PrGBM. Glioblastoma evolving from a previous lower grade astrocytoma are defined as ScGBM. Although PrGBM are indistinguishable from ScGBM by histology, the two types of tumors exhibit distinct genetic alterations and occur in different age groups. The mean age for PrGBM is ~55 years, whereas ScGBM typically occur in younger patients (<45 years). Thus, PrGBM and ScGBM can be considered as two different diseases (1), despite a similarly grim outcome with a median survival of <1 year after diagnosis and no effective therapy. PrGBM are characterized by amplification/rearrangement and overexpression of the EGFR gene (in 40 and 60% of the patients, respectively) often in association with deletion of the *INK4a/p14ARF* gene locus (2). The hallmarks for ScGBM are *TP53* mutations (60%) and overexpression of *PDGF* and *PDGF* receptor (3, 4). Development of mouse glioma models and developmental neurobiology have allowed for recent advances in the understanding of the molecular bases of these two distinct genetic pathways and their implication for tumor initiation and progression as well as the cell of origin (5). However, a substantial number of glial tumors cannot be characterized by either of the two pathways depicted above, suggesting additional not yet recognized pathogenetic pathways. The current knowledge of tumor genetics does not allow identifying clinically relevant factors predictive for outcome or response to therapy. More detailed knowledge of underlying mechanisms and their relevance for the cancer process will allow treating cancer specifically by targeting deregulated pathways, leading to rational design of future treatment modalities tailored according to the biology of the individual tumors (6).

An essential initial step toward this goal is the establishment of a taxonomy of tumors on the basis of their gene expression profiles. A search for alternative pathways must be based on identification of genes whose expression differs significantly between the various tumor classes. In particular, it is important to look for a group of (possibly) coregulated genes, some of which share some known biological function, and whose expression differentiates tumor classes. Identification of such groups leads to better understanding of the biological processes that underlie the distinction between the tumors and may provide clues for the roles of such genes in initiation and progression of cancer.

We aimed at identifying expression profiles that differentiate three groups of astrocytic glioma: (*a*) LGA; (*b*) their respective ScGBM; and (*c*) PrGBM. In a novel gene selection approach, we combined supervised statistical analysis with CTWC (7–9), an unsupervised method, to identify correlated groups of genes that distinguish between the various tumor subtypes. Here, we demonstrate that gliomas can be separated according to their gene expression profiles, with PrGBM exhibiting a much higher variation of expression profiles than LGA. A cluster of correlated genes was identified that separates PrGBM from the other tumors and contains genes that are known to

Received 2/22/03; revised 6/27/03; accepted 7/24/03.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by grants of the Swiss National Science Foundation and ONCOSUISSE (to N. d. T. and M. E. H.), the National Center of Competence in Research Molecular Oncology (to M. D., P. F., and M. E. H.), the Association des neuro-oncologues d'expression française (to S. G.), the Germany-Israel Science Foundation, the Israel Science Foundation, Minerva, and the Ridgefield Foundation (to G. G. and E. D.).

[2] Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org).

[3] To whom requests for reprints should be addressed, at Laboratory of Tumor Biology and Genetics, Department of Neurosurgery, Centre Hospitalier Universitaire Vaudois (CHUV), BH19-110, 1011 Lausanne. Phone: 41-(21) 314 2582; Fax: 41-(21) 314 2587; E-mail: monika.hegi@chuv.hospvd.ch.

[4] The abbreviations are: LGA, low-grade astrocytoma; PrGBM, primary glioblastoma; ScGBM, secondary glioblastoma; TA, tissue-array; PDGF, platelet-derived growth factor; CTWC, coupled two-way clustering; OAIII, anaplastic oligoastrocytoma; TNR, tenascin R; MDS, multidimensional scaling; EGFR, epidermal growth factor receptor; FDR, false discovery rate; VEGF, vascular endothelial growth factor; FLT1, fms-related tyrosine kinase 1; PTN, pleiotrophin; IGFBP2, insulin-like growth factor-binding protein 2; IGF-1, insulin-like growth factor-1; TGF, transforming growth factor; FGF2, basic fibroblast growth factor.

be involved in angiogenesis, such as *VEGF*, but also *IGFBP2*, whose implication in angiogenesis is novel. By using expression data of the most informative separating gene clusters, we were able to construct an almost perfect tumor classifier. Our main findings, based on cDNA arrays, were validated on an independent set of glioma and by other methods, respectively.

## MATERIALS AND METHODS

### Specimen and RNA Expression Analysis

**Tumor Biopsies and Cell Lines.** Tumor biopsies, obtained from patients who underwent surgery at University Hospitals in Lausanne or Geneva or the Cantonal Hospital in Fribourg (Switzerland), were shock frozen and stored at −70°C. The use of biopsies and respective clinical data have been approved by the local ethics committee and the respective federal agency. The tumors were diagnosed according to the WHO classification 2000 (1). Twenty-one biopsies originated from 20 patients enrolled in a prospective pilot trial for newly diagnosed glioblastoma (10). From the tumor bank, an additional 32 gliomas from 22 patients were included, comprising 24 LGA and 8 respective high-grade recurrent tumors (comprising two astrocytoma WHO grade III and 6 grade IV). These tumors have been analyzed previously for TP53 mutations (11). The tumor samples were organized into two data sets according to their date of analysis: (*a*) subsequently used as a training set, comprising 14 PrGBM, 5 ScGBM, and 12 LGA; and (*b*) used as validation set, 4 PrGBM, 4 ScGBM, and 12 LGA. A summary of the information on the individual tumors, including TP53 status, age, and gender of the patients, and their organization into data sets, is available in Table 1. Normal brain tissue for RNA isolation was obtained from a lobectomy after brain edema, and additional samples of human normal brain total RNA were obtained from Clontech (human total RNA panel IV).

TAs have been constructed from archived paraffin blocs at the University Hospital in Lausanne (1984–2000) as described (12). All cases have been reviewed according to the WHO classification 2000 (1) by the neuropathologist (R. C. J.). The GBM-TA comprises 190 GBM, and the non-GBM glioma-TA includes 158 gliomas WHO I–III of different glioma subtypes.

**Cell Lines.** Culture conditions for the glioblastoma cell line U87, LN229, LNZ308, and the TP53-inducible glioblastoma cell line 2024, which is derived from LNZ308 after introducing the Tet-On System, have been described before (13). Expression of wild-type TP53 was induced with 2 $\mu$g/ml doxocycline in cell line 2024 during 24 h before RNA isolation (2024+). For anoxia treatment, cells were cultured for 20 h in an anaerobic culture incubator ($O_2$ <1%; Scholzen, Microbiology Syst AG) filled with a mixture of $N_2$, $H_2$, and $CO_2$.

**Isolation of Total RNA.** Before RNA isolation, a section of the frozen tumor biopsy was reevaluated after H&E staining by the neuropathologist (R. C. J.) to estimate the proportion of solid tumor, contaminating normal tissue, or infiltration zone. Pieces comprising >30% normal tissue were excluded. The setting of thresholds on the tumor fraction is treated differently by various groups, an issue that has been reviewed recently by Ramaswamy and Golub (14), who suggest to use "tumor cell-enriched material." Fifty to 100 mg of frozen tumor tissue were homogenized in TRIzol solution (Life Technologies). RNA phase was purified in saturated phenol solution [60% phenol, 15% glycerol, and 0.1 M sodium acetate (pH 4)]. The quality of the RNA was evaluated on agarose gel. RNA from cell lines was isolated similarly.

***In Situ* Hybridization and Northern Blot Analysis.** *In situ* hybridization was performed according to the protocol supplied by Roche (Roche Applied Science) using cRNA probes labeled with digoxigenin during the *in vitro* transcription reaction. The plasmid pcDNA3 containing the 1400-bp human *IGFBP2* cDNA (a kind gift from S. Babajko; Ref. 15) was linearized with *Eco*RI or *Hind*III and transcribed with T7 or Sp6 RNA polymerase, respectively, to obtain sense or antisense probes. The plasmid pBluescript-KS-M13+ containing the 650-bp human *VEGF165* cDNA (pBspt-KS-VEGF165, a kind gift from K. Plate; Ref. 16) was linearized with *Eco*RI or *Bam*HI and transcribed with T7 or T3 RNA polymerase, respectively, to obtain sense and antisense probes. Probes were further reduced to an average size of 100 bp by limited alkaline hydrolysis. Northern blot analysis was performed as described before using 10 $\mu$g of total RNA (13). The membrane was sequentially hybridized to the plasmid-derived probe for *VEGF165* (*Eco*RI/*Bam*HI-fragment of pBspt-KS-VEGF165) and the PCR-derived probes for *IGFBP2, 3,* and *5* (respective primer sequences provided by Clontech). Probes were radioactively labeled using the random primed DNA labeling kit (Boehringer) using [$\alpha$-$^{32}$P]dCTP (3000 Ci/mmol, Amersham). Expression was quantified by phosphorimager (Fuji, BAS 1000).

**cDNA Synthesis and Hybridization on cDNA Array.** Atlas Human Cancer 1.2 Array membranes (Clontech) were used for all experiments described. These nylon filters are spotted with 1185 genes, including reference genes, and 1176 genes related to cancer. Three to 5 $\mu$g of DNase-treated total RNA were used to prepare a labeled first strand cDNA using the Clontech kit, basically as recommended. Briefly, RNA in a volume of 2 $\mu$l was mixed with 1 $\mu$l of specific CDS primers and 1 $\mu$l of RNasin (Promega; 40 units/ml) and denatured at 70°C. Subsequently, 1 $\mu$l of Superscript II reverse transcriptase (Life Technologies, Inc.; 200 units/ml) and 3.5–5 $\mu$l of $\alpha$-$^{32}$P-dATP (3000 Ci/mmol; Amersham) were added, and the reaction was performed at 48°C for 30 min. The probe was purified with Clontech Atlas Nucleospin extraction kit following the manufacturer's recommendations. Before use, the membranes were boiled in 0.5% SDS solution for stripping and also at first use. Membranes were exposed to an imaging plate (BAS MS 2040; Fuji) for 1–8 days. The Atlas Cancer Arrays were used three times.

**TP53-Mutation Analysis.** The tumors were screened for *TP53* mutations using the yeast functional assay as described (17, 18), followed by direct sequencing, if the test was positive (Microsynth, Balgach, Switzerland).

**Immunohistochemistry.** Immunohistochemical determination for EGFR (Novo Castra; NCL-EGFR; dilution 1:40) and tenascin R [Santa Cruz Biotechnology; Tenascin-R (N20) sc-9874; dilution 1:1000] on paraffin sections was performed according to standard procedures using a high temperature epitope retrieval technique in citrate buffer (pH 6.0; pressure cooker, 3 min). Semiquantitative evaluation was performed independently by two researchers.

### Data Analysis

**Preprocessing and Analysis of Expression Data.** Expression was quantified by phosphorimager (Fuji BAS1000) and analyzed with Atlasimage 1.5 software (Clontech). After background subtraction signals were normalized using the "sum method" comparing the sum of the intensity of all genes in the experiment to the sum calculated from the reference sample, yielding the coefficient of normalization, "c." The reference used in all experiments represents the calculated average of five independent expression profiles derived from "normal" brain. The ratio (R) was calculated for every gene as follows: $r = [(s - s0) \times c + K]/[(r - r0) + K]$, where $K = b \times [s0 + (c \times r0)]$; s, gene intensity in sample; s0, background in sample; c, coefficient of normalization obtained using the sum method; r, gene intensity in reference experiment; r0, background in reference experiment; and $b = 2$. For $s \gg s0$ and $r \gg r0$, this formula generates normalized ratios R shrunk toward 1 for low expressed genes in both experiments, sample and reference, respectively. Furthermore, it circumvents the problem of losing valuable information if the experiment or the reference display no expression for a given gene (division by 0).

**Distance and MDS.** This method projects the data points from the high dimension in which it is embedded to a low (two or three) dimensional space, in a way that best preserves their relative distances. Euclidian distances in the space of logarithms of the ratios R were used. MDS was performed with the implementation for the classical metric scaling (also known as principal coordinate analysis; Ref. 19) available in the mva package for R available online at the Comprehensive R Archive Network.[5]

**CTWC.** A variation filter was applied; only those genes were kept, for which the ratio of the maximal and minimal R values (obtained for the 36 experiments) exceeded 2. For each gene (row), the log of the ratio R was mean-centered (subtracting the average) and normalized. Euclidean distances, measured between all pairs of genes and between all pairs of tumors, served as the input to our clustering procedure. CTWC has been described elsewhere; see Getz *et al.* (7) for full details and comparisons with other methods; (8) for its applications to leukemia, colon, and breast cancer data analysis; and (9)

[5] The URLs referred to are: The Comprehensive R Archive Network: http://www.R-project.org; CTWC-Server: http://ctwc.weizmann.ac.il; homepage of complete CTWC data analysis: http://www.hospvd.ch/itbg; GeneCards, encyclopedia for genes, proteins, and diseases: http://genecards.weizmann.ac.il.

Table 1 *Summary information on 56 experiments analyzed by gene expression profiling*

This set comprises 53 gliomas from 44 patients and three experiments with cell lines.

| | | | | TP53 status | | | Data analysis[g] | | |
|---|---|---|---|---|---|---|---|---|---|
| Sample[a] | Pathology[b] | Gender | Age[c] | | Codon | TP53 mutation | CTWC | TRN-Set | VAL-Set |
| 1284 | PrGBM | F | 48 | wt | | | 1 | 1 | 0 |
| 1437 | PrGBM | M | 48 | wt | | | 1 | 1 | 0 |
| 1316 | PrGBM | M | 68 | wt | | | 1 | 1 | 0 |
| 1399 | PrGBM | M | 38 | wt | | | 1 | 1 | 0 |
| G204 | PrGBM | F | 51 | wt | | | 1 | 1 | 0 |
| 1430 | PrGBM | M | 37 | wt | | | 1 | 1 | 0 |
| G197 | PrGBM | M | 46 | wt | | | 1 | 1 | 0 |
| 1419 | PrGBM | F | 48 | wt | | | 1 | 1 | 0 |
| 1308 | PrGBM | F | 45 | wt | | | 1 | 1 | 0 |
| 1453 | PrGBM | F | 53 | mut | 244 | GGC to G**T**C | 1 | 1 | 0 |
| 1317 | PrGBM | M | 26 | mut | 175 | CGC to C**A**C | 1 | 1 | 0 |
| 1297 | PrGBM | M | 36 | mut | 163 | TAC to T**G**C | 1 | 1 | 0 |
| 1303 | PrGBM | M | 55 | mut | 273 | CGC to C**A**T | 1 | 1 | 0 |
| 1360 | PrGBM | M | 65 | mut | 173 | GTG to **A**TG | 1 | 1 | 0 |
| G205 | PrGBM | F | 56 | wt | | | 0 | 0 | 1 |
| G216 | PrGBM | M | 53 | wt | | | 0 | 0 | 1 |
| 1621 | PrGBM | M | 62 | wt | | | 0 | 0 | 1 |
| G226 | PrGBM | M | 45 | mut | 273 | CGT to C**A**T | 0 | 0 | 1 |
| 1342 | ScGBM | M | 51 | wt | | | 1 | 1 | 0 |
| 749 | ScGBM | F | 47 | mut | 175 | Ref. 11[f] | 1 | 1 | 0 |
| 946 | ScGBM | M | 41 | mut | 258, 267, 283 | Ref. 11[f] | 1 | 1 | 0 |
| 809 | ScGBM | M | 53 | mut | 273 | Ref. 11[f] | 1 | 1 | 0 |
| 978 | ScGBM | M | 29 | mut | 241 | Ref. 11[f] | 1 | 1 | 0 |
| 413 | ScGBM | F | 53 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 633 | ScGBM | M | 28 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 735 | ScGBM | M | 34 | mut | 261 | Ref. 11[f] | 0 | 0 | 1 |
| 722 | ScGBM | M | 39 | mut | 248 | Ref. 11[f] | 0 | 0 | 1 |
| 421 | LGA | M | 2 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 698 | LGA | M | 39 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 1070 | LGA | M | 58 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 80 | LGA | M | 28 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 246 | LGA | F | 53 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 328 | LGA | M | 28 | wt | | Ref. 11[f] | 1 | 1 | 0 |
| 416 | LGA | M | 29 | mut | 241 | Ref. 11[f] | 1 | 1 | 0 |
| 92 | LGA | M | 34 | mut | 261 | Ref. 11[f] | 1 | 1 | 0 |
| 289 | LGA | M | 27 | mut | 248 | Ref. 11[f] | 1 | 1 | 0 |
| 460 | LGA | F | 47 | mut | 175 | Ref. 11[f] | 1 | 1 | 0 |
| 736 | LGA | M | 41 | mut | 258, 267, 283 | Ref. 11[f] | 1 | 1 | 0 |
| 635 | LGA | M | 53 | mut | 273 | Ref. 11[f] | 1 | 1 | 0 |
| 676 | LGA | M | 41 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 355 | LGA | M | 35 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 374 | LGA | M | 15 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 875 | LGA | M | 50 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 501 | LGA | F | 57 | wt | | Ref. 11[f] | 0 | 0 | 1 |
| 898 | LGA | F | 35 | mut | 248 | Ref. 11[f] | 0 | 0 | 1 |
| 528 | LGA | M | 52 | mut | 248 | Ref. 11[f] | 0 | 0 | 1 |
| 551 | LGA | F | 31 | mut | 155 | Ref. 11[f] | 0 | 0 | 1 |
| 510 | LGA | F | 35 | mut | 248 | Ref. 11[f] | 0 | 0 | 1 |
| 210 | LGA | M | 39 | mut | 248 | Ref. 11[f] | 0 | 0 | 1 |
| 589 | LGA | M | 33 | mut | 220 | Ref. 11[f] | 0 | 0 | 1 |
| 552 | LGA | M | 26 | mut | 234 | Ref. 11[f] | 0 | 0 | 1 |
| 1497 | RecGBM | M | | wt | | | 1 | 0 | 0 |
| 1357 | OAIII[d] | M | 36 | mut | 273 | CGT to **T**GT | 1 | 0 | 0 |
| 2024− | CL | | | null | | | 1 | 0 | 0 |
| 2024+ | CL | | | wt[e] | | | 1 | 0 | 0 |
| U87 | CL | | | wt | | | 1 | 0 | 0 |

[a] Tumors from same patients: PrGBM to recurrent GBM: 1430/1497; LGA to recurrent LGA: 80/328; LGA to ScGBM: 736/946, 635/809, 416/978, 460/749, 92/735, 210/722; LGA to anaplastic astrocytoma (WHO grade III): 416/978, 246/413.

[b] PrGBM (WHO grade IV); ScGBM (WHO grade IV); LGA (WHO grade II); OAIII (WHO grade III).

[c] Age at diagnosis.

[d] 1357 was originally diagnosed as PrGBM.

[e] Induction of wild-type *TP53* with Tet-On system (13).

[f] Published previously by Ishii *et al.* (11).

[g] Samples used for respective data analysis: CTWC; TRN-set, training set; VAL-set, validation set for tumor predictor.

describing the use of the publicly available CTWC–Server.[5] CTWC starts with clustering all genes on the basis of the data from all tumors and clustering all tumors, using data from all genes. In both resulting dendrograms, we identify stable (*i.e.*, statistically significant) gene and sample clusters; these are denoted, respectively, as GX or SY (where X,Y are running indices). Note that G1 represents the set of all genes and S1 that of all samples. In the second iterative step, each stable gene cluster GX is used to characterize and cluster the members of every sample cluster SY and *vice versa*.

To use CTWC, we must be able to identify stable clusters. One of the few algorithms that provide a stability index to each cluster is Super Paramagnetic Clustering, a physics-based method that has been described in full detail in Blatt *et al.* (20). We register a cluster *C* as stable only if it exceeds a certain size and when *Stab(C)*, its stability index defined in terms of the range of resolution parameters, *T*, through which cluster *C* "lives" (21), exceeds a certain threshold.

**Classification.** For binary class comparisons, two standard statistical tests were used: (*a*) the two-sample *t* test (with unknown but equal variances); and (*b*) the Wilcoxon rank-sum test. To address contamination with false positive genes associated with multiple comparisons, we use the method of Benjamini and Hochberg (22) that bounds the average FDR. The outcome of this method

Table 2 *Summary of results for two-way comparisons between tumor classes using the tumors of the training set and all genes (1185)*

| Comparison tested (no. of tumors) | No. of genes selected (FDR q = 0.05)[a] | | | | Highest *P* |
|---|---|---|---|---|---|
| | t test | Rank-sum | Shared | Union | |
| PrGBM (14) *vs.* ScGBM + LGA (5 + 12) | 191 | 174 | 160 | 205 | 0.008 |
| PrGBM + ScGBM (14 + 5) *vs.* LGA (12) | 126 | 98 | 90 | 132 | 0.005 |
| PrGBM (14) *vs.* LGA (12) | 167 | 163 | 143 | 187 | 0.007 |

[a] FDR, false discovery rate.

is a list of differentiating genes; the expected fraction of false positive genes is at most *q*.

The class discrimination power of the selected sets of genes was validated by training a k-nearest neighbor (*k*-NN) classifier on one set of tumor tissue samples and testing the class prediction on an independent validation set. Computations were performed with the class package for R and with $k = 3$. Euclidian distance in the space of the ratios R was used, classification was decided by majority vote, with ties broken at random. When ties occurred, classification was repeated 100 times, and *P*s were averaged. *P*s for the significance of the deviation from independence between true and predicted labels were obtained with Fisher's exact test (ctest package for R) and with the alternative hypothesis set to "greater" in the 2 by 2 case.

## RESULTS

### Gene Expression Profiles Separate Tumor Classes

To characterize and classify gliomas by their gene expression profiles, RNA isolated from frozen gliomas and three glioblastoma cell lines was analyzed using cDNA arrays comprising 1185 genes.

**MDS.** The configuration of the 51 astrocytic gliomas in Euclidian space of overall gene expression, as visualized in Fig. 1 using MDS, clearly suggests that gene expression profiles contain information discriminating the three classes of astrocytic gliomas. The LGA and ScGBM show a higher degree of spatial intermingling, whereas the best separation is between LGA and PrGBM. This correctly reflects other biological characteristics of the ScGBM, in that they progress from LGA but share the malignancy grade (WHO IV) with PrGBM and are indistinguishable from them histopathologically. Interestingly, the pairwise distances in space of all genes (1185) are highest among PrGBM (mean, 10.06; SD, 2.1) and shortest between LGA (mean, 6.09; SD, 1.18), whereas ScGBM are intermediate (mean, 7.27; SD, 1.18). Using 358 genes that pass a variation filter (see CTWC, below) yields a nearly similar MDS picture (data not shown). Thus, MDS of gene expression suggests more biological heterogeneity of PrGBM, without firm evidence for obvious subclasses, reflecting well the heterogeneous morphological and clinical characteristics of glioblastoma multiforme.

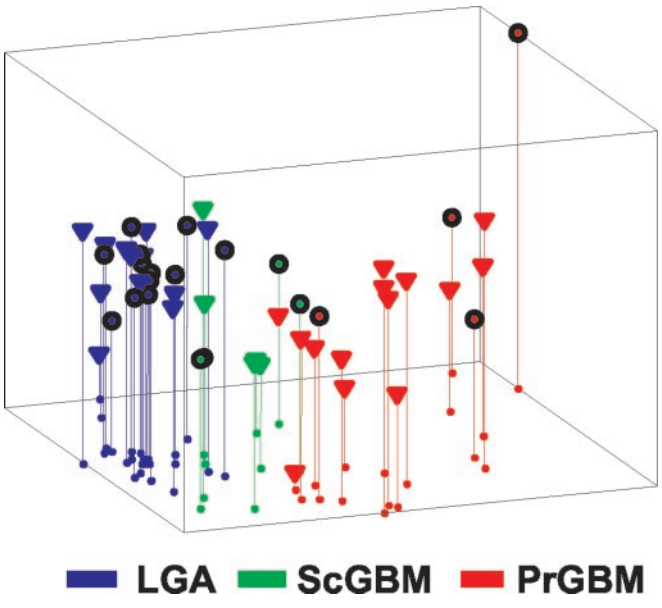



Fig. 1. MDS based on overall gene expression (1185 genes) of 51 astrocytic gliomas. The color code indicates the tumor subtype. ▼, samples from the training set; ●, gliomas from the validation set.

**Supervised Analysis.** We applied *t* test and Wilcoxon's rank-sum test to look for genes that differentiate two known classes, using all 1185 genes and the training set of 31 astrocytic gliomas. The threshold for the FDR was set at q = 0.05, *i.e.*, the expected number of false positives was kept at 5%. The results for three two-way comparisons are summarized in Table 2. The lists of differentiating genes determined by the two tests are very similar. Because of the low number of the ScGBM in the training set, comparisons involving ScGBM, ScGBM *versus* PrGBM, and ScGBM *versus* LGA yielded only very few separating genes (1 and 3, respectively). Full results of these supervised analyses, including a list of the respective genes selected by these binary comparisons, can be found in Table S1 of the supplementary information.

Our efforts to find genes correlated with *TP53* status failed, whether we used all tumors or only subsets thereof. Neither did we find any gene cluster separating groups of samples on the basis of their *TP53* status (see *TP53* status of biopsies in Table 1). Our results are in line with a previous report on a series of LGA, which also did not find evidence for genes correlated with *TP53* status (23). This result may not be surprising in consideration of the fact that the TP53 pathway has been found to be inactivated, in >70% of all astrocytic gliomas, regardless of tumor grade, by mutational alteration of *TP53* or alternative mechanisms, such as overexpression of mdm2, a negative regulator of TP53, or inactivation of p14ARF, a negative regulator of mdm2, by deletion or promoter methylation of the gene (24).

### A Cluster of Angiogenesis-related Genes Separates PrGBM from the Other Tumors

**First Level CTWC: Unsupervised Analysis.** The data derived from 36 experiments (Table 1; of 33 gliomas and three experiments with glioblastoma cell lines) were subjected to the variation filter described in "Materials and Methods," reducing the number of genes (rows) to 358. This filtering reduces the "noise" generated by genes that do not vary significantly over the samples. Changing the filtering parameters from 2 to 1.5 (or 3) changed the number of genes that passed to 601 (or 190), but the changes induced in the observed gene clusters were comparable with the fluctuations observed in repeated runs (with 358 genes). The first operation, G1(S1), clustered the set of all 358 genes (G1), using the data from the set of all 36 experiments (S1); the second, S1(G1), clustered the tumors S1 using their expression profiles measured for all of the genes G1. The two resulting dendrograms are shown in Fig. 2*A* together with the correspondingly reordered expression matrix. G1(S1) identified 15 stable gene clusters (Stability > 8, corresponding to $P < 0.01$) that are indicated in the gene dendrogram of Fig. 2*A* and denoted G2 to G16. The operation S1(G1) yields a dendrogram with two stable clusters, S2 and S3. Details

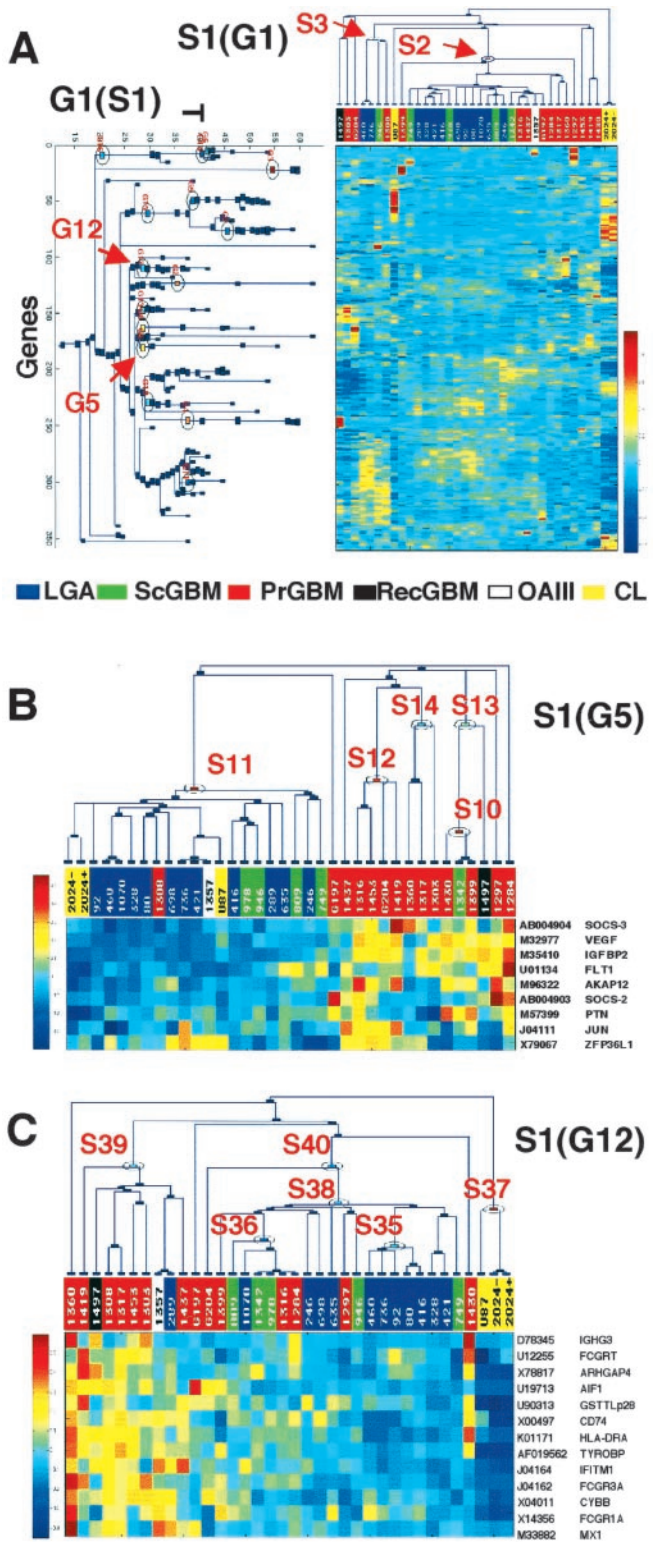



Fig. 2. Unsupervised analysis by CTWC. *A*, first level CTWC clusters G1 (the set of all 358 genes that have passed the filtering criteria), using all 36 experiments (S1), and clusters all samples (S1) according to the gene expression profiles of all genes G1. These clustering operations are called G1(S1) and S1(G1), respectively. G1(S1) yields a dendrogram of gene clusters, and S1(G1) is a dendrogram of sample clusters. The respectively reordered expression matrix is visualized using a color scale, representing centered and normalized values of the $\log_2$R. Fifteen stable gene clusters emerged and are marked with a ring (G2–G16). Clustering the samples according to all genes, S1(G1), yielded two stable sample clusters, S2 and a smaller cluster S3. The tumor samples are denoted at the *top*, using *colors* to represent tumor subtypes. *RecGBM*, recurrent glioblastoma; *OAIII*, oligoastrocytoma WHO grade III; *CL*, glioblastoma cell line. In *B*, in second level CTWC, we cluster all samples S1 according to selected gene clusters, S1(G5). Clustering S1 according to gene cluster G5 [which was obtained from G1(S1); *A*] yields a dendrogram of the clusters obtained in the CTWC can be viewed and searched online at the respective Web page.[5]

**Second Level CTWC: Clustering Tumors Using Selected Gene Clusters.** We used all of the 15 stable gene clusters found above, one by one, to recluster S1. Whenever a stable partition of the samples was found, we checked whether it divided the tumors according to some biologically relevant attribute. In particular, we looked for gene clusters that partition the tumors into GBM *versus* LGA.

**Separation into PrGBM *Versus* LGA and ScGBM on the basis of Angiogenic Activity in S1(G5).** In the first level of CTWC, we identified G5 as a stable cluster of nine genes (Fig. 2, *A* and *B*). This cluster comprises hallmarks of angiogenesis, such as *VEGF*, *FLT1* (fms-related tyrosine kinase 1; also called *VEGFR1*), and pleiotrophin (25). Some of the other genes have also been related to some aspects of angiogenesis before (Table 3), namely, *IGFBP2* (insulin-like growth factor binding protein 2), which has been suggested to be activated by hypoxia-inducible factor-1, although in an indirect manner because of the absence of a defined hypoxia-response element (26). When the expression levels of only these genes were used to characterize the tumors, a large and stable cluster, S11, of 21 tumors emerged (Fig. 2*B*). This cluster contained all of the 12 LGA and 4 of 5 ScGBM. Of the remaining 5 samples of S11, three were cell lines, whose expression profiles were consistently different. Hence, the expression levels of G5 gave rise to a nearly perfect separation of clinically defined tumor entities, namely PrGBM from LGA and ScGBM. These genes were significantly up-regulated in PrGBM (Fig. 2*B*) as compared with LGA and ScGBM.

## Other Separating Gene Clusters

**Separation by Immune Response-related Genes in S1(G12).** The gene cluster G12 contains 13 genes that are almost exclusively related to the immune system and inflammation (Table 3). Four of which are known to be regulated by γ-IFN (*FCGR1A*, *CYBB*, *IFITM1*, and *MX1*), and some of the Fc γ receptors and *DAP12* represent markers for natural killer cells, although their expression is not exclusive. The dendrogram obtained by clustering the tumors using these genes is presented in Fig. 2*C*. Of its two large stable clusters, S39 contains nine tumors, eight of which are PrGBM; the expression level of the G12 genes in these tumors is high. The other large cluster, S38, contains all but one of the LGA and ScGBM samples, together with four PrGBM. Three of these four tumors are from patients whose survival was relatively long (>18 months; Ref. 10). Hence, the immune system-related genes of G12 appear to have lower expression levels in LGA and ScGBM and in those PrGBM that have longer survival.

**Gene Selection by Combining Supervised Analysis with Clustering.** In a second step, we used the results of the supervised analysis described above to identify those clusters in the dendrogram yielded by G1(S1) that are rich in discriminating genes. Of the 205 genes that were found to differentiate PrGBM from ScGBM and LGA (Table 2; Table S1 of supplementary information), 91 passed the variation filter and were included in G1. As depicted in Fig. 3*A*, most separating genes belonged to one of four marked clusters. The same four clusters

for the samples. Note the separation of LGA and ScGBM from PrGBM. PrGBMs exhibit overexpression of the genes of G5 that are related to angiogenesis. In *C*, clustering S1 according to G12 yields a dendrogram S1(G12) separating a group of PrGBM in S39. G12 contains mostly genes related to the immune system. In the dendrograms, a *box* represents a cluster. The T value at which the box is placed corresponds to the temperature at which the cluster disintegrates. The sizes of the clusters are not reflected by their boxes. The parameters for a stable gene cluster have been set at a stability threshold of 8, a maximal dropout of 3 at a single step of T, and the minimal cluster size at 5. The criteria for stable sample clusters are: Stab(C) > 8, maximal dropout at a single step of T is 1, and minimal cluster size is 3.

Table 3 *Genes[a] comprised in stable clusters G5(S1) and G12(S1) related to specific biological processes*

| Gene cluster | GenBank acc. no. | Gene symbol | Location[b] | S[c] | Gene description | Putative function |
|---|---|---|---|---|---|---|
| G5 | | | | | | |
| 1 | AB004904 | SOCS-3 | 17q25.3 | 1 | suppressor of cytokine signaling 3; STAT-induced STAT inhibitor 3 (STATI3) | SOCS3 is involved in negative regulation of cytokines that signal through the JAK/STAT pathway. Inhibits cytokine signal transduction by binding to tyrosine kinase receptors including gp130, LIF, erythropoietin, insulin, and leptin receptors. Interacts with multiple activated proteins of the tyrosine kinase signaling pathway, including IGF1 receptor, insulin receptor, and JAK2 (40). |
| 2 | M32977 | VEGF | 6p21.1 | 1 | VEGF-A | Growth factor active in angiogenesis, vasculogenesis, and endothelial cell growth. It induces endothelial cell proliferation, promotes cell migration, inhibits apoptosis, and induces permeabilization of blood vessels. |
| 3 | M35410 | IGFBP2 | 2q35 | 1 | IGFBP2 | IGF-binding proteins prolong the half-life of the IGFs and have been shown to either inhibit or stimulate the growth-promoting effects of the IGFs on cell culture. They alter the interaction of IGFs with their cell surface receptors. |
| 4 | U01134 | FLT1 | 3q12.2 | 1 | fms-related tyrosine kinase 1; vascular endothelial growth factor receptor 1 (VEGFR1) | Receptor for VEGF, VEGFB, and PGF. Has a tyrosine-protein kinase activity. The VEGF-kinase ligand/receptor signaling system plays a key role in vascular development and regulation of vascular permeability. |
| 5 | M96322 | AKAP12 | 6q25.1 | 1 | A kinase anchor protein 12, gravin | Anchoring protein that mediates the subcellular compartmentation of protein kinase (PKA) and protein kinase C (PKC). May play a role in wound repair and vascular development (48). |
| 6 | AB004903 | SOCS-2 | 12q22 | 1 | Suppressor of cytokine signaling 2, STAT-induced STAT inhibitor 2 (STATI2) | Negative feedback system that regulates cytokine signal transduction. SOCS2 appears to be a negative regulator in the growth hormone/IGF1 signaling pathway (40). |
| 7 | M57399 | PTN | 7q33 | 0 | Pleiotrophin; heparin-binding growth factor 8, neurite growth-promoting factor 1 | Heparin-binding mitogenic protein. Has neurite extension activity. Angiogenic properties have been demonstrated (25). |
| 8 | J04111 | JUN | 1p32.1 | 0 | c-jun proto-oncogene; transcription factor AP-1 | Transcription factor, interacts with c-fos to form a dimer. Interacts with SMAD3/SMAD4 heterodimers. Interacts with TCF20. Cooperates with HIF-1 in hypoxia-induced gene transcription (49). |
| 9 | X79067 | ZFP36L1 | 14q24.1 | 0 | Zinc finger protein 36, C3H type-like 1, TIS11B protein; EGF response factor 1 | Probable regulatory protein involved in regulating the response to growth factors. |
| G12 | | | | | | |
| 1 | D78345 | IGHG3 | 14q32.33 | 1 | Human DNA for Ig $\gamma$ heavy-chain, membrane-bound-type and secrete-type | $\sigma$-region located between C mu and C $\Delta$ genes of human immunoglobulin heavy chain. |
| 2 | U12255 | FCGRT | 19q13.33 | 0 | IgG receptor Fc large subunit P51; FcRN | Binds to the Fc region of monomeric immunoglobulins $\gamma$. Mediates the uptake of IgG from milk. Possible role in transfer of immunoglobulin G from mother to fetus. |
| 3 | X78817 | ARHGAP4 | Xq28 | 0 | $\rho$-GAP hematopoietic protein C1 (RGC1); KIAA0131, ($\rho$ GTPase activating protein 4) | Inhibitory effect on stress fiber organization. May down-regulate $\rho$-like GTPase in hematopoietic cells. Predominantly in hematopoietic cells. |
| 4 | U19713 | AIF1 | 6p21.33 | 1 | allograft inflammatory factor 1 (AIF1); glutathione-*S*-transferase (GST) homologue, glutathione transferase $\omega$. | May play a role in macrophage activation and function. |
| 5 | U90313 | GSTTLp28 | 10q25.1 | | Acts as a small stress response protein likely involved in cellular redox homeostasis. | |
| 6 | X00497 | CD74 | 5q33.1 | 0 | invariant polypeptide of major histocompatibility complex, class II antigen-associated | Plays a critical role in MHC class II antigen processing. |
| 7 | K01171 | HLA-DRA | 6p21.32 | 1 | Major histocompatibility complex, class II, DR $\alpha$, HLA class II histocompatibility antigen $\alpha$ chain | MHC class II receptor activity |
| 8 | AF019562 | TYROBP | 19q13.12 | 1 | TYRO protein tyrosine kinase binding protein, DNAX activation protein 12 (DAP12) | Noncovalently associates with membrane glycoproteins of the killer-cell inhibitory receptor (KIR) family without an ITIM in their cytoplasmic domain. |
| 9 | J04164 | IFITM1 | CHR 11 | 0 | IFN-induced transmembrane protein 1 (9–27), Leu13, CD225 | Control of cell growth. Component of a multimeric complex involved in the transduction of antiproliferative and homotypic adhesion signals. |
| 10 | J04162 | FCGR3A | 1q23.3 | 1 | Fc fragment of $\gamma$G, low affinity IIIa, leukocyte IgG receptor (Fc-$\gamma$-R), CD16 | Receptor for the Fc region of IgG. Mediates antibody-dependent cellular cytotoxicity (adcc) and other antibody-dependent responses, such as phagocytosis. |
| 11 | X04011 | CYBB | Xp11.4 | 0 | Cytochrome b-245, $\beta$ polypeptide (chronic granulomatous disease), GP91-PHOX | Critical component of the membrane-bound oxidase of phagocytes that generates superoxide. |
| 12 | X14356 | FCGR1A | 1q21.2 | 0 | Fc fragment of IgG, high affinity Ia, CD64 | Binds to the Fc region of immunoglobulins $\gamma$. |
| 13 | M33882 | MX1 | 21q22.3 | 0 | IFN-regulated resistance GTP-binding protein MXA (IFI-78K); IFN-induced protein P78 | Shows activity against influenza virus and vsv, a rhabdovirus. |

[a] Information on genes listed in this table is taken from GeneCards.[5]
[b] Ensembl cytogenetic band.
[c] Separating gene according to statistical criteria, see "Materials and Methods" (Table S1 of supplementary information for complete gene list).

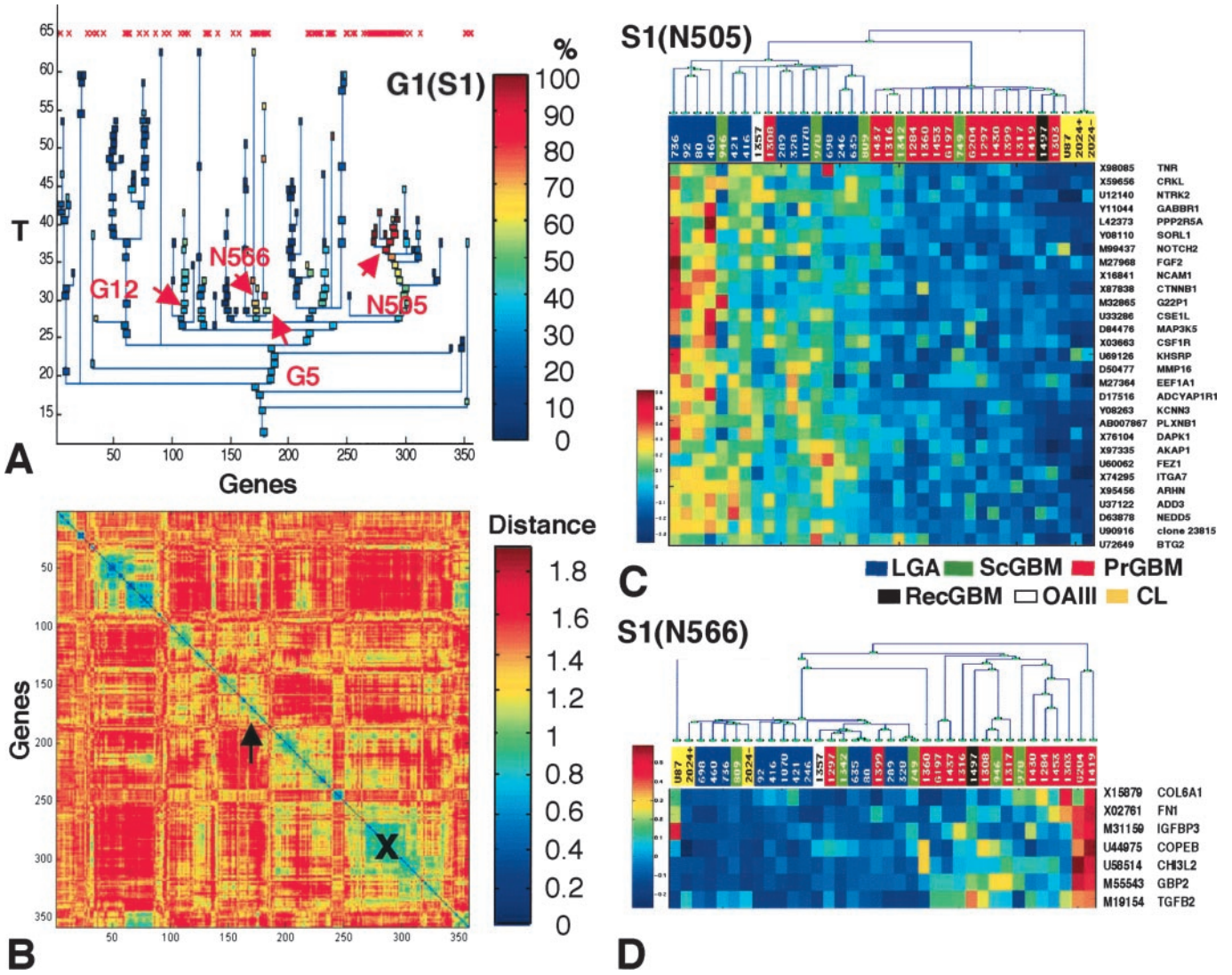



Fig. 3. Combination of supervised and unsupervised analysis. In *A*, the genes found by pairwise comparison of PrGBM *versus* LGA and ScGBM were identified in the gene dendrogram G1(S1) (see also Fig. 2*A*) and marked on *top* with a *red x*. The percentage of these genes in a given gene cluster is presented in a color code. Most of the identified genes are grouped in four to five clusters. Two of them had been found previously (G5 and G12) *arrow*. The two new clusters identified are particularly rich in these genes, denoted as N505 and N566. They were not stable according to the criteria set; however, the distance matrix (*B*) recognizes these nodes as comprising closely related genes, indicated by an *arrow* for N566 and an *X* for N505. The distance D is visualized as a color code, with *blue* indicating short distance. Note the order of the genes, and the scale is the same for *A* and *B*. Reclustering of all experiments S1 according to N505 and N566, respectively, is shown in *C* and *D*. Both nodes yield an almost perfect separation of the tumors according to their classification. This is not surprising because in both N505 and N566, all but one of the genes were identified as separating in the binary comparison.

were identified also as rich in genes that partition samples according to the other supervised binary comparisons mentioned above. Two of the identified clusters were indeed G5 (six of nine genes are among the 91 separating ones) and G12 (5 of 13). However, two additional clusters were identified as rich in separating genes; these are marked in Fig. 3*A* as nodes N505 (28 of 29) and N566 (6 of 7). These sets of genes (Table 4) were not selected as stable clusters by CTWC because they were either too small or they "leaked" too many genes as the control parameter *T* was increased. However, as clearly shown in the distance matrix (Fig. 3*B*), the genes of both of these nodes had relatively highly correlated expression profiles.

These additional clusters contain some interesting genes; members of the Wnt- and Notch-pathway, genes involved in adhesion/migration, and apoptosis-related genes in node N505 (Fig. 3*C*; Table 4) and in node N566 genes like *TGFβ2* and *IGFBP3* that have been reported before to be overexpressed in high-grade gliomas (Fig. 3*D*; Table 4).

Next, we used these two additional clusters to partition all of the samples S1. As shown in Fig. 3*C*, node N505 gives rise to a nearly perfect separation of the samples into three clusters: (*a*) "Non-PrGBM"; of 17 samples, comprising 12 of 12 LGA, 3 of 5 ScGBM, and 1 PrGBM and the OAIII, both of which (1308 and 1357) joined the non-PrGBM cluster in S1(G5); (*b*) mainly PrGBM, 13 of 14 PrGBM, the RecGBM (1497), 2 of 5 ScGBM, one of which (1342) had joined the PrGBM in S1(G5), 1 cell line (U87); and (*c*) a cluster with the remaining cell line under two conditions (wild-type *TP53* on and off; 2024+, 2024−). Hence, node N505 gives a clear separation into LGA *versus* PrGBM, with the ScGBM split evenly. As we also see in Fig. 3*D*, node N566 identifies a cluster of 20 samples comprising 12 of 12 LGA; thus, this node also separates the samples largely into LGA *versus* PrGBM, with the ScGBM distributed evenly.

### Constructing a Tumor Classifier

A gene expression based tumor classifier was built to validate our findings. A successful test of such a classifier based on the selected genes and applied to an independent validation set of 20 samples

Table 4 *Genes[a] comprised in nodes N505(S1) and N566(S1) rich in separating genes*

| Gene cluster | GenBank acc. no. | Gene symbol | Location[b] | S[c] | Gene description | Putative function |
|---|---|---|---|---|---|---|
| N505 | | | | | | |
| 1 | X98085 | TNR | 1q25.1 | 1 | tenascin-R; restrictin; janusin | Extracellular matix protein, secreted by oligodendrocytes during myelination, some astrocytes and neurons. Can modify adhesive substrate properties of fibronectin. Involved in cell adhesion, migration, and differentiation. |
| 2 | X59656 | CRKL | 22q11.21 | 1 | v-crk sarcoma virus CT10 oncogene homologue (avian)-like; CRK-like protein | May mediate the transduction of intracellular signals. |
| 3 | U12140 | NTRK2 | 9q21.33 | 1 | neurotrophic tyrosine kinase, receptor, type 2; brain-derived neurotrophic factor (BDNF)/NT-3 growth factors receptor; TRKB tyrosine kinase receptor; GP145-TRKB | Receptor for brain-derived neurotrophic factor (BDNF), neurotrophin-3 and neurotrophin-4/5 but not nerve growth factor (NGF) |
| 4 | Y11044 | GABBR1 | 6p22.1 | 1 | γ-aminobutyric acid (GABA) B receptor, 1 | Receptor for GABA. |
| 5 | L42373 | PPP2R5A | 1q32.3 | 1 | protein phosphatase 2, regulatory subunit B (B56), alpha isoform); protein phosphatase 2A B56-α, PP2A | The b regulatory subunit might modulate substrate selectivity and catalytic activity, and also might direct the localization of the catalytic enzyme to a particular subcellular compartment. |
| 6 | Y08110 | SORL1 | 11q23.3 | 1 | sortilin-related receptor L(DLR class) A repeats-containing); low-density lipoprotein receptor-related protein LR11 | Likely to be a multifunctional endocytic receptor, that may be implicated in the uptake of lipoproteins and of proteases. Could play a role in cell-cell interaction. Expressed mainly in brain. |
| 7 | M99437 | NOTCH2 | 1p11.2 | 1 | Notch homologue 2, neurogenic locus notch protein | Functions as a receptor for membrane-bound ligands Jagged1, Jagged2, and Δ1 to regulate cell-fate determination. Affects the implementation of differentiation, proliferation, and apoptotic programs. |
| 8 | M27968 | FGF2 | 4q27 | 0 | fibroblast growth factor, basic; FGFB; heparin-binding growth factor 2 precursor (HBGF2) | The heparin-binding growth factors are angiogenic agents *in vivo* and are potent mitogens for a variety of cell types *in vitro.* There are differences in the tissue distribution and concentration of these two growth factors. |
| 9 | X16841 | NCAM1 | 11q23.1 | 1 | neural cell adhesion molecule 1; NCAM120; CD56 | Cell adhesion molecule involved in neuron-neuron adhesion, neurite fasciculation, outgrowth of neurites, etc. |
| 10 | X87838 | CTNNB1 | 3p22.1 | 1 | β-beta catenin | Involved in the regulation of cell adhesion and in signal transduction through the wnt pathway. |
| 11 | M32865 | G22P1 | 22q13.2 | 1 | thyroid autoantigen 70 kDa; Ku 70-kDa subunit; | Single stranded DNA-dependent ATP-dependent helicase. Has a role in chromosome translocation. |
| 12 | U33286 | CSE1L | 20q13.13 | 1 | chromosome segregation 1-like (yeast); cellular apoptosis susceptibility protein (CAS); | Export receptor for importin α. Mediates importin-α reexport from the nucleus to the cytoplasm after import substrates have been released into the nucleoplasm. Highly expressed in proliferating cells. |
| 13 | D84476 | MAP3K5 | 6q23.3 | 1 | mitogen-activated protein kinase kinase kinase 5; MAP/ERK kinase kinase 5; MAPKKK5; MEKK5; ASK1 | Phosphorylates and activates two different subgroups of MAP kinase kinases, MKK4/SEK1 and MKK3/MAPKK6 (OR MKK6). Overexpression induces apoptotic cell death. |
| 14 | X03663 | CSF1R | 5q33.1 | 1 | colony-stimulating factor I receptor precursor; fms proto-oncogene (c-fms); CD115 | This protein is the receptor for CSF-1, it is a protein tyrosine-kinase transmembrane receptor. |
| 15 | U69126 | KHSRP | 19p13.3 | 1 | KH-type splicing regulatory protein; fuse-binding protein 2 (FBP2) | A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer. |
| 16 | D50477 | MMP16 | 8q21.3 | 1 | matrix metalloproteinase 16 (membrane-inserted); membrane-type matrix metalloproteinase 3 (MT-MMP3) | Endopeptidase, degrades components of the extracellular matrix, such as collagen type III and fibronectin. Activates progelatinase A. Involved in the matrix remodeling of blood vessels. No effect on type I, II, IV, and V collagen. |
| 17 | M27364 | EEF1A1 | 6q14.1 | 1 | eukaryotic translation elongation factor 1 α 1; EF1 α | This protein promotes the GTP-dependent binding of aminoacyl-trna to the a-site of ribosomes during protein biosynthesis. |
| 18 | D17516 | ADCYAP1R1 | 7p14.3 | 1 | adenylate cyclase activating polypeptide 1 (pituitary) receptor type I; PACAP I receptor; PACAPR1 | Receptor for pacap-27 and pacap-38. The activity of this receptor is mediated by G proteins, which activate adenylyl cyclase. |
| 19 | Y08263 | KCNN3 | 1q22 | 1 | potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3; AAD14 | Forms a voltage-independent potassium channel activated by intracellular calcium. |
| 20 | AB007867 | PLXNB1 | 3p21.31 | 1 | plexin B1; plexin 5; KIAA0407 | A family of transmembrane proteins with homology to the MET-hepatocyte growth factor receptor. |
| 21 | X76104 | DAPK1 | 9q21.33 | 1 | death-associated protein kinase 1; DAP kinase 1 | Involved in mediating IFN-γ-induced cell death. |
| 22 | X97335 | AKAP1 | 17q23.2 | 1 | A-kinase anchoring protein 1; PRKA | Binds to type I and II regulatory subunits of protein kinase A and anchors them to the cytoplasmic face of the mitochondrial outer membrane. |
| 23 | U60062 | FEZ1 | 11q24.2 | 1 | fasciculation and elongation protein zeta 1; zygin I | May be involved in axonal outgrowth as component of the network of molecules that regulate cellular morphology and axon guidance machinery. |
| 24 | X74295 | ITGA7 | 12q13.2 | 1 | integrin α 7B precursor; IGA7B | Integrin α-7/β-1 is the primary laminin receptor on skeletal myoblasts and adult myofibers. |
| 25 | X95456 | ARHN | 17q23.3 | 1 | ras homologue gene family, member N; Rho7 protein | May be specifically involved in neuronal and hepatic functions. |
| 26 | U37122 | ADD3 | 10q25.1 | 1 | adducin 3 γ | Membrane cytoskeleton-associated protein that promotes the assembly of the spectrin-actin network. Binds to calmodulin. |
| 27 | D63878 | NEDD5 | 2q37.3 | 1 | neural precursor cell expressed, developmentally down-regulated 5 | Mammalian septin is a novel cytoskeletal component interacting with actin-based structures |
| 28 | U90916 | clone 23815 | | 1 | Human clone 23815 | Soares library 1NIB from IMAGE Consortium |
| 29 | U72649 | BTG2 | 1q32.1 | 1 | btg protein; NGF-inducible antiproliferative protein PC3 | Antiproliferative protein. |

Table 4 *Continued*

| Gene cluster | GenBank acc. no. | Gene symbol | Location[b] | S[c] | Gene description | Putative function |
|---|---|---|---|---|---|---|
| N566 | | | | | | |
| 1 | X15879 | COL6A1 | 21q22.3 | 0 | collagen, type VI, $\alpha$-1 | Collagen VI acts as a cell-binding protein. |
| 2 | X02761 | FN1 | 2q35 | 1 | fibronectin 1 | Fibronectins bind cell surfaces and various compounds including collagen, fibrin, heparin, DNA, and actin. Fibronectins are involved in cell adhesion, cell motility, opsonization, wound healing, and maintenance of cell shape. |
| 3 | M31159 | IGFBP3 | 7p13 | 1 | IGFBP 3 | IGF-binding proteins prolong the half-life of the IGFs and have been shown to either inhibit or stimulate the growth-promoting effects of the IGFs on cell culture. They alter the interaction of IGFs with their cell surface receptors. |
| 4 | U44975 | COPEB | 10p15.2 | 1 | core promoter element binding protein; DNA-binding protein CPBP | Transcriptional activator (by similarity). Binds a GC box motif. Could play a role in B-cell growth and development. |
| 5 | U58514 | CHI3L2 | 1p13.3 | 1 | chitinase 3-like 2 | Not detected in brain, spleen, pancreas, and liver. Belongs to family 18 of glycosyl hydrolases. |
| 6 | M55543 | GBP2 | 1p22.2 | 1 | guanylate binding protein 2, IFN-inducible | Binds GTP, GDP, and GMP. Induction by IFN-$\gamma$ during macrophage activation. |
| 7 | M19154 | TGFB2 | 1q41 | 1 | TGF-$\beta$2; TGFB2 | TGF-$\beta$2 has suppressive effects on interleukin-2 dependent t-cell growth. |

[a] Information on genes listed in this table is taken from GeneCards.[5]
[b] Ensembl cytogenetic band.
[c] Separating gene according to statistical criteria, see "Materials and Methods" (Table S1 of supplementary information for complete gene list).

indicates that the separation found previously is robust and was not caused by random fluctuations. For each separating cluster of genes (G5, G12, node N505, and node N566) and their combination, we generated a *k*-nearest neighbor classifier (k = 3) and tested its ability to perform class partitions. In all cases, the separation power was found to be significant (Table 5). In addition, we tested whether a three-way separation of the tumor types can be performed based on the combination of the four gene clusters. This gave rise to only one to two errors out of the 20 gliomas, thus validating our claim that the tumor types have different gene expression profiles that can be used to distinguish between them ($P = 2.49E-06$). When using the gene sets separately, the distinction of PrGBM *versus* ScGBM was the least successful. Most classification errors disappeared when all four selected clusters were used. This suggests that distinct sets discriminate between different aspects of the classes. In fact, genes of the same cluster were generally highly correlated across the samples of the same class as a result of the selection procedure, whereas genes of different clusters were not. Although the best performance was obtained by combining all genes, the G5 set alone provided already a good separation power in all cases, including three-way classification where only one to two (8%) errors were made (Table 5).

## Validation of Gene Expression Data by Other Methods

Genes found most significantly up-regulated in these glioma as compared with normal brain comprise genes involved in the cell cycle, angiogenesis, migration, and immune response, such as vimentin, fibronectin, *EGFR*, *VEGF*, *CDK4*, *PDGFRA*, *SPARC*, *HLA-G*, and *HLA-DRA*. Similar lists have been published before (23, 27–29). An overview of the most differentially expressed genes is available in Table S2 of supplementary information.

To further validate our findings obtained by gene expression profiling, we confirmed a gene known to be overexpressed, *EGFR*, on respective paraffin sections by immunohistochemistry. Furthermore, we validated differential expression of *TNR* on a larger panel of tumors using immunohistochemistry on tissue arrays.

A good correlation between *EGFR* overexpression as measured on cDNA arrays and by immunohistochemical detection of the EGFR on respective paraffin sections (no/low expression *versus* high expression) was observed using a subset of 26 gliomas. The mean expression levels were 265 (SD 101) for samples with no/low expression as determined by immunohistochemistry and 1052 (SD 1369), respectively, for biopsies with high immunopositivity (unpaired *t* test, $P = 0.035$). Please note that expression of EGFR as determined by immunohistochemistry can be focal.

Relative overexpression of *TNR* in LGA as compared with PrGBM and ScGBM was indicated by supervised analysis and its presence in the separating gene cluster N505. We chose to validate TNR, a multifunctional extracellular matrix glycoprotein that is of interest for further evaluations because of its role in adhesion, migration, and differentiation during central nervous system morphogenesis, in particular, differentiation of oligodendrocytes (30). This member of the tenascin family is exclusively expressed in the central nervous system

Table 5 *Verification of k-NN-class prediction using selected gene clusters on validation set of astrocytic gliomas*

| | Classification Two-Way, k = 3 | | | | | | | | | | Classification[c] Three-Way, k = 3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ScGBM *vs.* LGA | | PrGBM *vs.* ScGBM | | PrGBM *vs.* LGA | | PrGBM *vs.* (ScGBM + LGA) | | LGA *vs.* (ScGBM + PrGBM) | | LGA *vs.* PrGBM *vs.* ScGBM | |
| Comparisons Gene clusters used by the classifier[a] | 4 *vs.* 12 No. of errors (%) | $P$[b] | 4 *vs.* 4 No. of errors (%) | $P$ | 4 *vs.* 12 No. of errors (%) | $P$ | 4 *vs.* (4 + 12) No. of errors (%) | $P$ | 12 *vs.* (4 + 4) No. of errors (%) | $P$ | 12 *vs.* 4 *vs.* 4 Mean no. of errors (%) | Mean $P$ |
| G5 | 2 (12%) | 0.05 | 1 (13%) | 0.1 | 1 (6%) | 0.007 | 1 (5%) | 0.004 | 1 (5%) | 0.0001 | 1–2 (8%) | 6E-05 |
| N505 | 2 (12%) | 0.03 | 3 (37%) | 1 | 2 (12%) | 0.03 | 4 (20%) | 0.06 | 3 (15%) | 0.003 | 6 (30%) | 0.002 |
| N566 | 2 (12%) | 0.05 | 2 (25%) | 0.5 | 1 (6%) | 0.007 | 2 (10%) | 0.01 | 3 (15%) | 0.004 | 4 (20%) | 0.003 |
| G12 | 3 (19%) | 0.1 | 2 (25%) | 0.5 | 2 (12%) | 0.03 | 3 (15%) | 0.03 | 7 (25%) | 0.02 | 5 (27%) | 0.05 |
| G5 + G505 + G566 + G12 | 1 (6%) | 0.007 | 0 (0%) | 0.03 | 0 (0%) | 0.0005 | 0 (0%) | 0.0002 | 0 (0%) | 8E-06 | 1–2 (7%) | 2E-06 |

[a] The classifier was built on the training set of 31 astrocytic gliomas, employing the respective gene set and the k-nearest neighbor (k-NN, k = 3) method.
[b] Fisher's exact test.
[c] Classification was repeated 100 times, and *P*s were averaged, see "Materials and Methods" for explanation.

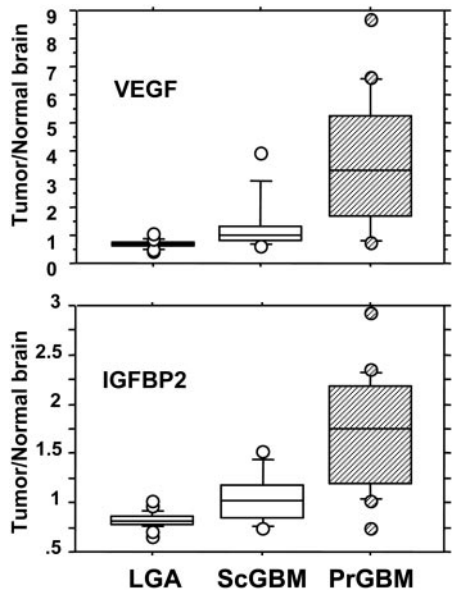



Fig. 4. Angiogenic activity correlates with tumor subtype. The relative expression of *VEGF* and *IGFBP2* as compared with normal brain is shown for the three tumor subtypes, using all samples of the training and validation set. Binary comparisons using the unpaired *t* test between any two tumor subtypes showed highly significant differences for *VEGF* and *IGFBP2*, respectively. LGA *versus* PrGBM: *VEGF*, $P < 0.0001$; *IGFBP2*, $P < 0.0001$; LGA *versus* ScGBM: *VEGF*, $P = 0.0061$; *IGFBP2*, $P = 0.0003$; PrGBM *versus* ScGBM: *VEGF*, $P = 0.0099$; *IGFBP2*, $P = 0.0027$.

in particular by oligodendrocytes. TNR expression was validated on a large panel of gliomas by immunohistochemistry on our GBM-TA and non-GBM glioma-TA, respectively. The criteria for semiquantitative analysis were no/low expression as opposed to high expression. Twenty-nine "pure" LGA were available for analysis on the array applying the same pathology criteria as for the LGA included for gene expression profiling. LGA exhibited significantly higher expression of TNR (83%, 24 of 29) than GBM (63%, 84 of 133 available for analysis; Fisher's exact test, $P = 0.03$).

**IGFBP2 Is Coexpressed with VEGF Exerting Most Prominent Expression in Pseudopalisading Cells Surrounding Tumor Necrosis.** Most of the genes in G5 are involved in angiogenesis in a broader sense as detailed above. Within this cluster, *IGFBP2* expression displayed the highest correlation ($r = 0.85$) with *VEGF* expression as shown in Fig. 2*B*. Both genes exhibited significant differential expression between any two tumor subclasses as visualized in a box-plot in Fig. 4, using all 51 astrocytic gliomas of the test and the validation set. To further understand the biology of *IGFBP2*, we addressed the question whether the two genes are involved in the same process. Therefore, *in situ* hybridization was performed on serial frozen sections of glioblastomas for *IGFBP2* and *VEGF*, respectively. Expression of *IFGBP2* and *VEGF* was most prominent in pseudopalisading cells surrounding tumor necrosis of GBM (Fig. 5; *VEGF in situ* hybridization; data not shown). Thus, the *IGFBP2* expression pattern in the glioblastoma tissue is superposable to the reported pattern for *VEGF* (16).

**Anoxia Induces Expression of *IGFBPs* in Glioblastoma Cell Lines.** Subsequently, we investigated the regulation of *IGFBP2* and two other family members, *IGFBP3* and *5*, under anoxic conditions. In PrGBM, increased expression of *IGFBP2* was often paralleled by overexpression of either *IGFBP3* (comprised in N566; Fig. 3*D*) or *IGFBP5* (comprised in the list of separating genes, see Table S1 of supplementary information). Glioblastoma cell lines LNZ308 and LN229 displayed *IGFBP2* up-regulation concomitant with *VEGF* after 20 h of anoxia (Fig. 6). Similarly, *IGFBP5* was induced in these two cell lines, whereas U87 that is wild-type for TP53 exerted most prominent induction of *IGFBP3*, a TP53 target gene. Thus, *IGFBP3* and 5 might also be induced in the same process in gliomas. Induction of different members of the family in these cell lines is in line with a certain redundancy of these genes observed in *IGFBP2* knockout mice. These show increased levels of other *IGFBPs* and display a much less dramatic phenotype than the one initially predicted based on the fetal *IGFBP2* expression pattern (31).

## DISCUSSION

Our characterization of astrocytic gliomas by their gene expression profiles revealed consistent inherent differences between LGA, ScGBM, and PrGBM. Thus, the previous histological and clinical recognition of these three glioma entities can be strongly supported by cDNA-array data as reflected by MDS of overall gene expression (Fig. 1). The LGA were found to be the most closely related group and well separated from PrGBM that displayed the most heterogeneous expression profiles. The ScGBM share some of the features of both subgroups, reflecting the common pathogenetic pathway with LGA and the malignant behavior with PrGBM.

Analysis of the expression profiles for correlated genes separating

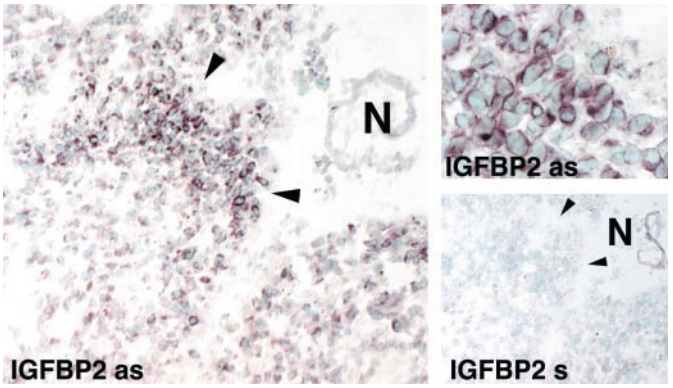



Fig. 5. *IGFBP-2* mRNA is localized to pseudopalisading cells in human glioblastoma. Hybridization with an *IGFBP-2* antisense cRNA probe yields a strong signal in the pseudopalisading cells, as indicated with *arrows*, surrounding tumor necrosis (*N*) of a GBM, same region in higher magnification is displayed in the *top right panel*. *Bottom right panel*, *IGFBP-2* sense negative control.

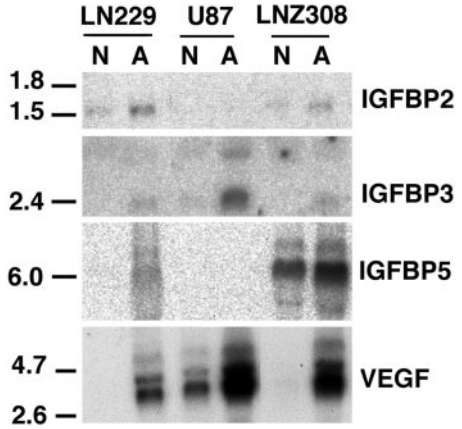



Fig. 6. Induction of *IGFBP2*, *3*, or *5* by anoxia in glioblastoma cell lines. Glioblastoma cell lines LN229, U87, and LNZ308 were cultured under normoxic (*N*) or anoxic (*A*) conditions for 20 h. Expression *of IGFBP2*, *3*, and *5* was evaluated by Northern blot analysis and compared with induction of *VEGF*. At least one of the analyzed *IGFBP* family members was induced on hypoxic treatment. Note, U87 the only cell line in this series harboring wild-type TP53 exhibits substantial induction of the TP53 target gene *IGFBP3*. Fragment sizes in kilobases are marked on the *left*.

tumor subtypes allowed identification of sets of genes implicating particular biological features that are associated with tumor-specific subtypes. Most strikingly, a cluster (G5) of correlated genes suggests that inherent angiogenic activity, manifested in overexpression of known angiogenic factors, including the most potent, *VEGF*, and *FLT1* and *PTN*, distinguishes PrGBM from the other two groups. In fact, this feature alone was sufficient to classify most astrocytic gliomas (92%) correctly according to their subtype. Glioblastoma are among the most vascularized tumors with characteristic, structurally and functionally abrogated blood vessels of multilayered glomeruloid pattern. Angiogenesis, an essential requisite for tumor growth, is a tightly controlled, complex process involving various pro and antiangiogenic factors that need to act in a concerted fashion to yield functional blood vessels (32). The angiogenic factors VEGF, angiopoietins, and Ephrins act specifically on endothelial cells, because their respective receptors FLT1 and KDR, Tie receptors, and Ephrin receptors are exclusively expressed on these cells. In contrast, other angiogenic factors highly abundant in glioma, such as FGF2, PDGF, and TGF-$\beta$, act also on other cell types (33).

According to our analysis, PrGBM appear to have higher angiogenic activity mediated by the genes comprised in G5 than LGA but also higher activity than ScGBM. The difference between PrGBM and ScGBM is unexpected, because they have the same malignancy grade (WHO IV) and cannot be discriminated by classical histology. This may reflect the fact that PrGBM are rapidly growing tumors that cannot keep up with their increasing need for blood supply and thus may suffer from more severe hypoxic conditions triggering angiogenic activity. In contrast, ScGBM may progress over years from LGA and thus may use different pathways for sustained angiogenesis. As compared with PrGBM, LGA display relative overexpression of the angiogenic factor *FGF2* that emerged in the separating gene node N505 (Fig. 3*C*; all LGA *versus* all PrGBM, $P < 0.01$, unpaired *t* test).

An important role for *FGF2* in LGA is also supported by findings from expression profiling efforts reported from small sets of pediatric gliomas (6 LGA *versus* 7 high-grade astrocytoma; Ref. 34) revealing significantly higher *FGF2* expression in LGA. Differential expression of *VEGF*, however, did not reach statistical significance, possibly because of the small numbers of samples. Interestingly, in this study (34) that appeared during revision of this manuscript, the authors used the opposite strategy to analyze their gene expression data obtained from these pediatric gliomas. A knowledge-based list of 133 genes related in some way to angiogenesis was constructed and successfully used to separate LGA from GBM according to hierarchical clustering and PCA. This supervised gene reduction yielded similar separation power as obtained using their full set of 9198 genes present on the chip, thus again emphasizing the important role of angiogenesis in tumor progression.

The biological differences between mechanisms of angiogenesis may have important implications for response to antiangiogenic therapy, which is about to enter the clinics (35).

An important implication of the CTWC method concerns investigation of the genes that belong to identified clusters, here in particular G5, for their biological function, assuming that coexpressed genes might be coregulated and therefore might be involved in the same biological process. The fact that *IGFBP2* is strongly correlated with the angiogenesis genes came as a surprise that will be worth detailed additional studies. The presented association supported by our finding of coexpression of *IGFBP2* and *VEGF* in pseudopalisading cells surrounding necrosis in GBM (Fig. 5) provides strong evidence that *IGFBP2* overexpression in astrocytic gliomas may reflect its implication in angiogenesis. These correlations add to the fact the *IGFBP2* can be induced under anoxic conditions in glioblastoma cell lines (Fig. 6) and mouse embryonic stem cells (26).

Overexpression of *IGFBP2* in GBM has first been discovered by Fuller *et al.* (36) using gene expression profiling. This finding was confirmed (23, 27, 34, 37) and extended to other tumor types. Subsequent immunohistochemical investigations for IGFBP2, a secreted protein, suggested a significant association with the histopathological malignancy grade in glioma (37) with more pronounced immunopositivity in macrophage/microglial and glioma cells around focal necrosis (38). The IGFBPs were first identified and characterized based on their role to bind and modulate the *in vivo* bioactivity of the mitogenic and anabolic peptides IGF-I and IGF-II. More recently, it has become clear that IGFBPs are multifunctional proteins with IGF-dependent and -independent functions in controlling growth and metabolism. The intrinsic bioactivity of IGFBPs, in a positive or negative fashion, depends not only on the cell type but also on their differentiation state and physiological/pathophysiologic condition (39). Furthermore, additional genes of cluster G5 have been linked to the IGF/IGFBP system (*SOCS2* and *3*, *ZFP36L1*; Refs. 40 and 41). For IGF-I itself, a role in angiogenesis has been established (26, 42). Thus, our observations support the idea to develop a therapeutic approach targeting the IGF-I/IGFBP system in cancer.

Another separating gene cluster emerging from CTWC analysis associated with known biological function is G12 (Fig. 2; Table 3). This gene cluster partitioned a subgroup of PrGBM with increased expression from LGA and comprises only genes related to the immune system. Half of the genes are known to be IFN inducible and are likely expressed by lymphoid cells and/or macrophages and microglia; others represent markers of natural killer cells. These results confirm the observation by Fathallah-Shaykh *et al.* (29) who reported elevated expression of MHC class I and II genes. However, some of the genes are potentially expressed by tumor cells according to our preliminary results. It is worth to note that studies trying to relate immune response in gliomas with survival have not yielded conclusive results. The notion that gliomas produce an immuno-compromised environment by secreting cytokines, such as TGF$\beta$-2 (43, 44), is also supported by this study; TGF$\beta$-2 emerged in node N566 that separates PrGBM from LGA (Fig. 3*D*). Along the same lines, we found HLA-G to be overexpressed (over normal brain, mean 2.9) in PrGBM. This nonclassical class I antigen is mainly expressed at the materno-fetal interface during pregnancy where it plays an important role in maternal-fetal tolerance. HLA-G expression has been associated with tumor progression in several tumor types and has been suggested to contribute to escape from immune surveillance (45, 46). Furthermore, functional assays with human glioma cell lines indicated that few HLA-G-positive cells within a population of HLA-G-negative tumor cells *in vitro* are sufficient to exert a significant immune inhibitory effect (46).

The tumor classifier using expression data of all four identified discriminatory clusters yielded correct tumor prediction in 93% of an independent set of astrocytic gliomas. This remarkable success has to be appreciated in light of the inherent inaccuracy of the initial tumor diagnosis for gliomas, because of sampling errors and subjective histological criteria that lead to striking interobserver differences among neuropathologists (47). In this context, three tumors originally designated as PrGBM attracted our attention (1357, 1297, and 1308), because they appear evenly in clusters characteristic for PrGBM and LGA, respectively, depending on the analysis (Figs. 2 and 3). Central pathology review revised the first (1357) as OAIII, whereas the other two cases were confirmed. Tumor 1297, clinically defined as a PrGBM, presented with genetic characteristics of a ScGBM, namely TP53 mutation and most prominent overexpression of PDGF receptor. Finally, glioblastoma 1308 exhibited overexpression of EGFR, characteristic for PrGBM. However, the biopsy used for gene expression profiling comprised a part of the tumor displaying only features of

WHO grade II (the most malignant part of tumor defines diagnosis). This demonstrates that regions of a glioblastoma with the appearance of a lower grade astrocytoma exert already features only found in PrGBM, exemplified here as overexpression of the EGFR gene and the genes comprised in G12 and N566, characteristic for PrGBM. In contrast, after reclustering S1 (all samples) according to G5, comprising genes indicating high angiogenic activity, this tumor (PrGBM 1308) partitions with LGA. This is biologically convincing, because lower grade astrocytoma have no necrosis and thus are expected to suffer from less hypoxic stress. This observation might be of clinical relevance, knowing the difficulties in tumor sampling at surgery in particular stereotactic biopsies. Thus, the true tumor grade may be underestimated. It follows, the identification of markers specific for glioblastoma that are detectable in apparent lower grade areas of such tumors would allow to improve the diagnostic reliability and ultimately the choice of therapy. Interestingly, in MDS, both 1297 and 1308 were located at the interface of ScGBM and PrGBM.

Combined unsupervised and supervised analysis is a novel approach of gene selection that allows identification of clusters rich in genes informative (by supervised analysis) for tumor classification. CTWC was particularly suitable for this task, because it is designed to go beyond clustering of all genes on the basis of the data from all tumors, and clustering of all tumors, using data from all genes. This is of importance because most of the genes for which expression levels have been measured are irrelevant for the partition sought. CTWC proposes identification of correlated groups (clusters) of genes and using only data from one such group at a time to recluster the tumors, or *vice versa*.

Our gene selection approach has some interesting features, and allowed us: (*a*) to reduce the many discriminatory genes obtained by binary class comparisons; (*b*) enabled an almost correct classification of tumor entities, namely discrimination of LGA from ScGBM, and ScGBM from PrGBM, although pairwise comparisons yielded only few separating genes (partially because of small numbers of ScGBM in the training set, $n = 5$; threshold of FDR of $q < 0.05$); (*c*) by using the signal of a group of correlated genes, the noise of the individual measurements averages out and is reduced; and most importantly (*d*) the identified gene clusters yielded information on the biological context of coexpressed and possibly coregulated genes. Such insight may give some indication on biological processes determining tumor entities. This fact is highlighted in this study in particular by identification of G5, with high discriminatory power on its own, featuring angiogenesis-related genes.

Analyzing gene expression profiles has yielded biologically relevant information using a limited set of genes (1185). This approach is a first step toward molecular diagnostics for astrocytic gliomas that may improve tumor diagnoses in the future by adding objective criteria. Gene expression profiles may ultimately provide a tool for the identification of patients who are most likely to benefit from targeted therapy. It follows that such approaches for outcome prediction need to be established in clinical trials that in turn will allow rational design of future therapies.

Here, we discussed mostly findings demonstrating similarities within and differences between tumor subtypes for classification purposes. Numerous interesting clusters emerging from this analysis are awaiting attention to gain additional insights into the biology of astrocytic gliomas, *e.g.*, in regard to differentiation and migration (*e.g.*, *TNR*, node 505), and the involvement of the wnt/notch pathways as indicated in N505, and other clusters that have not been discussed in this study [found in G1(S2), see full CTWC analysis on the web[5]].

## ACKNOWLEDGMENTS

We thank our colleagues who made this study possible by participating actively in the clinical trial and providing primary tumor samples, namely Drs. J-G. Villemure, F. Porchet, O. Vernet, P. Otten, A. Reverdin, and B. Rilliet. We also thank Dr. P. Reymond for helpful discussions; Drs. S. Ostermann, M. Albertoni, and G. Pizzolato for their collaboration; and Dr. P. Walker for critical reading of this manuscript.

## REFERENCES

1. Kleihues, P., and Cavenee, W. K. Pathology & Genetics. Tumours of the Nervous System. Lyon, France: IARC, 2000.
2. Hegi, M. E., zur Hausen, A., Rüedi, D., Malin, G., and Kleihues, P. Hemizygous or homozygous deletion of the chromosomal region containing the p16INK4a gene is associated with amplification of the EGF receptor gene in glioblastomas. Int. J. Cancer, *73:* 57–63, 1997.
3. Watanabe, K., Tachibana, O., Sato, K., Yonekawa, Y., Kleihues, P., and Ohgaki, H. Overexpression of the EGF receptor and *p53* mutations are mutually exclusive in the evolution of primary and secondary glioblastomas. Brain Pathol., *6:* 217–224, 1996.
4. Hermanson, M., Funa, K., Koopmann, J., Maintz, D., Waha, A., Westermark, B., Heldin, C-H., Wiestler, O. D., Louis, D. N., von Deimling, A., and Nistér, M. Association of loss of heterozygosity on chromosome 17p with high platelet-derived growth factor $\alpha$ receptor expression in human malignant gliomas. Cancer Res., *56:* 164–171, 1996.
5. Zhu, Y., and Parada, L. F. The molecular and genetic basis of neurological tumours. Nat. Rev. Cancer, *2:* 616–626, 2002.
6. Shawver, L. K., Slamon, D., and Ullrich, A. Smart drugs: tyrosine kinase inhibitors in cancer therapy. Cancer Cell, *1:* 117–123, 2002.
7. Getz, G., Levine, E., and Domany, E. Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. USA, *97:* 12079–12084, 2000.
8. Getz, G., Gal, H., Kela, I., Notterman, D. A., and Domany, E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics, *19:* 1079–1089, 2003.
9. Getz, G., and Domany, E. Coupled two-way clustering server. Bioinformatics, *19:* 1153–1154, 2003.
10. Stupp, R., Dietrich, P.-Y., Ostermann Kraljevic, S., Pica, A., Maillard, I., Maeder, P., Meuli, R., Janzer, R., Pizzolato, G., Miralbell, R., Porchet, F., Regli, L., de Tribolet, N., Mirimanoff, R. O., and Leyvraz, S. Promising survival for patients with newly diagnosed glioblastoma multiforme treated with concomitant radiation plus temozolomide followed by adjuvant temozolomide. J. Clin. Oncol., *20:* 1375–1382, 2002.
11. Ishii, N., Tada, M., Hamou, M. F., Janzer, R. C., Meagher-Villemure, K., Wiestler, O. D., Tribolet, N., and Van Meir, E. G. Cells with TP53 mutations in low grade astrocytic tumors evolve clonally to malignancy and are an unfavorable prognostic factor. Oncogene, *18:* 5870–5878, 1999.
12. Kononen, J., Bubendorf, L., Kalloniemi, A., Bärlund, M., Schraml, P., Leighton, S., Torhoorst, J., Mihatsch, M. J., Sauter, G., and Kalloniemi, O-P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat. Med., *4:* 844–847, 1998.
13. Albertoni, M., Shaw, P. H., Nozaki, M., Godard, S., Tenan, M., Hamou, M-F., Fairlie, W. D., Breit, S. N., Paralkar, V. M., de Tribolet, N., Van Meir, E. G., and Hegi, M. E. Anoxia induces the macrophage inhibitory cytokine-1 (MIC-1) in glioblastoma cells independently of p53 and HIF-1. Oncogene, *21:* 4212–4219, 2002.
14. Ramaswamy, S., and Golub, T. R. DNA microarrays in clinical oncology. J. Clin. Oncol., *20:* 1932–1941, 2002.
15. Menouny, M., Binoux, M., and Babajko, S. IGFBP-2 expression in a human cell line is associated with increased IGFBP-3 proteolysis, decreased IGFBP-1 expression and increased tumorigenicity. Int. J. Cancer, *77:* 874–879, 1998.
16. Plate, K. H., Breier, G., Weich, H. A., and Risau, W. Vascular endothelial growth factor is a potential tumour angiogenesis factor in human gliomas in vivo. Nature (Lond.), *359:* 845–848, 1992.
17. Flaman, J-M., Frebourg, T., Moreau, V., Charbonnier, F., Martin, C., Chappuis, P., Sappino, A-P., Limacher, J-M., Bron, L., Benhattar, J., Tada, M., Van Meir, E. G., Estreicher, A., and Iggo, R. D. A simple p53 functional assay for screening cell lines, blood, and tumors. Proc. Natl. Acad. Sci. USA, *92:* 3963–3967, 1995.
18. Nozaki, M., Tada, M., Kashiwazaki, H., Hamou, M-F., Diserens, A-C., Shinoe, Y., Sawamura, Y., Iwasaki, Y., de Tribolet, N., and Hegi, M. E. p73 is not mutated in meningiomas as determined with a functional yeast assay but p73 expression increases with tumor grade. Brain Pathol., *11:* 296–305, 2001.
19. Venables, W. N., and Ripley, B. D. (eds.). Modern Applied Statistics with S-PLUS, Ed. 3, p. 333. New York: Springer Verlag, 1999.
20. Blatt, M., Wiseman, S., and Domany, E. Superparamagnetic clustering of data. Phys. Rev. Lett., *76:* 3251–3254, 1997.
21. Levine, E., and Domany, E. Resampling method for unsupervised estimation of cluster validity. Neural. Comput., *13:* 2573–2593, 2001.
22. Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, *57:* 289–300, 1995.
23. Huang, H., Colella, S., Kurrer, M., Yonekawa, Y., Kleihues, P., and Ohgaki, H. Gene expression profiling of low-grade diffuse astrocytomas by cDNA arrays. Cancer Res., *60:* 6868–6874, 2000.
24. Ichimura, K., Bolin, M. B., Goike, H. M., Schmidt, E. E., Moshref, A., and Collins, V. P. Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. Cancer Res., *60:* 417–424, 2000.

25. Choudhuri, R., Zhang, H. T., Donnini, S., Ziche, M., and Bicknell, R. An angiogenic role for the neurokines midkine and pleiotrophin in tumorigenesis. Cancer Res., *57:* 1814–1819, 1997.
26. Feldser, D., Agani, F., Iyer, N. V., Pak, B., Ferreira, G., and Semenza, G. L. Reciprocal positive regulation of hypoxia-inducible factor 1$\alpha$ and insulin-like growth factor 2. Cancer Res., *59:* 3915–3918, 1999.
27. Rickman, D. S., Bobek, M. P., Misek, D. E., Kuick, R., Blaivas, M., Kurnit, D. M., Taylor, J., and Hanash, S. M. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res., *61:* 6885–6891, 2001.
28. Ljubimova, J. Y., Lakhter, A. J., Loksh, A., Yong, W. H., Riedinger, M. S., Miner, J. H., Sorokin, L. M., Ljubimov, A. V., and Black, K. L. Overexpression of $\alpha$4 chain-containing laminins in human glial tumors identified by gene microarray analysis. Cancer Res., *61:* 5601–5610, 2001.
29. Fathallah-Shaykh, H. M., Rigen, M., Zhao, L. J., Bansal, K., He, B., Engelhard, H. H., Cerullo, L., Roenn, K. V., Byrne, R., Munoz, L., Rosseau, G. L., Glick, R., Lichtor, T., and DiSavino, E. Mathematical modeling of noise and discovery of genetic expression classes in gliomas. Oncogene, *21:* 7164–7174, 2002.
30. Pesheva, P., and Probstmeier, R. The yin and yang of tenascin-R in CNS development and pathology. Prog. Neurobiol., *61:* 465–493, 2000.
31. Pintar, J. E., Schuller, A., Cerro, J. A., Czick, M., Grewal, A., and Green, B. Genetic ablation of IGFBP-2 suggests functional redundancy in the IGFBP family. Prog. Growth Factor Res., *6:* 437–445, 1995.
32. Yancopoulos, G. D., Davis, S., Gale, N. W., Rudge, J. S., Wiegand, S. J., and Holash, J. Vascular-specific growth factors and blood vessel formation. Nature (Lond.), *407:* 242–248, 2000.
33. Zadeh, G., and Guha, A. Neoangiogenesis in human astrocytomas: expression and functional role of angiopoietins and their cognate receptors. Front. Biosci., *8:* E128–E137, 2003.
34. Khatua, S., Peterson, K. M., Brown, K. M., Lawlor, C., Santi, M. R., LaFleur, B., Dressman, D., Stephan, D. A., and MacDonald, T. J. Overexpression of the EGFR/FKBP12/HIF-2$\alpha$ pathway identified in childhood astrocytomas by angiogenesis gene profiling. Cancer Res., *63:* 1865–1870, 2003.
35. Kerbel, R., and Folkman, J. Clinical translation of angiogenesis inhibitors. Nat. Rev. Cancer, *2:* 727–739, 2002.
36. Fuller, G. N., Rhee, C. H., Hess, K. R., Caskey, L. S., Wang, R., Bruner, J. M., Yung, W. K., and Zhang, W. Reactivation of insulin-like growth factor binding protein 2 expression in glioblastoma multiforme: a revelation by parallel gene expression profiling. Cancer Res., *59:* 4228–4232, 1999.
37. Sallinen, S. L., Sallinen, P. K., Haapasalo, H. K., Helin, H. J., Helen, P. T., Schraml, P., Kallioniemi, O. P., and Kononen, J. Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. Cancer Res., *60:* 6617–6622, 2000.
38. Elmlinger, M. W., Deininger, M. H., Schuett, B. S., Meyermann, R., Duffner, F., Grote, E. H., and Ranke, M. B. In vivo expression of insulin-like growth factor-binding protein-2 in human gliomas increases with the tumor grade. Endocrinology, *142:* 1652–1658, 2001.
39. Hoeflich, A., Reisinger, R., Lahm, H., Kiess, W., Blum, W. F., Kolb, H. J., Weber, M. M., and Wolf, E. Insulin-like growth factor-binding protein 2 in tumorigenesis: protector or promoter? Cancer Res., *61:* 8601–8610, 2001.
40. O'Shea, J. J., Gadina, M., and Schreiber, R. D. Cytokine signaling in 2002: new surprises in the Jak/Stat pathway. Cell, *109:* S121–S131, 2002.
41. Corps, A. N., and Brown, K. D. Insulin and insulin-like growth factor I stimulate expression of the primary response gene cMG1/TIS11b by a wortmannin-sensitive pathway in RIE-1 cells. FEBS Lett., *368:* 160–164, 1995.
42. Warren, R. S., Yuan, H., Matli, M. R., Ferrara, N., and Donner, D. B. Induction of vascular endothelial growth factor by insulin-like growth factor 1 in colorectal carcinoma. J. Biol. Chem., *271:* 29483–29488, 1996.
43. Bodmer, S., Strommer, K., Frei, K., Siepl, C., de Tribolet, N., Heid, I., and Fontana, A. Immunosuppression and transforming growth factor-beta in glioblastoma. Preferential production of transforming growth factor-beta 2. J. Immunol., *143:* 3222–3229, 1989.
44. Walker, P. R., and Dietrich, P-Y. Immune escape of gliomas. Prog. Brain Res., *132:* 685–698, 2001.
45. Carosella, E. D., Paul, P., Moreau, P., and Rouas-Freiss, N. HLA-G and HLA-E: fundamental and pathophysiological aspects. Immunol. Today, *21:* 532–534, 2000.
46. Wiendl, H., Mitsdoerffer, M., Hofmeister, V., Wischhusen, J., Bornemann, A., Meyermann, R., Weiss, E. H., Melms, A., and Weller, M. A functional role of HLA-G expression in human gliomas: an alternative strategy of immune escape. J. Immunol., *168:* 4772–4780, 2002.
47. Coons, S. W., Johnson, P. C., Scheithauer, B. W., Yates, A. J., and Pearl, D. K. Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary glioma. Cancer (Phila.), *79:* 1381–1393, 1997.
48. Grove, B. D., and Bruchey, A. K. Intracellular distribution of gravin, a PKA and PKC binding protein, in vascular endothelial cells. J. Vasc. Res., *38:* 163–175, 2001.
49. Alfranca, A., Gutierrez, M. D., Vara, A., Aragones, J., Vidal, F., and Landazuri, M. O. c-Jun and hypoxia-inducible factor 1 functionally cooperate in hypoxia-induced gene transcription. Mol. Cell. Biol., *22:* 12–22, 2002.

# Publication 9:

## Expression profiles of acute lymphoblastic and myeloblastic leukaemia with ALL-1 rearrangements

Authors: T. Rozovskaia, O. Ravid-Amir, S. Tillib, G. Getz, E. Feinstein, H. Agrawal, A. Nagler, E.F. Rappaport, I. Issaeva, Y. Matsuo, U.R. Kees, T. Lapidot, F. Lo Coco, R. Foa, A. Mazo, T. Nakamura, C.M. Croce, G. Cimino, E. Domany and E. Canaani. Published in: *Proc. Natl. Acad. Sci.* **24**, 5853–7858 (2003).

# Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements

T. Rozovskaia[a], O. Ravid-Amir[b], S. Tillib[c,d], G. Getz[b], E. Feinstein[e], H. Agrawal[b], A. Nagler[f], E. F. Rappaport[g], I. Issaeva[a], Y. Matsuo[h], U. R. Kees[i], T. Lapidot[j], F. Lo Coco[k], R. Foa[k], A. Mazo[c], T. Nakamura[c], C. M. Croce[c,l], G. Cimino[k], E. Domany[b,l], and E. Canaani[a,l]

Departments of [a]Molecular Cell Biology, [b]Physics of Complex Systems, and [j]Immunology, Weizmann Institute of Science, Rehovot 76100, Israel; [c]Kimmel Cancer Center, Jefferson Medical College, Philadelphia, PA 19107; [d]Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia; [e]Quark Biotech, Inc., Cleveland, OH 44106; [f]Sheba Medical Center, Tel-Hashomer 52621, Israel; [g]Children's Hospital, Philadelphia, PA 19104; [h]Fujisaki Cell Center, Fujisaki, Okayama 702-8006, Japan; [i]Institute for Child Health Research, Subiaco, Western Australia 6008, Australia; and [k]Department of Cellular Biotechnology and Hematology, University La Sapienza, 00161 Rome, Italy

Contributed by C. M. Croce, April 10, 2003

**The ALL-1 gene is directly involved in 5–10% of acute lymphoblastic leukemias (ALLs) and acute myeloid leukemias (AMLs) by fusion to other genes or through internal rearrangements. DNA microarrays were used to determine expression profiles of ALLs and AMLs with ALL-1 rearrangements. These profiles distinguish those tumors from other ALLs and AMLs. The expression patterns of ALL-1-associated tumors, in particular ALLs, involve oncogenes, tumor suppressors, antiapoptotic genes, drug-resistance genes, etc., and correlate with the aggressive nature of the tumors. The genes whose expression differentiates between ALLs with and without ALL-1 rearrangement were further divided into several groups, enabling separation of ALL-1-associated ALLs into two subclasses. One of the groups included 43 genes that exhibited expression profiles closely linked to ALLs with ALL-1 rearrangements. Further, there were evident differences between the expression profiles of AMLs in which ALL-1 had undergone fusion to other genes and AMLs with partial duplication of ALL-1. The extensive analysis described here pinpointed genes that might have a direct role in pathogenesis.**

Chromosome band 11q23 is a region of recurrent rearrangements in human acute leukemias. These rearrangements, usually in the form of reciprocal chromosome translocations, affect 5–10% of children and adults with acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML). The most common translocations are t(4;11) and t(9;11), accounting for 40% and 27%, respectively, of all 11q23 rearrangements. There is a strong association between leukemia phenotype and particular rearrangements. Thus, t(4;11) occurs nearly exclusively in ALL, and 85% of cases with t(9;11) are AMLs (1, 2). Essentially all 11q23 abnormalities involve the ALL-1 gene (also termed MLL, HRX, or HTRX), which rearranges with >30 partner genes to produce fusion proteins composed of ALL-1 N terminus and the C terminus of the partner protein (3, 4). A second and less frequent type of ALL-1 rearrangement does not involve partner genes but rather partial duplications of ALL-1 N-terminal segments (5). ALL-1-associated leukemias show some unusual and intriguing features (reviewed in refs. 6 and 7). First, they predominate infant acute leukemias, amounting to 80% of infants with ALL and 65% of those with AML. Second, they account for the majority of therapy-related (secondary) leukemias, developing in 5–15% of primary cancer patients treated with drugs, such as etoposide (VP16), that inhibit DNA topoisomerase II. Third, in infant leukemia and in therapy-related leukemia the disease arises after a brief latency. In fact, studies of monozygotic twins and newborns with leukemia and analysis of neonatal blood spots from children who were diagnosed with leukemia indicate that in most or all infant leukemias ALL-1 rearrangements occur *in utero*. The short latency suggests that ALL-1 fusion proteins induce leukemia with few, if any, additional mutations. Fourth, prognosis of patients with 11q23 abnormalities is dismal. Recent large studies indicated that <25% of infants and adults >40 years old with ALL and t(4;11), or with AML and t(9;11), were curable (1, 2, 8).

The unique biological and clinical features of 11q23-associated leukemias, in conjunction with their induction by altered versions of ALL-1, a highly intricate chromatin modifier (9), prompted us to look for molecular clues for those features by examining the expression profiles of these leukemias.

## Materials and Methods

**Patients, Specimens, and DNA Microarrays.** Apart from two individuals, all patients with 11q23 abnormalities were adults. The samples were provided by the GIMEMA Italian Multicenter Study Group. Informed consent was obtained from the patients. Also included in the analysis were four AML cell lines with t(9;11) (MONO-MAC-1, MONO-MAC-6, THP-1, and MOLM-13) and one with t(6;11) (ML-2), and two ALL cell lines with t(4;11), RS-(4;11) and B-1. Genes picked up in the supervised analysis, as well as most of those pointed out as separating ALLs with and without t(4;11) in nonsupervised analysis, had similar expression profiles in cell lines and primary tumors. The primary tumors included 12 ALLs with t(4;11) obtained from 10 adults, one child, and one infant, and 10 AMLs of adults, including 5 with t(9;11), 3 with ALL-1 partial duplication, and single cases of t(10;11) and t(11;19). Controls comprised 10 AMLs of adults, 11 ALLs of adults, and 2 ALLs of children. Details regarding the patients the sample identifies may be found in Table 2 and *Supporting Text*, which are published as supporting information on the PNAS web site, www.pnas.org. Bone marrow samples were obtained from newly diagnosed patients. The samples were composed of at least 80% blasts and were subjected to Ficoll gradient centrifugation before extraction of RNA. Details of microarray analysis may be found in *Supporting Text*.

**Preprocessing and Filtering of Data.** The expression data were organized in a matrix of $n_s = 52$ columns (hybridizations) and 12,600 rows (genes on the chip). Denote by $A_{gs}$ the "average difference" of gene $g$ in sample $s$. First, we thresholded the data; we set $T_{gs} = A_{gs}$ for sizeable values, $A_{gs} \geq 10$, and replaced low values, $A_{gs} < 10$, by $T_{gs} = 10$. Next, log was taken, $E_{gs} = \log_2 T_{gs}$, and the genes were filtered on the basis of their variation across the samples. Denote by $\bar{E}_{gs}$ the average of the $E_{gs}$ values obtained for gene $g$ over all $n_s$ tumor samples and by $\sigma_g$ their standard deviation. Only those genes that satisfied $\sigma_g > 1.1$ were studied; 3,090 genes passed this filtering procedure. After removal of non-human Affymetrix controls and genes appearing on only one of the versions of the U95A chip, we were left with 3,064 genes (of 12,600). All further analysis was done on these genes.

**Supervised Analysis.** We used supervised analysis (hypothesis testing) to identify genes, one at a time, whose expression levels can be

Abbreviations: ALL, acute lymphoblastic leukemia; AML, acute myeloblastic leukemia; FDR, false discovery rate.

[l]To whom correspondence may be addressed. E-mail: croce@calvin.jci.tju.edu, fedomany@wisemail.weizmann.ac.il, or eli.canaani@weizmann.ac.il.

used to separate tumors into two known classes $A$ and $B$ of $n_A$ and $n_B$ samples, respectively. We used the Wilcoxon rank sum test to find genes differentially expressed between the two groups of samples (e.g., AML samples with chromosome translocations versus those without; other comparisons are listed in *Results*). We used the rank sum test because it is nonparametric, i.e., it does not assume normal distribution of the data. For each gene $g$ we make the null hypothesis, according to which all $n_A + n_B$ expression levels were drawn from the same distribution. The test produces a statistic $W_g$ and a $P$ value for each gene $g$. A $P$ value of $P_g = 0.05$ means that the probability of erroneously concluding that a gene does separate the two groups is 5%, which is the standard value used in the literature. However, here we deal with multiple comparisons; at this level of $P_g$ it is expected that 150 of 3,000 random, identically distributed genes will be falsely identified as separating the two groups of samples. To control the number of false positives, we used the false discovery rate (FDR) method (10). The $N$ tested genes are ordered according to their increasing $P_g$ values, and a parameter $q$, which controls the fraction of false positives, is set. We then identify the minimal index $j$ such that for all $i > j$ we have $P_i \geq 1 \times q/N$. The null hypothesis is rejected for all genes with index $i \leq j$. This procedure yields a list of genes for which the expected fraction of false positives is $q$.

**Unsupervised Analysis: Clustering.** We used the coupled two-way clustering method (11). We assume that each group of genes is important for one particular process of interest. Thereby, the noise generated by the large majority of genes that are not relevant for that process is eliminated; furthermore, by using a group of correlated genes, the noise of the individual measurements is averaged out and reduced. The relevant subsets of genes and samples are identified by means of an iterative process, which uses, at each iteration level, *stable* gene and sample clusters that were generated at the previous step. Before each clustering operation the rows of the data matrix (genes) are centered (mean = 0) and normalized (SD = 1). The ability to focus on stable clusters that were generated by any clustering operation is essential for the coupled two-way clustering method; otherwise, there would be a computationally unfeasible number of gene/sample cluster pairs to test (11). Because most clustering methods do not have a reliable inherent stability measure for clusters, we used Superparamagnetic Clustering (SPC), a physics-based algorithm (12) that does provide a stability index, $\Delta T(C)$, to each cluster $C$. SPC was tested on data from a large number of problem areas including image analysis, speech recognition, computer vision, and gene expression (refs. 11 and 12 and references therein). A parameter $T$ controls the resolution at which the data are viewed; as $T$ increases, clusters break up and the outcome is a dendrogram. A cluster $C$ is "born" at $T = T_1(C)$, the value of $T$ at which its "parent" cluster breaks up into two or more subclusters, one of which is $C$. As $T$ increases further, to $T_2(C) > T_1(C)$, $C$ itself breaks up and "dies"; $\Delta T(C) = T_2(C) - T_1(C)$ is the stability index provided by SPC. The larger $\Delta T(C)$ is, the more statistically significant and stable (against noise in the data and fluctuations) is the cluster $C$ (13).

## Results

**Expression Profiles of ALLs with t(4;11).** Leukemic cells of ALLs with t(4;11) display features of precursor B cells with IgH rearrangements, negative for CD10 and positive for CD19, but also show some characteristics of myeloid cells (1). This and their capability to differentiate *in vitro* into monocyte-like cells had suggested that the leukemic clones originate from an early precursor cell. Hence, this leukemia is classified as pro-B cell ALL. To determine whether the expression repertoire of ALLs with t(4;11) is unique, we compared it to the transcription profiles of a set of ALL samples lacking t(4;11). These consisted of $CD10^-$ pro-B cell ALLs, Ph chromosome-positive early pre-B cell ALLs, $CD10^+$ early pre-B cell ALLs, and T cell ALLs. Supervised analysis indicated that at a FDR of 0.05 there were 130 overexpressed and 107 underexpressed genes in ALLs with t(4;11), in comparison to ALLs lacking the abnormality (Fig. 1*A*). To evaluate the consistency of the pattern, the relative expression of each gene in all of the samples was displayed in the form of bars (see examples in Fig. 1*B*). The top genes on the lists of overexpressed or underexpressed genes in ALLs with t(4;11) are shown in Table 1. The complete lists may be

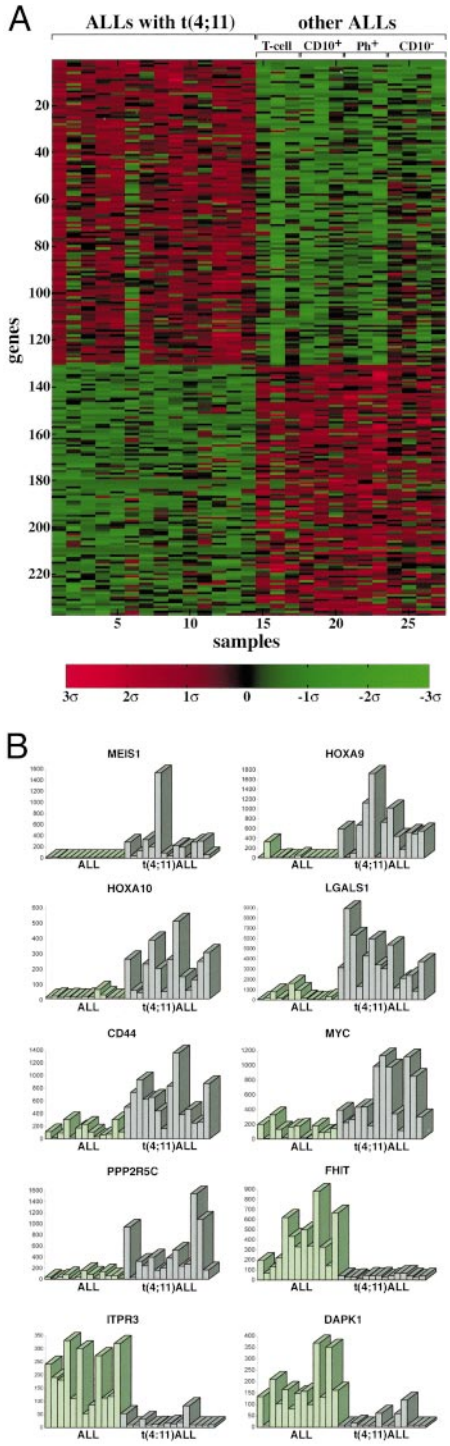



**Fig. 1.** Supervised analysis of genes distinguishing ALLs with ALL-1 rearrangement [t(4;11)] from other ALLs (*A*), and relative levels of expression of selected genes (*B*). Expression levels greater and smaller than the mean 0 are shown in red and green, respectively.

**Table 1. Genes most correlated with ALLs carrying the t(4;11) aberration, compared to other ALLs**

| No.* | Also scored as t(4;11)-specific† | GenBank accession no. | | *P* value | Fold change | Confidence interval |
|---|---|---|---|---|---|---|
| | | | Overexpressed in t(4;11) ALLs | | | |
| 1 | √ | D16532 | VLDLR, very low density lipoprotein receptor | 0.000004 | 17.51 | (10.67–28.74) |
| 2 | √ | U85707 | MEIS1, myeloid ecotropic viral integration site 1 homolog (mouse) | 0.000004 | 14.50 | (7.64–27.51) |
| 3 | √ | AC004080 | HOXA10, homeo box A10 | 0.000022 | 10.80 | (6.17–18.90) |
| 4 | √ | AI535946 | LGALS1, lectin, galactoside-binding, soluble 1 (galectin1) | 0.000024 | 23.00 | (8.82–59.98) |
| 5 | √ | M54992 | CD72 antigen | 0.000037 | 4.35 | (2.80–6.74) |
| 6 | √ | U41813 | HOXA9, homeo box A9 | 0.000041 | 20.12 | (7.22–56.09) |
| 7 | | AF098641 | CD44, CD44 isoform (Indian blood group system) | 0.000056 | 4.43 | (2.65–7.41) |
| 9 | √ | AA099265 | RECK, reversion-inducing-cystein-rich protein with kazal motifs | 0.000063 | 3.58 | (2.03–6.31) |
| 10 | √ | M14087 | HL14, $\beta$-galactoside-binding lectin | 0.000068 | 6.94 | (3.50–13.76) |
| 11 | √ | Z69030 | PPP2R5C, protein phosphatase 2, regulatory subunit B (B56), $\gamma$ isoform | 0.000069 | 7.55 | (3.60–15.83) |
| 13 | | D83767 | D8S2298E (reproduction 8) | 0.000086 | 3.16 | (2.01–4.97) |
| 14 | | AF016004 | GPM6B, glycoprotein M6B | 0.000095 | 13.06 | (5.98–28.53) |
| 15 | | X96753 | CSPG4, chondroitin sulfate proteoglycan 4 (melanoma-associated) | 0.000097 | 7.97 | (3.56–17.85) |
| 16 | √ | D78177 | QPRT, quinolinate phosphoribosyltransferase | 0.000104 | 7.31 | (3.86–13.86) |
| 17 | √ | V00568 | MYC, v-myc myelocytomatosis viral oncogene homolog (avian) | 0.000126 | 5.93 | (2.68–13.12) |
| 18 | | X61118 | LMO2, LIM domain only 2 (rhombotin-like 1) | 0.000126 | 3.85 | (2.05–7.22) |
| 20 | √ | M58597 | FUT4, fucosyltransferase 4 [$\alpha$-(1,3) fucosyltransferase, myeloid-specific] | 0.000187 | 3.52 | (2.17–5.71) |
| | | | Underexpressed | | | |
| 131 | √ | U46922 | FHIT, fragile histidine triad | 0.000010 | −8.18 | (−5.16)–(−12.97) |
| 132 | √ | U70321 | TNFRSF14, tumor necrosis factor receptor superfamily, member 14 | 0.000012 | −24.73 | (−12.05)–(−50.72) |
| 133 | √ | U01062 | ITPR3, inositol 1,4,5-triphosphate receptor type 3 | 0.000013 | −10.69 | (−6.49)–(−17.63) |
| 134 | √ | M16594 | GSTA2, glutathione *S*-transferase A2 | 0.000017 | −3.48 | (−2.27)–(−5.34) |
| 135 | √ | U03858 | FLT3LG, fms-related tyrosine kinase 3 ligand | 0.000024 | −2.24 | (−1.56)–(−3.20) |
| 136 | √ | AB007895 | KIAA0435 | 0.000037 | −4.38 | (−2.47)–(−7.77) |
| 137 | | J05257 | DPEP1, dipeptidase 1 (renal); renal metabolism of glutathione | 0.000046 | −3.24 | (−2.07)–(−5.07) |
| 138 | √ | X53586 | ITGA6, integrin $\alpha$6 | 0.000056 | −15.57 | (−6.59)–(−36.79) |
| 139 | √ | J03600 | ALOX5, arachidonate 5-lipoxygenase | 0.000056 | −4.57 | (−2.63)–(−7.94) |
| 141 | √ | L34059 | CDH4, cadherin 4, type 1, R-cadherin (retinal) | 0.000069 | −6.89 | (−3.48)–(−13.64) |
| 142 | | AF041434 | PTP4A3, protein tyrosine phosphatase type IVA, member 3 | 0.000085 | −4.11 | (−2.36)–(−7.18) |
| 143 | √ | X76104 | DAPK1, death-associated protein kinase 1 | 0.000093 | −7.90 | (−3.94)–(−15.85) |

*Gene numbers at the left match numbers in Table 3 and appear in the same order as in Fig. 1. Missing numbers (8, 12, 19, and 140) correspond to genes that were present more than once on the array and already appear in the table.

†Also included within the group of 43 (three genes appear twice) genes, associated with specific features of t(4;11) ALLs, in Fig. 2.

found in Table 3, which is published as supporting information on the PNAS web site. The clear difference in expression profiles between ALLs with the t(4;11) abnormality and other types of ALLs establishes that the former belong to a unique and distinguishable class of ALL.

Examination of the genes whose expression pattern distinguishes t(4;11) ALLs from other ALLs reveals a substantial number of genes associated with growth control, cell transformation, or malignancy. Those genes may be classified into several functional categories.

1. Overexpressed oncogenes: (*i*) HOX A9 and MEIS1, which form a sequence-specific DNA-binding complex (14) and are frequently coactivated in spontaneous AML of BXH-2 mice (15) [forced coexpression of the two genes in murine bone marrow cells rapidly induces AML (16)]; (*ii*) HOX A10, which induces AML in mice (17); (*iii*) LMO2 (RMBT2), whose overexpression, resulting from chromosome translocations, is associated with T cell ALL (18); (*iv*) MYC, which has a critical role in cell proliferation and is deregulated in human lymphomas and other tumors (19); (*v*) LGALS1 (galectin1), which cooperates with RAS in cell transformation (20) and inhibits T cell proliferation and survival (21); and (*vi*) PDGFR$\beta$ (platelet-derived growth factor receptor $\beta$), which is a tyrosine kinase and is deregulated through chromosome translocations and gene fusions in chronic myeloproliferative diseases (22).
2. Overexpressed genes involved in drug resistance: (*i*) CD44, which is associated with aggressive B-CLL (23) and conferring resistance to several widely used anticancer drugs (24); (*ii*) DHFR (dihydrofolate reductase), which confers resistance to methotrexate; (*iii*) BLMH (bleomycine hydrolase); and (*iv*) CAT (catalase), which protects from oxidative stress.
3. Overexpressed genes involved in protection from apoptosis and in survival: (*i*) CDC2 (cell division cycle 2; p34; CDK1), which preserves the viability of cancer cells in response to microtubule poisons and anticancer drugs like vincristine and taxol, by increasing expression of the apoptosis inhibitor survivin (25); (*ii*) PPP2R5C (phosphatase 2A), which is implicated in regulation of growth, transcription, and signal transduction, and is required for survival and protects from apoptosis in *Drosophila* (26); (*iii*) MAP3K5 (mitogen-activated protein kinase kinase kinase 5), which is involved in activation of the p38 MAP kinase required for initiation of the $G_2$/M checkpoint (27) and is selectively activated in non-small cell lung cancer (28).
4. Underexpressed proapoptotic genes: (*i*) ITPR3 (inositol 1,4,5-triphosphate receptor type 3), which mediates the release of intracellular calcium and consequently actively promotes apoptosis (29); (*ii*) IGFBP3 (IGF-binding protein 3), which has proapoptotic activity both dependent and independent of p53 (30); and (*iii*) JUN, which is implicated as positive modulator of apoptosis induced in hematopoietic progenitor cells of the myeloid linkage (31) [down-regulation of JUN might account for the failure of glucocorticoid therapy (32)].

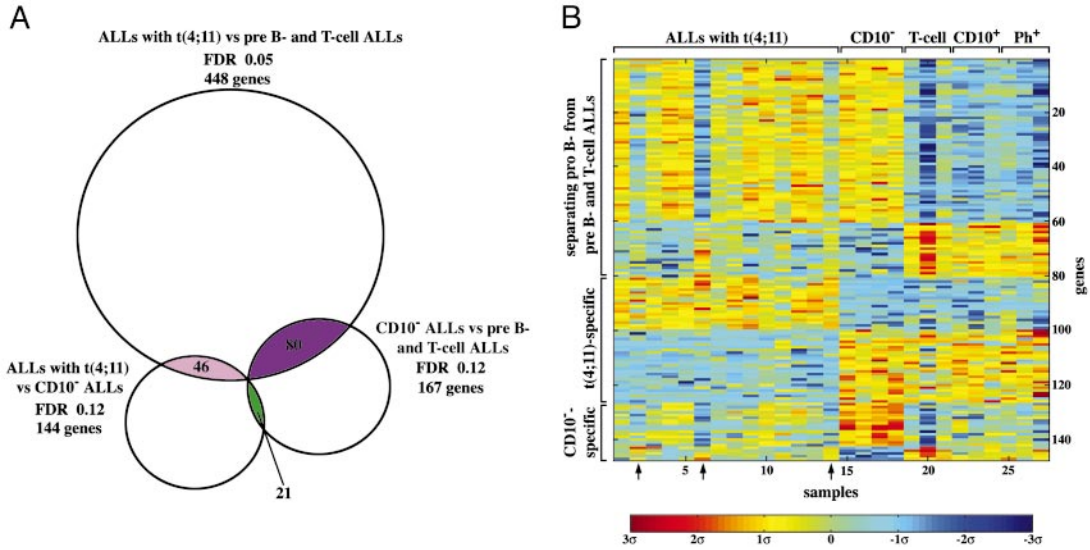



**Fig. 2.** Intersections of genes separating three types of ALLs (see text). (*A*) Three groups of genes, encompassing 77 (three appear twice), 43 (three appear twice), and 20 (one appears twice) genes, were found to participate each in two separations. (*B*) The expression matrix of these three groups. Levels of expression higher or lower than the mean 0 are shown in red/yellow and blue, respectively. Arrows point to samples with variant expression profile (see text).

5. Underexpressed tumor suppressors and growth inhibitors: (*i*) FHIT (fragile histidine triad), which is a target of chromosome aberrations and inactivated in many cancers, including lung, esophagus, stomach, breast, kidney, and leukemias (33); (*ii*) DAPK1 (death-associated protein kinase 1), mediating IFN′s activity and countering oncogene-induced transformation by activation of a p19ARF/p53 apoptotic checkpoint (34); and (*iii*) MADH1 (mothers against decapentaplegic homologue 1; SMAD1), which is a transcription modulator mutated in various forms of cancer (35).
6. Overexpressed genes acting in cell cycle progression and cell proliferation: (*i*) CCNA1 (cyclin A1), which functions in S phase and mitosis and the expression of which is elevated in a variety of tumors including AMLs (36); (*ii*) BMYB (myb-like 2), which is required for proliferation of hematopoietic cells (37) and directly activates the antiapoptotic gene ApoJ/clusterin (38); and (*iii*) CDKN3 (cyclin-dependent kinase inhibitor 3), which interacts with cyclin-dependent kinases and is overexpressed in breast and prostate cancer (39).

Our battery of ALLs lacking t(4;11) consisted of tumors at various stages of differentiation, including pre-B, pro-B, and T cell ALLs. Therefore, the differences in expression found should be due in part to the differences in differentiation stage between t(4;11) to the other ALLs. Hence, we now tried to (*i*) identify those genes whose expression pattern is directly correlated with the t(4;11) abnormality, either resulting from the abnormality or specifically associated with the cell type in which the chromosome translocation occurred; (*ii*) separate the genes above from genes whose expression reflects (sensitive to) the differences between the early and late differentiation stages (pro-B vs. pre-B and T cell tumors); (*iii*) identify genes associated with unique features of CD10$^-$ ALLs. To this end we defined three groups of ALL samples: (*i*) t(4;11) tumors (pro-B cells), (*ii*) CD10$^-$ tumors (pro-B cells), and (*iii*) the rest of the ALLs (pre-B and T cells). Three distinct supervised analyses were performed, which separate (*i*) t(4;11) ALLs from the rest of ALLs, (*ii*) t(4;11) ALLs from CD10$^-$ ALLs, and (*iii*) CD10$^-$ ALLs from the rest of the ALLs.

The genes that participate in one or more separations were identified (see the Venn diagram of Fig. 2*A*). Three overlapping groups were found, containing 77 (three genes appear twice), 43 (three genes appear twice), and 20 (one gene appears twice) genes. Lists of the genes are in Tables 4–6, which are published as supporting information on the PNAS web site. Each of these groups contained genes that are overexpressed or underexpressed; the expression matrix of the three groups is shown in Fig. 2*B*. Seventy-seven genes separate both pro-B cell t(4;11) ALLs from pre-B and T cell ALLs, as well as the latter from pro-B cell CD10$^-$ ALLs. Having been picked in both separations, this group of 77 genes distinguishes pro-B cell ALLs [both with and without the t(4;11) chromosome translocation] from pre-B and T cell ALLs. The 43 genes of the second intersection simultaneously separate t(4;11) ALLs from CD10$^-$ ALLs and from pre-B and T cell ALLs. Being singled out in both separations, this group of genes is neither associated with the differences between pro-B vs. pre-B and T cells and tumors, nor does it involve specific features of CD10$^-$ tumors. Rather, the expression of these 43 genes is affected directly by the t(4;11) abnormality and probably by other unique features of the pro-B cells in which the t(4;11) aberration occurred. The majority of these 43 genes also appear in Table 1. The last group of 20 genes separates CD10$^-$ from t(4;11) ALLs, as well as from pre-B and T cell ALLs. Being selected in both separations, these 20 genes are likely to be associated with unique features of CD10$^-$ tumors.

Inspection of Fig. 2*B* points to three t(4;11) tumors, samples 2, 6, and 14, that show a variant transcription profile. Although the expression pattern of the 43 genes, specifically correlated with t(4;11) ALLs, is similar in these three tumors and in the rest of t(4;11) ALLs (see genes 81–126 of Fig. 2*B*), the transcription profile of the three tumors with regard to genes 1–80 (which distinguish pro-B from pre-B and T cell tumors) is closer to pre-B and T cell ALLs, unlike the profile of the other 11 t(4;11) samples. The three tumors also show some quantitative variation from the other t(4;11) ALLs in transcription of the genes whose expression is associated with CD10$^-$ ALLs (genes 127–147; Fig. 2*B*). These results suggest the existence of two subfamilies of ALLs with the t(4;11) chromosome translocation, distinguished by their expression patterns.

Finally, we applied the coupled two-way clustering method (11, 12) in an unsupervised analysis. A group of 25 genes was found to be consistently underexpressed in ALLs with t(4;11) compared with the other ALLs (Fig. 3; Table 7, which is published as supporting information on the PNAS web site). The cluster of samples with low expression of these genes includes 13/14 of t(4;11) and 3/4 of CD10$^-$ ALLs. This is consistent with the close similarity in biological and clinical features between these two types of tumors. A

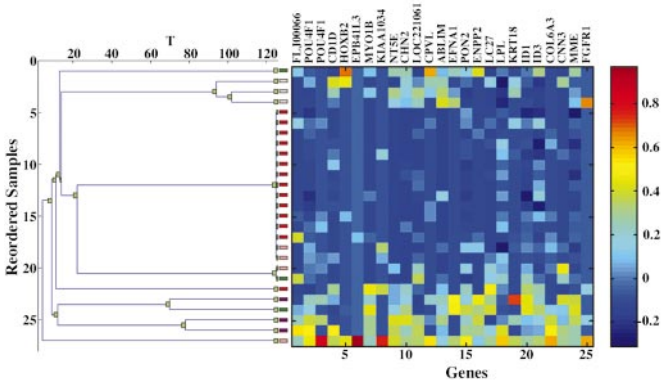



**Fig. 3.** Clustering the ALL samples on the basis of their expression levels over a cluster of 25 genes, G7 (which was obtained by the coupled two-way clustering). (*Left*) The resulting dendrogram; each leaf corresponds to an ALL sample, with t(4;11) ALLs colored red and CD10$^-$ ALLs rose. (*Right*) The expression matrix, with rows corresponding to samples and columns to genes. 13/14 of t(4;11) samples and 3/4 of CD10$^-$ ALLs are in the central cluster of samples with low expression levels.

second group of 132 genes separated the seven cell lines included in the analysis from the 45 primary tumors. All these genes were underexpressed in the cell lines (Fig. 5 and Table 8, which are published as supporting information on the PNAS web site).

**Transcription Profile of AMLs with ALL-1 Rearrangements.** AMLs with 11q23 translocations and ALL-1 rearrangements were compared in their expression profiles to AMLs with normal karyotypes. At a FDR of 0.15 (85% confidence) we identified 66 genes overexpressed or underexpressed in AMLs with 11q23 abnormalities (Fig. 4; Table 9, which is published as supporting information on the PNAS web site). Three primary AMLs with ALL-1 partial duplication (5) were compared with the other AMLs with regard to expression of the 66 genes. Two of the three tumors resembled AMLs without 11q23 abnormalities, whereas the third appeared closer to the tumors with chromosome translocations (Fig. 4). The similarity between AMLs without 11q23 aberrations and AMLs with ALL-1 partial duplications was further evidenced in the failure to separate the two groups at an acceptable FDR. (In parallel, AMLs with 11q23 abnormalities were separated from AMLs with ALL-1 partial duplications at a FDR of 0.3; some of this analysis is shown in Fig. 6 and Table 10, which are published as supporting information on the PNAS web site.) These results, if confirmed with additional samples, suggest molecular differences between AMLs triggered by recombination of the ALL-1 gene to partner genes and AMLs triggered by ALL-1 partial duplications. The variations might be reflected in biological and clinical features.

Examination of the list of genes most correlated with AMLs carrying 11q23 abnormalities (Table 9) shows that some are involved in cancer, proliferation, or apoptosis. These include the overexpressed insulin receptor, which enhances DNA synthesis and inhibits apoptosis (40), the overexpressed repair gene RAD 51, which is up-regulated in breast and pancreatic cancers (41) and probably increases drug resistance, the overexpressed PPP2R5C phosphatase, the underexpressed JUNB, which up-regulates the tumor suppressor gene p16 and represses cyclin D1 (42) and whose knockout in mice induces myeloproliferative disease (43), the underexpressed tumor suppressor FHIT, the underexpressed double stranded RNA-activated protein kinase proapoptotic PRKR, which acts in the context of IFN′s pathway and up-regulates FAS and BAX (44), and the underexpressed DEFA1 (defensin), which is involved in immune response.

Having identified genes differentially expressed in ALLs with t(4;11) compared with ALLs without t(4:11) and in AMLs with

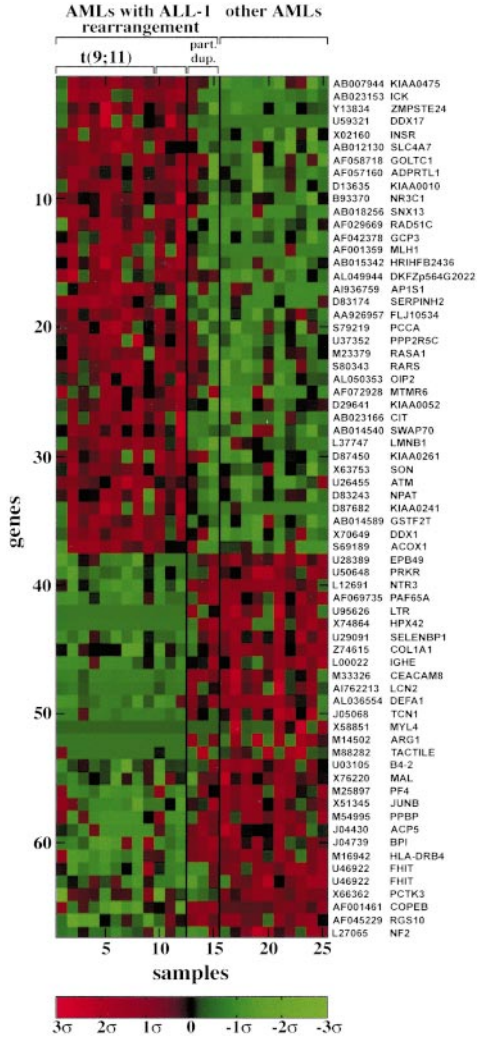



**Fig. 4.** Genes distinguishing AMLs with 11q23 chromosome translocations and ALL-1 rearrangements (samples 1–12) from other AMLs (samples 16–25). Samples 13–15 of AMLs with ALL-1 partial duplication were not included in the supervised analysis but were added later for the purpose of comparison.

11q23 abnormalities compared with AMLs without such abnormalities, we intersected the results of these two tests (we used a FDR level of 0.15 for both) to find the genes in common. We identified 50 (two appear twice) such genes that were overexpressed or down-regulated in the relevant tumors (Fig. 7 and Table 11, which are published as supporting information on the PNAS web site). For all these genes the difference was high for one type of tumors (e.g., ALLs), but modest for the second type (e.g., AMLs). The genes that were overexpressed in the samples with ALL-1 rearrangements included the phosphatase PPP2R5C and the MCM4 gene, whose product is an essential component of the prereplicative complex (45). The underexpressed genes included FHIT and JUNB.

## Discussion

Our results indicate distinct transcription profiles of ALL-1-associated tumors. This is likely to be reflected in the unusual clinical and biological characteristics of these tumors, such as short latency, poor prognosis, expression of myeloid genes in ALL, etc. Some of the genes pinpointed in our study of ALLs with t(4;11), which were mostly in adults, were also indicated (Table 3) in our

previous preliminary analysis (46) and in recent investigations that dealt with ALLs from infants and children (47, 48).

Examining the genes overexpressed or underexpressed in tumors with ALL-1 rearrangements (in particular in ALLs) indicates a constellation of expression patterns previously associated with and/or highly favorable for malignant transformation and cancer. This includes activation of oncogenes (MYC, HOX A9 and MEIS1, LMO2, etc.), inactivation of tumor suppressor genes such as FHIT and DAPK1, suppression of apoptosis by down-regulation of proapoptotic genes and up-regulation of survival genes, suppression of host immune response (up- and down-regulation of galectin 1 and defensin, respectively), up-regulation of genes conferring drug resistance, such as CD44, DHFR (dihydrofolate reductase), and BLMH (bleomycine hydrolase), and overexpression of genes involved in cell proliferation (e.g., cyclin A1 and myb-like 2). Some of the overexpressed genes we identified, like VLDL, PDGFR$\beta$ (platelet-derived growth factor receptor $\beta$), HOX A9, MEIS1, and insulin receptor, are also to be found expressed in normal hematopoietic stem cells (49) but the majority of genes are not. We suggest that at least some of the genes alluded to by our study contribute directly to the aggressive nature of the disease and to its known resistance to therapy.

In an attempt to identify genes whose expression correlates more strictly with the t(4;11) genotype, we removed genes that distinguish pro-B from pre-B and T cell tumors, as well as genes associated with the $CD10^-$ phenotype. The 43 genes left (Table 5) are closely linked to the t(4;11) genotype and would be good candidates for future biological experiments. Examination (Fig. 8, which is published as supporting information on the PNAS web site) of the expression level of the 43 genes in two other studies (47, 48) demonstrates (in particular the investigation of Armstrong *et al.*) that nearly all of these genes separate ALLs with and without ALL-1 rearrangements. Therefore, the expression profile of the 43 genes distinguishes both adults (most patients of ours) and children (the other two studies) with ALLs and ALL-1 rearrangements. Another approach taken to identify genes more likely to be associated with the pathogenesis was based on the (unproven) speculation that ALL-1 fusion proteins trigger malignancy by a similar mechanism in both ALLs and AMLs. Thus, we looked for genes that behave in similar fashion (up- or down-regulated) in ALLs and AMLs with ALL-1 rearrangements (Fig. 7). At the top of the list we find PPP2R5C, FHIT, and JUNB.

Compartmentalization of the genes into two groups whose expression distinguishes t(4;11) from other ALLs resulted in the unexpected identification of two subclasses of t(4;11) tumors (Fig. 2*B*). The subclasses are discerned by the expression profile of the 77 genes separating pro-B from pre-B and T cell tumors. Because t(4;11) tumors are generally considered to be pro-B cell ALLs, it is perplexing that, with regard to genes separating pro-B from pre-B and T cell ALLs, the smaller subclass of t(4;11) appears close to pre-B and T cell tumors. Comparison of the clinical records of the corresponding two subclasses of patients (Table 2; samples ht17, ht21, and ht27 in this table show the variant profile) indicates that in the first group there are 2/3 long-term survivors, but in the second group the outcome is worse (2/9). How widespread the distribution of t(4;11) patients into two groups is and whether there is a significant correlation with survival remain to be determined.

The supervised analysis of AMLs with ALL-1 rearrangements vs. control AMLs showed a less uniform pattern, as well as a lower number of separating genes. This suggests that the two groups of tumors are more heterogeneous. Unexpectedly, two of the three AMLs with ALL-1 partial duplications showed expression profiles resembling AML controls. The generality of this observation should be decided by analyzing additional tumors.

O.R.-A. thanks Amnon Amir for helpful thoughts and ideas. These studies were supported by National Cancer Institute Grant CA 50507 and by grants from the Israel Academy of Science, the Binational Science Foundation (U.S. and Israel), the Israel Cancer Research Fund, the Ridgefield Foundation, the Minerva Foundation, and the Germany–Israel Science Foundation.

1. Johansson, B., Moorman, A. V., Haas, O. A., Watmore, A. E., Cheung, K. L., Swanton, S. & Secker-Walker, L. M. (1998) *Leukemia* **12,** 779–787.
2. Swansbury, G. J., Slater, R., Bain, B. J., Moorman, A. V. & Secker-Walker, L. M. (1998) *Leukemia* **12,** 792–800.
3. Gu, Y., Nakamura, T., Alder, H., Prasad, R., Canaani, O., Cimino, G., Croce, C. M. & Canaani, E. (1992) *Cell* **71,** 701–708.
4. Tkachuk, D. C., Kohler, S. & Cleary, M. L. (1992) *Cell* **71,** 691–700.
5. Schichman, S. A., Canaani, E. & Croce, C. M. (1995) *J. Am. Med. Assoc.* **273,** 571–576.
6. DiMartino, J. F. & Cleary, M. L. (1999) *Br. J. Haematol.* **106,** 614–626.
7. Biondi, A., Cimino, G., Pieters, R. & Pui, C. H. (2000) *Blood* **96,** 24–33.
8. Pui, C. H., Gaynon, P. S., Boyett, J. M., Chessells, J. M., Baruchel, A., Kamps, W., Silverman, L. B., Biondi, A., Harms, D. O., Vilmer, E., *et al.* (2002) *Lancet* **359,** 1909–1915.
9. Nakamura, T., Mori, T., Tada, S., Krajewski, W., Rozovskaia, T., Wassell, R., Dubois, G., Mazo, A., Croce, C. M. & Canaani, E. (2002) *Mol. Cell* **10,** 1119–1128.
10. Benjamini, Y. & Hochberg, Y. (1995) *J. R. Stat. Soc. B* **57,** 289–300.
11. Getz, G., Levine, E. & Domany, E. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12079–12084.
12. Blatt, M., Wiseman, S. & Domany, E. (1996) *Phys. Rev. Lett.* **76,** 3251–3254.
13. Levine, E. & Domany, E. (2001) *Neural Comput.* **13,** 2573–2593.
14. Shen, W. F., Rozenfeld, S., Kwong, A., Komuves, L. G., Lawrence, H. J. & Largman, C. (1999) *Mol. Cell. Biol.* **19,** 3051–3061.
15. Nakamura, T., Largaespada, D. A., Shaughnessy, J. D., Jr., Jenkins, N. A. & Copeland, N. G. (1996) *Nat. Genet.* **12,** 149–153.
16. Kroon, E., Krosl, J., Thorsteinsdottir, U., Baban, S., Buchberg, A. M. & Sauvageau, G. (1998) *EMBO J.* **17,** 3714–3725.
17. Thorsteinsdottir, U., Sauvageau, G., Hough, M. R., Dragowska, W., Lansdorp, P. M., Lawrence, H. J., Largman, C. & Humphries, R. K. (1997) *Mol. Cell. Biol.* **17,** 495–505.
18. Boehm, T., Foroni, L., Kaneko, Y., Perutz, M. F. & Rabbitts, T. H. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4367–4371.
19. Nesbit, C. E., Tersak, J. M. & Prochownik, E. V. (1999) *Oncogene* **18,** 3004–3016.
20. Paz, A., Haklai, R., Elad-Sfadia, G., Ballan, E. & Kloog, Y. (2001) *Oncogene* **20,** 7486–7493.
21. Rabinovich, G. A., Baum, L. G., Tinari, N., Paganelli, R., Natoli, C., Liu, F. T. & Iacobelli, S. (2002) *Trends Immunol.* **23,** 313–320.
22. Cross, N. C. & Reiter, A. (2002) *Leukemia* **16,** 1207–1212.
23. Eistere, W., Hilbe, W., Stauder, R., Bechter, O., Fend, F. & Thaler, J. (1996) *Br. J. Haematol.* **93,** 661–669.
24. Fujita, Y., Kitagawa, M., Nakamura, S., Azuma, K., Ishi, G., Higashi, M., Kishi, H., Hiwasa, T., Koda, K., Nakajima, N. & Harigaya, K. (2002) *FEBS Lett.* **528,** 101–108.
25. O'Connor, D. S., Wall, N. R., Porter, A. C. & Altieri, D. C. (2002) *Cancer Cell* **2,** 43–54.
26. Li, X., Scuderi, A., Letsou, A. & Virshup, D. M. (2002) *Mol. Cell. Biol.* **22,** 3674–3684.
27. Bulavin, D. V., Higashimoto, Y., Popoff, I. J., Gaarde, W. A., Basrur, V., Potapova, O., Appella, E. & Fornace, A. J. (2001) *Nature* **411,** 102–107.
28. Greenberg, A. K., Basu, S., Hu, J., Yie, T. A., Tchou-Wong, K. M., Rom, W. N. & Lee, T. C. (2002) *Am. J. Respir. Cell Mol. Biol.* **26,** 558–564.
29. Blackshaw, S., Sawa, A., Sharp, A. H., Ross, C. A., Snyder, S. H. & Khan, A. A. (2000) *FASEB J.* **14,** 1375–1379.
30. Furstenberger, G. & Senn, H. J. (2002) *Lancet Oncol.* **3,** 298–302.
31. Liebermann, D. A., Gregory, B. & Hoffman, B. (1998) *Int. J. Oncol.* **12,** 685–700.
32. Pallardy, M. & Biola, A. (1998) *C. R. Seances Soc. Biol. Ses Fil.* **192,** 1051–1063.
33. Pekarsky, Y., Zanesi, N., Palamarchuk, A., Huebner, K. & Croce, C. M. (2002) *Lancet Oncol.* **3,** 748–754.
34. Raveh, T., Droguett, G., Horwitz, M. S., DePinho, R. A. & Kimchi, A. (2001) *Nat. Cell Biol.* **3,** 1–7.
35. Hata, A., Shi, Y. & Massague, J. (1998) *Mol. Med. Today* **4,** 257–262.
36. Yam, C. H., Fung, T. K. & Poon, R. Y. (2002) *Cell Mol. Life Sci.* **59,** 1317–1326.
37. Arsura, M., Introna, M., Passerini, F., Mantovani, A. & Golay, J. (1992) *Blood* **79,** 2708–2716.
38. Cervellera, M., Raschella, G., Santilli, G., Tanno, B., Ventura, A., Mancini, C., Sevignani, C., Calabretta, B. & Sala, A. (2000) *J. Biol. Chem.* **275,** 21055–21060.
39. Lee, S. W., Reimer, C. L., Fang, L., Iruela-Arispe, M. & Aaronson, S. A. (2000) *Mol. Cell. Biol.* **20,** 1723–1732.
40. Tseng, Y. H., Ueki, K., Kriauciunas, K. M. & Kahn, C. R. (2002) *J. Biol. Chem.* **277,** 31601–31611.
41. Maacke, H., Opitz, S., Jost, K., Hamdorf, W., Henning, W., Kruger, S., Feller, A. C., Lopens, A., Diedrich, K., Schwinger, E.& Sturzbecher, H. W. (2000) *Int. J. Cancer* **88,** 907–913.
42. Shaulian, E. & Karin, M. (2001) *Oncogene* **20,** 2390–2400.
43. Passegue, E., Jochum, W., Schorpp-Kistner, M., Mohle-Steinlein, U. & Wagner, E. F. (2001) *Cell* **104,** 21–32.
44. Gil, J. & Esteban, M. (2000) *Apoptosis* **5,** 107–114.
45. You, Z., Ishimi, Y., Masai, H. & Hanaoka, F. (2002) *J. Biol. Chem.* **277,** 42471–42479.
46. Rozovskaia, T., Feinstein, E., Mor, O., Foa, R., Blechman, J., Nakamura, T., Croce, C. M., Cimino, G. & Canaani, E. (2001) *Oncogene* **20,** 874–878.
47. Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., *et al.* (2002) *Cancer Cell* **1,** 133–143.
48. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. & Korsmeyer, S. J. (2002) *Nat. Genet.* **30,** 41–47.
49. Ivanova, N. B., Dimos, J. T., Schaniel, C., Hackney, J. A., Moore, K. A. & Lemischka, I. R. (2002) *Science* **298,** 601–604.

# Publication 10:

## Is there a Unique Gene-Expression Signature of Survival in Breast Cancer?

Authors: I. Kela, G. Getz, L. Ein-Dor, D. Givol and E. Domany. *Submitted*. (2003).

# Is there a Unique Gene-Expression Signature of Survival in Breast Cancer?

I. Kela[1], G. Getz[1], L. Ein-Dor[1], D. Givol, and E. Domany

December 27, 2003

**Predicting the metastatic potential of primary malignant tissues has direct bearing on choice of therapy. Different microarray studies yielded several gene sets whose expression profiles could predict survival [1, 2, 3]. Surprisingly, the overlap between these sets of genes is almost zero; this fact is frustrating and not understood. We elucidate this disagreement by analyzing breast cancer data from a single source [1], thereby eliminating variability due to using different technologies on different patients. The good predictor of outcome reported by Van't Veer et al. uses 70 genes whose expression was most correlated with survival over a particular training set of patients. Our analysis shows that one could have easily singled out several alternative sets of 70 genes, that could have been used to construct classifiers with the same success rate as those of Van't Veer et al. This result extends the list of 70 to a much longer one, which includes known prognostic markers. The main implication is that the identities of the top 70 genes are fairly arbitrary; by focusing only on these, one can miss important key players.**

Predicting survival of cancer patients in general, and of breast cancer patients in particular, is of great importance in choosing an appropriate treatment. Use of microarray analysis to find gene expression profiles related to survival has become a popular tool to address this problem. The two main statistical strategies adopted in this field are unsupervised and supervised

[1]These authors contributed equally.

analysis. In the former, tumors are partitioned into subtypes, based on expression similarities between the samples. Then the relation of each subclass to survival is checked, and survival prospects are assigned to tumors based on their subclass affiliation. Implementing this approach using hierarchial clustering [3], breast carcinoma tissues were classified into five different subtypes, each with a distinctive expression profile. The clustering analysis was based on two sets of genes; 456 "intrinsic" genes [4] and a list of 264 cDNA clones, that exhibited high correlation with patient survival. Clustering the samples based on these two sets of genes yielded the same five subtypes, which were found to be informative with respect to survival. A recent study [5] has demonstrated the robustness of these survival related subclasses, by applying the same analysis procedure on two independent breast carcinoma data sets.The second strategy uses supervised approaches to identify genes associated with survival, regardless of having different subtypes in the set of studied samples. Using this method, Van't Veer et al. [1] constructed a classifier capable of predicting disease outcome. The genes, whose expression levels were used by the classifier were selected on the basis of their correlation with survival. 231 correlated ($|\text{coeff}| > 0.3$) genes were identified, rank-ordered and optimized by cross validation (using a leave-one-out method), yielding an optimally separating list of 70 genes. The predictions of the classifier were tested, and an error rate of 10% (2/19) was reported. A follow-up study proved the efficiency of this classifier as survival predictor, by testing its performance on a large set of 295 tumor specimens, with an error rate of 18.6% (55/295). These proved feasibility of predicting disease outcome on the basis of gene expression profiles. Another work by Ramaswamy et al. [2] identified a set of 128 genes separating metastatic from primary tumors. A refined set of 17 metastases associated genes, while tested on a large set of primary solid tumors from six diverse types, were found to well distinguish between good and poor prognosis patients.

The success of these studies may be interpreted as indicating that we may be close to identifying *the* genes whose expression levels determine outcome. Frustratingly, the sets of survival related genes identified by the three studies mentioned above have only a few in common. The overlap between the 456 of Sorlie et al. and the 231 of van't Veer et al. is 17 genes, whereas the intersect between Sorlie's set and Golub's set is two. Only one gene appears in both Van't Veer's and Golub's sets.[2] In this work we explore this surprising

[2]The listed comparisons included only genes that have a symbol.

phenomenon, and suggest an explanation for the lack of agreement between the sets of genes.

**Many genes are related to survival.** Obviously, the lack of agreement between predictive lists obtained by different groups can be attributed to reliance on different chips, different methods of sample preparation, mRNA extraction and analysis of the data and, most importantly - to genuine differences between the patients (tumor grade, stage etc). In order to eliminate as many of the biological and technological sources of variation, we chose to focus on data from one experiment [1]. The data consist of 96 samples and 5852 genes (see Methods). Patient $i$ is assigned a survival index $s_i = 1$ if a metastasis free time interval (MFTI) of more than 5 years (after first diagnosis) has been recorded ( $s_i = 0$ otherwise) (see Methods and [1]). Survival is represented by a 96-component binary vector, and each gene - by 96 expression values. A general picture of the gene expression "world" and its relation to survival is presented in the "globe" shown in Fig 1. The (normalized) survival vector $\mathbf{s}$ is assigned to the north pole, and two additional genes, BUB1 (known to be negatively correlated with survival) and ESR1 (positively correlated with survival) also reside on the surface, at positions ($\mathbf{b}$ and $\mathbf{e}$) that reflect their correlations to survival and between themselves. All other genes are placed according to their relation to these three vectors (see Methods). The 5852 genes comprise an oblate spheroid shaped cloud, tilted with respect to the equator, extending from near $\mathbf{e}$ to $\mathbf{b}$. This is a striking geometrical manifestation of the fact that the expression vectors of very many genes (1234 - at an FDR of 10%, see Methods) are related to survival. If we use a random survival vector and repeat the procedure that generated Fig. 1, the genes form a similar oblate spheroid, but placed at the equator of the "globe" - with no tilt. (see Fig. **??**, Supplementary Information). The distribution of correlations with the random survival is significantly narrower than that obtained for the real data (see Fig. 1).

According to our model, if the experiment is repeated on a different group of patients (with the same clinical characteristics), the overall appearance of the new "globe" will be quite similar, but the positions of individual genes will swarm around. This swarming suffices to change drastically the relative ranking of the genes and to allow a different group of non survival-related genes to "sneak" into the list of correlated ones by chance. In what follows, we substantiate this model and explain its consequences.

**Small number of patients causes a gene's rank to fluctuate.** Say we measure the correlation $r$ of a gene's expression with survival on the basis

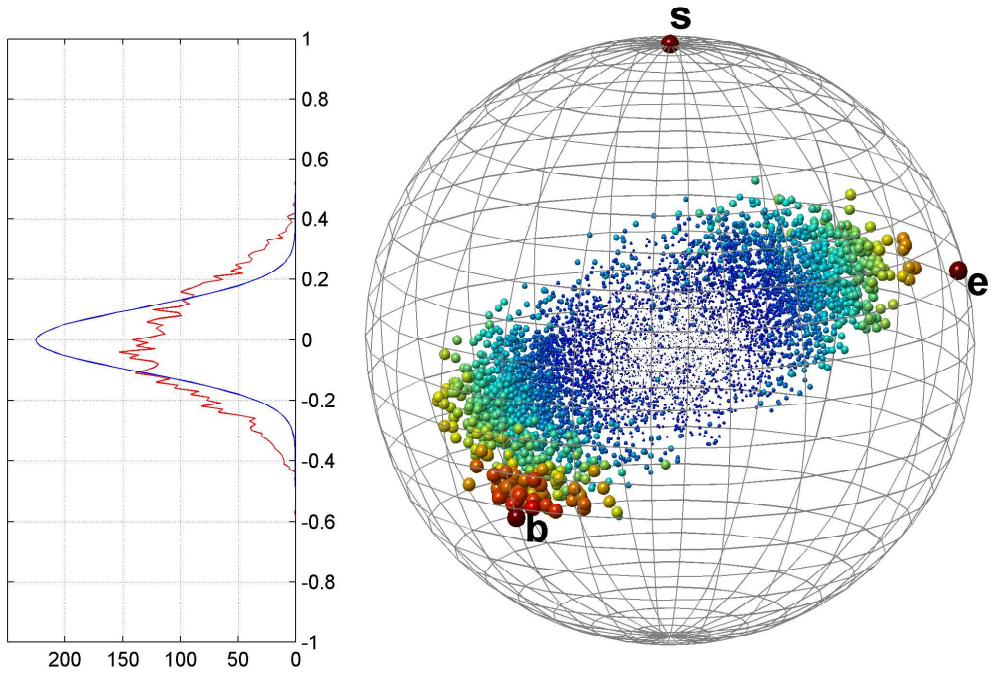



Figure 1: Globe of genes in the "world" spanned by survival **s**, BUB1 **b** and ESR1 **e**. Survival is located at the north pole while BUB1 and ESR1 are on the sphere's surface and their relative locations are determined by their angles with survival and with each other. All other genes were also normalized (in 96 dimensions), their projections on **s,b,e** were calculated and then used to position the genes. Each gene is represented by a spot whose size and color illustrate how close the gene is to the surface (large red spots are close and small blue are far). Genes at the vicinity of BUB1 have a negative correlation with survival, whereas genes around ESR1 have a positive correlation. The genes create an elongated structure at an angle $< \pi/2$ with **s**, implying that a large fraction of genes exhibit a non-vanishing correlation with survival. Left hand side: the red curve is the histogram of the genes' correlation with the real survival vector (i.e. the gene's projection onto the vertical **s** axis), while the blue curve is the histogram of the genes' correlations with a random permutation of the survival vector.

of a sample of $N$ patients drawn at random from a group with similar clinical characteristics. If a different set of $N$ is drawn, we expect the correlation to be different, due to the effect of the finite size of our sample. If these statistical fluctuations are sizeable, they may cause a gene of high ranked correlation in one sample of $N$ patients to drop to a low rank in another sample. The smaller $N$ the larger will be the fluctuation of $r$. In order to estimate the effect of these fluctuations on the composition of gene lists such as those of [1], we repeatedly selected different subgroups of 77 samples out of the 96 (in each group we maintained the overall good/poor prognosis ratio) and for each subgroup identified the 70 genes that have the highest correlation with survival. The significant variation of the membership of the top 70 genes is clearly shown in Fig. 3 of the Supplementary Information. Note that every pair of these subgroups have at least 58 samples in common, which significantly reduces the fluctuations of $r$ and variation of the genes' ranks. In spite of this, the average overlap between two such gene groups is only 33.7/70. To estimate the "true" fluctuations of $r$ for independent subgroups of 77 we used bootstrapping [6] in which the subgroups are drawn from the 96 with repeats (see methods). This reduces the expected overlap of two top-70-gene lists to 12.2/70. Figure 2 shows how the ranking by correlations with survival, measured over ten such subgroups, varies; genes high ranked over one subgroup are likely to become low ranked (around 800) in another. Consequently, even when the same technology is used, analysis of gene expression, measured over different sets of 77 patients drawn from a clinically similar pool of patients, will yield different lists of genes as top ranked with respect to correlation with survival.

**Many sets of genes can be used to predict survival.** Van't Veer et al. ranked the genes according to their correlation with survival, measured across 78 training samples. They found that the expression levels of the top 70 genes attain the best performance in predicting good or poor prognosis. Two of their predictions on the test-set of the remaining 19 samples were wrong. This is probably an under-estimate of the error rate since in a later study [7] they tested their classifier on 295 different patients and obtained a higher overall error rate of 18.6% . Note that such an error rate is equivalent to about 4 errors in a test set of 19 samples. In light of the fact that many genes are correlated with survival, and that the differences between those correlations are rather small, one would expect that many sets of genes could perform similarly. We tested this hypothesis in the following manner. We selected the same 77 patients as [1] and ranked all genes according to their

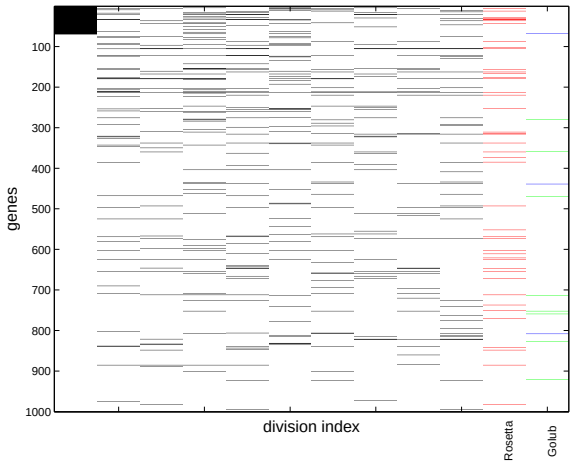



Figure 2: Distribution of the top 70 genes, determined by correlation of gene expression with survival, measured over 10 randomly chosen subgroups of $N = 77$ patients (using bootstrapping). Each row represents a gene and each column - a subgroup. Genes are ordered according to their correlation with survival over the first subgroup (from top to bottom). In each column (based on a different patient subgroup) the top 70 ranked genes are colored black. The genes that were top ranked over one subgroup can have a much lower rank when other subgroups are used to measure the correlation with survival. The two rightmost columns mark the genes that appeared in Golub et al.'s [2] list of 128, and the 70 genes chosen by Van't Veer et al. [1]

correlation with survival, as measured over this training set. In the spirit of the classifier established by van't Veer et al., we built a series of classifiers, each based on a different group of 70 genes; the first classifier used the top 70 genes, the second - the genes ranked 71-140, etc. For each classifier we measured the training and the test error, and found seven other sets of 70 genes (that appeared way down in the correlation-ranked list) producing classifiers with the same prognostic capabilities as those based on the top 70. The location of those 7 sets in the globe (Fig. 1) are shown in Fig. 1 of the supplementary information. The Kaplan-Meier plots corresponding to these 7 classifiers are shown in Fig. 2 of the Supplementary Information. Note that the data set is characterized by correlation between ER$\alpha$ status and disease outcome. To ensure that the additional classifiers we found, do discriminate between good and poor prognosis, and not between (ER) positive and negative tumors [8], we reduced the data set to 68 samples (see methods), eliminating correlations between ER$\alpha$ status of the tumors and clinical outcome of the patients. Applying the seven classifiers to the reduced data set, their predictive performance was not hurt, indicating that the classifiers succeed in predicting outcome where ER fails [9]. The classifier gene sets generate a long list of genes, pointing to the wide survival impact on gene expression, rather than reflecting the obvious influence of ER$\alpha$ status on the gene expression profile.

Since a gene is ranked differently when it's correlation with survival is measured on different subgroups of $N = 77$ patients (see Fig. 2), one expects the resulting training and test errors to be subgroup dependent as well. To control this effect, we repeated the procedure described above for 1000 different subdivisions of 77/19 samples; for each subdivision we ranked the genes and measured the training and test errors for each of the classifiers (using top 70 genes, next 70 and so on). Figure 3 shows the training and test errors for a single "experiment" and also the average of 1000 such curves (the grey areas are the 95% confidence intervals). The average test error increases very slowly as the genes used for the classifier become less correlated with survival, while the average training error exhibits a slightly higher sensitivity. Yet, the difference between the average number of training errors of the classifier that uses the top 70 genes and the one constructed from genes ranked around 1000 is less than 1.5. This implies that not only the first classifier (the one suggested by van't Veer et al), but also much lower ranked classifiers are capable of predicting survival with a similar quality. To give a quantitative meaning to this claim we measured, for each of the 1000 training sets, the

number of classifiers for which *both* training and test errors were at least as low as those of the first classifier. The distribution of the number of such classifiers is exhibited in the inset of Figure 3. The results show that more than 70% of the 1000 training sets produced at least one classifier with the same (or better) performance than the one based on its own top 70 genes, and the average number of such classifiers is 4. The surprising summary of these observations is that (a) the list of "top 70 genes" of highest correlation with survival depends strongly on the training set of (77) patients on which the correlation was measured and (b) even with a fixed training set one could have easily singled out a different set of 70 much lower ranked genes with similar prognostic performance.

In this work we showed that small sample size is a major factor in explaining the inconsistency between published lists of genes related to survival. Many genes are related to survival, with statistically significant, but not very high correlations, that decrease very slowly with rank. A possible biological scenario behind this is the following: even within a clinically homogenous group of patients the individual variations and heterogeneities associated with markers for outcome are large. Perhaps one has to divide the patients into smaller subgroups on the basis of some yet unknown attribute [3], and then for each subgroup of tumors one will be able to find it's much sought "primary master genes" that control metastatic potential. Such a master gene will have very high correlation with survival *in it's own subgroup* and, possibly, very low correlation in other subgroups. The large fluctuations in the correlation of such a gene's expression with survival, measured over small samples, are due to the fluctuating fraction of how many members of the gene's subgroup are in the sample. It is important to note that such a master gene will not necessarily be top-ranked with respect to correlation measured in a very large sample of patients, composed of a mixture of subgroups. Our picture implies that one needs much larger numbers of patients to identify such survival-wise-homogenous subgroups and their associated master genes. Hence one should separate two issues: the quest for survival-related master genes and the construction of prognostic tools on the basis of a short list of genes. One can produce fairly reliably prognostic tools; many genes are related to survival, and using a large enough subset of them will compensate for the fluctuations in the predictive power of individual genes for individual patients. Membership in a prognostic list, however, is not necessarily indicative of the gene's importance for cancer pathology. Rather, in order to study the biology one must scan the entire, wide list of

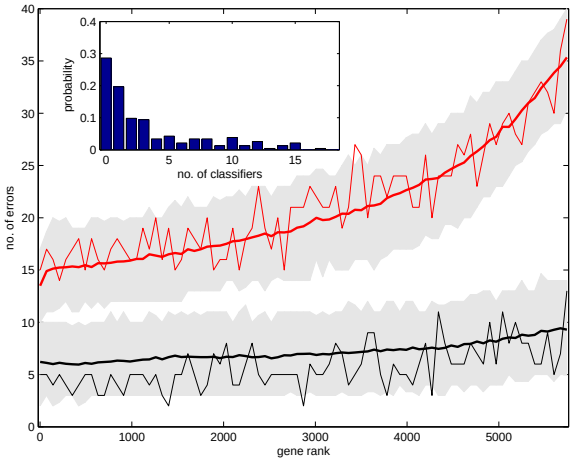



Figure 3: 1000 partitions of the patients into a training set of 77 and test set of 19 were generated. Over each training set the correlation with survival was calculated for every gene and the genes were ranked accordingly, as indicated on the horizontal axis. Consecutively ranked groups of 70 genes were used to construct a classifier of short/long survival. The fluctuating thin red and black curves present, for a single partition, the number of errors over the training and test sets, produced by classifiers, versus the rank of the 70 genes on which the classifier was based. The average of 1000 such curves is plotted as thick red and black lines; the grey areas around these two curves are the 95% confidence interval of the training and test errors. Inset: histogram of the number of classifiers whose training and test errors are at least as low as those of the first classifier based on the 70 genes with highest correlation to survival, obtained for 1000 partitions. Note that more than 70% of the training sets produce at least one classifier with the same performance as the top 70; the expected number of such classifiers is around 4.

survival-related genes. By focusing only on those genes, that were singled out by one group as it's preferred prognostic tool, one may miss important key players. The picture we obtained for breast cancer may emerge also in other types of cancer.

## Methods

### Public data-set

The data [1] contain gene expression profiles of primary breast tumors, from 96 sporadic young patients with grade $T1/T2$ tumors less than 5cm in size, and N0 (no lymph node metastases). 34 of the 96 sporadic patients were treated by modified radical mastectomy and 62 underwent breast-conserving treatment, including axillary lymph node dissection followed by radiotherapy. Hybridization ratios were measured with respect to a reference made by pooling equal amounts of cRNA from all the sporadic carcinomas, on microarrays containing 25,000 human genes [10].

### preprocessing of data

The full expression matrix of van't Veer et al had 24481 rows (genes) and 117 columns (samples). We applied a filtering criteria yielding 5852 genes that exhibited two-fold change of expression with a p-value less than 0.01 in 5 or more samples. We discarded from the set a single sample (sample 54) which contained more than 20% missing values. We based our analysis on 96 'sporadic' patients free of $BRCA1/2$ germline mutations.

### correlation analysis

For each gene we test the null hypothesis that its gene expression profile is uncorrelated with the survival vector (over all 96 samples). Correction for multiple comparisons was performed using the False Discovery Rate (FDR) method [11]. Bounding the expected false discovery rate by 10% yielded a list of 1234 genes for which the null hypothesis can be reject.

### Eliminating correlation between ER$\alpha$ status and outcome

First, the 96 samples were divided into 61 (ER) positive tumors and 35 (ER) negative tumors. Then each of those groups was divided into two subgroups: good prognosis patients and poor prognosis patients. Since there are more good prognosis than poor prognosis patients in the (ER) positive group (39 compared to 22), we randomly chose 22 from the 39 good prognosis

samples such that the new (ER) positive group contained an equal number, 22, of good and poor prognosis patients. We repeated this procedure for the (ER) negative group, by selecting at random 12 among the 23 poor prognosis patients, thus reducing their number to 12, the number of poor prognosis (ER) positive patients. The resultant dataset was composed of 44 (ER) positive and 24 (ER) negative tumors, and had no correlation between ER$\alpha$ status and disease outcome. The new training and test sets were the old training and test set without the samples that were removed from the data set.

**Creating the globe figure**

Gene $i$ is represented as a centered and normalized unit vector in a 96 dimensional space; $\langle \hat{g}_{ij} \rangle_j = 0$, $\|\hat{\mathbf{g}}_i\|^2 = 1$. We identify two specific genes; BUB1, $\hat{\mathbf{b}} = \hat{\mathbf{g}}_{\text{BUB1}}$ and ESR1, $\hat{\mathbf{e}} = \hat{\mathbf{g}}_{\text{ESR1}}$. The 96-dimensional survival vector, $\mathbf{s} = (s_j)$, is a binary vector indicating whether patient $j$ has an MFTI (metastasis free time interval) greater or equal to 5 years. Next, we center and normalize $\mathbf{s}$ to obtain $\hat{s}$. The correlation between any two unit vectors, either a gene and survival or two genes, is simply the cosine of the angle between the vectors which is calculated by their dot-product; e.g. $\text{Corr}(\text{BUB1}, \text{survival}) = \hat{\mathbf{b}} \cdot \hat{\mathbf{s}}$. In order to generate the "globe" figure (Fig. 1) we generated a set of three orthogonal unit vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$, creating a space on which the projection of $\hat{\mathbf{s}}$, $\hat{\mathbf{b}}$ and $\hat{\mathbf{e}}$ maintain their unit length and thus will reside on the surface of the "globe". Moreover, we require that $\hat{\mathbf{z}} = \hat{\mathbf{s}}$ in order that the north pole and the survival vector will coincide. This is performed using a Gram-Schmidt orthonormalization method [12]. In general, the projections of all other genes, $\hat{\mathbf{g}}_i$, on this space fall inside the unit sphere. The coordinates of any gene in this "globe" can be used to calculate its correlations with the chosen vectors; survival, BUB1 and ESR1. Specifically, the latitude which can be easily read from Figure 1, is the angle between a gene and survival.

**Dividing the data into ten different division of 77/19**

To examine how different experiments of 77 samples, influence the composition of the 70 most correlated genes with survival, we used the bootstrapping method [6]. Bootstrapping is a computer simulation enabling to overcome finite size effects. It assumes that the sample is a good approximation of the population. Generating a large number of new samples from the original sample set, enables to estimate the statistics parameters of the population. To keep the good/poor prognosis ratio of the original training set (33/44) we

divided the 96 samples into a poor prognosis set of 45 samples, and a good prognosis set of 51. We chose with repetitions a random set of 33 samples from the poor prognosis set, and 44 from the good prognosis. We repeated this procedure ten times and found the top 70 genes for each 'training set' composition.

**Generating the distribution of the training and the test errors** We checked how the classifier, established by van't Veer et al. performs on different sets of 70 genes, as described in the paper.

# References


[1] van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, and Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[2] Ramaswamy S, Ross KN, Lander ES, and Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet*, 33:49–54, 2003.

[3] Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, and Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98:10869–10874, 2001.

[4] 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, and Botstein D. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[5] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, and Botstein D. Repeated observation of

breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100:8418–8423, 2003.

[6] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, USA, 1993.

[7] van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, and Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347:1999–2009, 2002.

[8] Gruvberger SK, Ringner M, Eden P, Borg A, Ferno M, Peterson C, and Meltzer P. Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res*, 5:2326, 2002.

[9] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Bernards R, and Friend SH. Expression profiling predicts outcome in breast cancer. *Breast Cancer Res*, 5:57–58, 2003.

[10] Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, and Linsley PS. Expression profiling using microarrays fabricated by an inkjet oligonucleotide synthesizer. *Nat Biotechnol*, 19:342–347, 2001.

[11] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, 57:289–300, 1995.

[12] Gil Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley-Cambridge Press Box 812060 Wellesley MA 02482, 2003.

# Supplementary Information

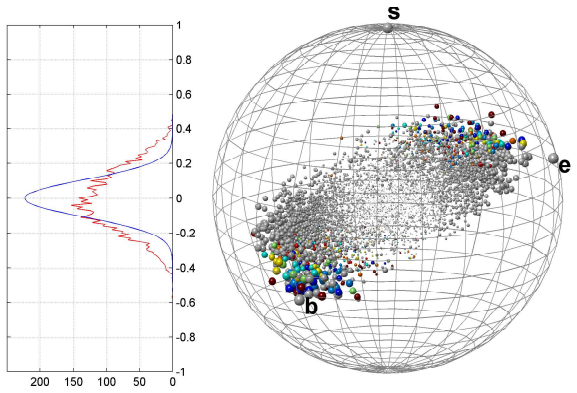



Figure 1: Genes' location in the world created by Survival (S), ER1 (e) and BUB1 (b). Grey spheres represent genes not appearing neither in van't Veer et al.'s list nor in the 6 alternative lists presented in the paper. Colored spheres denote location of genes belonging to the extended list of survival related genes, including van't Veer et al.'s gene (red), and the 6 alternative classifier gene sets (each denoted by a different color - orange to blue). The sphere size indicates how close it is to the surface. Large spheres are closer than small ones.

**Kaplan Meier analysis of the additional seven classifiers**

To demonstrate the efficiency of the seven alternative classifiers in discriminating between poor and good prognosis patients, we carried out a univariate Kaplan-Meier analysis with time to development of distant metastasis as a variable. The analysis was performed on all 96 patients. Each of the classifiers divided the patients into those with good prognosis signature, and those with poor prognosis signature. As shown in Fig. 2, in all seven classifiers, the probability of remaining disease free is significantly higher in patients classified as having good prognosis signature. The Kaplan-Meier plots of the seven classifiers appear aside to those of van't Veer et al. (right top corner), showing that their classification performance is at the same level as van't Veer et al.'s classifier.

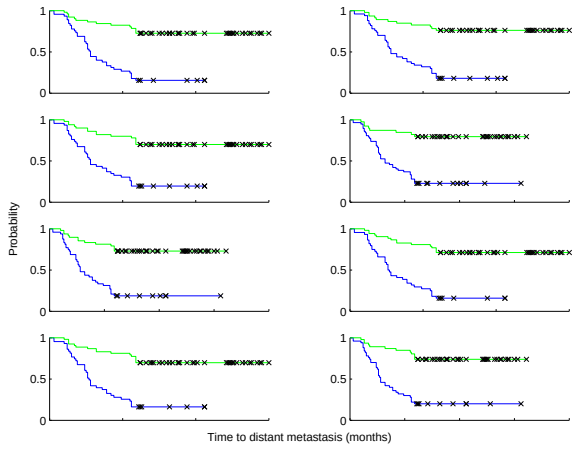



Figure 2: Kaplan-Meier analysis of van't Veer et al.'s classifier ( top right curves), and of the seven alternative classifiers (the other curves) as obtained from classifying all 96 samples. Green lines describe the probability of remaining free of metastasis in the group of samples classified as having good prognosis signature, while the blue lines describe the poor prognosis group. All p values are smaller then $10^{-7}$

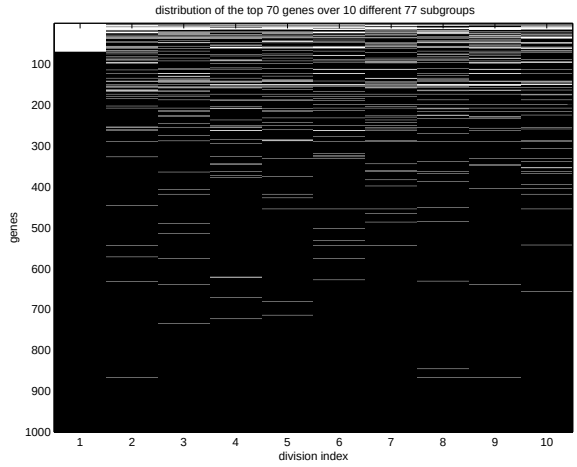



Figure 3: Distribution of the top 70 genes, determined by correlation of gene expression with survival, measured over 10 randomly chosen subgroups of $N = 77$ patients. Each row represents a gene and each column - a subgroup. Genes are ordered according to their correlation with survival over the first subgroup (from top to bottom). In each column (based on a different patient subgroup) the top 70 ranked genes are colored black. The genes that were top ranked over one subgroup can have a much lower rank when other subgroups are used to measure the correlation with survival.

# Chapter 4

# Protein Structures Applications

## 4.1 Introduction

This Chapter serves as an introduction to Publications (11), (12) which deal with prediction of protein fold assignment using structural similarity measures. This work is a continuation of my previous work (several topics from my M.Sc. [111]).

One of the main outcomes of the genome project is the list of $\sim$ 30000 genes in the human DNA. The goal of the next step is to assign functions to those genes (and proteins) and to understand their network of interactions. A key-step in understanding the function of a protein is to know its three-dimensional structure. Predicting the three-dimensional structure of a protein from its primary sequence (coded in the DNA) is a long-standing challenge named the *protein folding* problem. This challenge has not been overcome yet but slow and continuous progress is being made [159]. Recently, a new protein was designed to fold into a certain structure [160] which demonstrates the growing knowledge in this field.

In order to assist to assign a function to a predicted fold and to help building models for possible folds one needs to study the fold space of proteins. Three-dimensional structures of 20000 proteins have already been solved by means of X-ray crystallography or NMR. These are stored in the Protein Data Bank (PDB) [161] and are used to study the different protein structures. There are different methods to organize the proteins, all with a common purpose, to classify the proteins into classes with similar properties. Since during evolution protein structures are much more conserved than sequences and functions [162], proteins are usually classified first by their structural similarity (phenetic classification) and then by the similarity of their sequences or by the similarity of their functions (phylogenetic classification) [163].

The three most widely used databases that organize the known protein structures in a hierarchical classification are SCOP [164], CATH [165] and FSSP [166]. Each group has its own way to compare and classify proteins using a different mix of automatically calculated similarity scores and visual inspection. On one extreme, FSSP is based on a fully automated structure comparison method that calculates pairwise similarity scores

(Z-scores) between proteins. CATH uses a structural similarity score for lower levels of the hierarchy and manual inspection for higher ones. On the other extreme, SCOP is constructed manually, by visual inspection and comparison of not only structures but also sequences and functions. In our work we first showed that these three classification schemes are consistent and then, presented an automated procedure to assign CATH and SCOP classifications to proteins whose FSSP score is available. As the FSSP database is updated weekly, this method makes it possible to update also CATH and SCOP with the same frequency. To make our predictions available to the structural biology community we have set up a website that accepts the name of a protein (that appears in FSSP) and returns the predicted SCOP and CATH classifications. Recently we updated the web site with the latest versions of FSSP, CATH and SCOP.

Since our aim is to classify proteins that appear in FSSP and not in CATH or SCOP based on the ones which are classified, we are faced with a problem of partial labels which calls for semi-supervised methods (see Sec. **??**). Here we use a heuristic method to find the ground state, the only configuration and hence the typical one at $T = 0$, of the inhomogeneous Potts ferromagnet with external fields that was introduced in Section **??**. Each protein corresponds to a Potts spin with the number of states, $q$, set to be equal to the number of known fold (or architecture) types. The interactions between the spins are taken as the similarity Z-scores, $J_{ij} = Z_{ij}$, where $i$ and $j$ represent proteins. Spins of known classification are assigned a state according to their type and are frozen to that state. Therefore, the free spins feel a field generated by neighbors of known type. To predict the assignment of the unknown spins we search for the configuration with lowest energy of this Potts Hamiltonian. We developed a heuristic method to find a low energy configuration which for some cases is guaranteed to yield the global minimum.

## 4.2 FSSP to CATH and FSSP (F2CS)

Publication (12) describes the F2CS server which is based on the algorithm described in Publication (11). The site is available at
`http://www.weizmann.ac.il/physics/complex/compphys/f2cs/` and is constantly updated based on the latest versions of FSSP and Dali Database (a new database that will replace FSSP [167]) which supply the Z-scores and the new proteins for which we report our predictions, and CATH and SCOP which provide the training set – those proteins for which the classification is known. In the current version, the predicted set consists of 4014 single domain chains that do not appear in CATH and 511 ones which are new to SCOP.

Figure 4.1 illustrates the prediction process for CATH chains by depicting the all-against-all Z-score matrix among the union of the training and predicted sets. The training set consists of 3750 single domain chains from CATH that were chosen as representatives in FSSP/DD and the predicted set contains 4014 chains which are present in FSSP but not yet in CATH. In figure 4.1, the first 3750 are ordered according to their known CATH classification whereas the order of the 4014 is according to our predicted classification. One

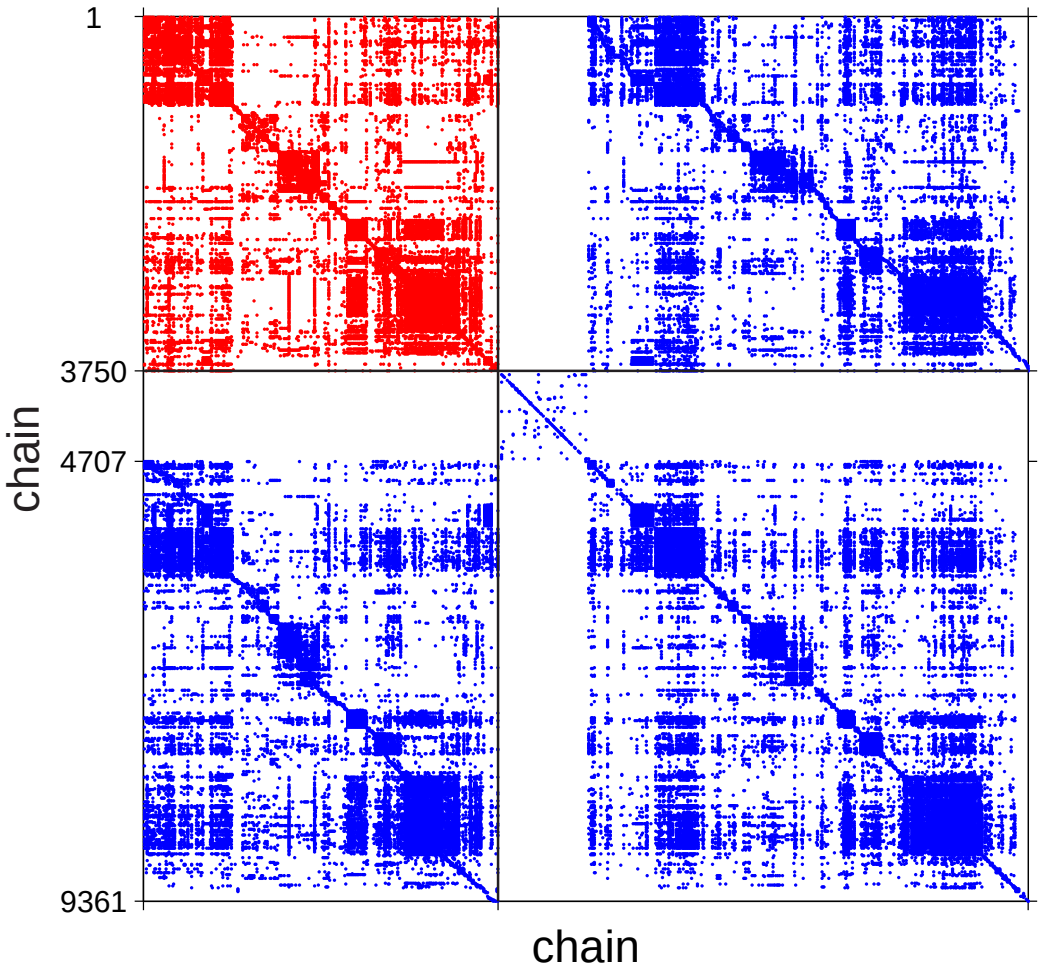



Figure 4.1: All-against-all Z-score matrix among the 9361 protein chains; The first 3750 are FSSP/DD representatives with known CATH classification and the remaining are those for which the classification is predicted. The order of the chains is according to their CATH classification whether it is known or predicted. The structure among the ones with known classification is propagated to the predicted set using our CO classification method.

can see that the original order in the training set (upper left submatrix colored in red) is propagated by our assignment procedure to the predicted set. For 957 chains, ranging from 3751 to 4707, we do not have any prediction since there is no path of neighboring chains, i.e. chains with $Z \geq 2$, extending from them to a chain of known classification and therefore no class information can permeate to them.

In the latest version of the site we added a capability to insert a Z-score file of a new structure and obtain the classification for it. A crystallography or NMR group which solves a new three-dimensional structure can input its coordinates to the DALI server [168] and obtain its Z-scores with the FSSP/DD representative set. One can insert these Z-scores to our server and it provides our predicted CATH and SCOP classification of the new protein structure.

In Publication (11) we demonstrate that, as a side product, the algorithm can highlight misclassification in the database by identifying those proteins which are classified differently from the majority of their neighbors. This can help fix errors in the databases.

# Publication 11:

## Automated assignment of SCOP and CATH protein structure classification from FSSP scores

Authors: G. Getz, M. Vendruscolo, D. Sachs and E. Domany
Published in: *PROTEINS: Structure, Function, and Genetics* **46**, 405–415 (2002).

PROTEINS: Structure, Function, and Genetics 46:405–415 (2002)
DOI 10.1002/prot.1176

# Automated Assignment of SCOP and CATH Protein Structure Classifications From FSSP Scores

**Gad Getz,[1] Michele Vendruscolo,[2] David Sachs,[3] and Eytan Domany[1]***
[1]*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel*
[2]*Oxford Centre for Molecular Sciences, New Chemistry Laboratory, Oxford, United Kingdom*
[3]*Department of Physics, Princeton University, Princeton, New Jersey*

***ABSTRACT*** **We present an automated procedure to assign CATH and SCOP classifications to proteins whose FSSP score is available. CATH classification is assigned down to the topology level, and SCOP classification is assigned to the fold level. Because the FSSP database is updated weekly, this method makes it possible to update also CATH and SCOP with the same frequency. Our predictions have a nearly perfect success rate when ambiguous cases are discarded. These ambiguous cases are intrinsic in any protein structure classification that relies on structural information alone. Hence, we introduce the "twilight zone for structure classification." We further suggest that to resolve these ambiguous cases, other criteria of classification, based also on information about sequence and function, must be used. Proteins 2002;46:405–415.**


**Key words: protein structure; protein databases; CATH; FSSP; SCOP; classification; clustering**

## INTRODUCTION

The first step to analyze the vast amount of information provided by genome sequencing projects is to organize proteins (the gene products) into classes with similar properties. Because during evolution protein structures are much more conserved than sequences and functions,[1] proteins are usually classified first by their structural similarity (phenetic classification) and then by the similarity of their sequences or by the similarity of their functions (phylogenetic classification).[2]

A reliable structural classification scheme is useful for several reasons. Perhaps the most exciting perspective is the possibility to routinely assign a function to newly identified genes.[3] This goal may be achievable because a classified database provides a library of representative structures to perform prediction of protein structure by homology[4,5] or by threading,[6–8] and it allows for the identification of distant evolutionary relationships.[9] In addition, given a particular protein, it provides a tool to identify other proteins of similar structure and function.[10] The knowledge of the structure helps to reveal the mechanism of molecular recognition involved in catalysis, signaling, and binding[2] and may lead to the rational design of new drugs.[11] At a more abstract level, the physical principles dictating structural stability of proteins are revealed by their folded state. Therefore, most of the recently proposed methods to derive energy functions to perform protein fold predictions rely in different ways on structural data.[12,13]

The most comprehensive repository of three-dimensional structures of proteins is the Protein Data Bank (PDB).[14] The number of released structures is increasing at the pace of about 50 per week, and >12,000 complete sets of coordinates were available at the time of writing. Many research groups maintain web-accessible hierarchical classifications of PDB entries. The most widely used are FSSP,[15] CATH,[16] SCOP,[17] HOMSTRAD,[18] MMDB,[19] and 3Dee[20] (see Table I for a list of abbreviations). Here we consider three of these: the FSSP, the CATH, and the SCOP databases. Each group has its own way to compare and classify proteins; these three classification schemes are, however, consistent with each other to a large extent.[21,22]

### FSSP Database

The FSSP (Fold classification based on Structure-Structure alignment of Proteins) uses a fully automated structure comparison algorithm, DALI (Distances ALIgnment algorithm),[23,24] to calculate a pairwise structural similarity measure (the S-score) between protein chains.

The algorithm searches for that amino acid alignment between the two protein chains that yields the most similar pair of $C_\alpha$ distance maps. In general, the more geometrically similar two chain structures are, the higher their S-score is. The mean and standard deviations of the S-scores obtained for all the pairs of proteins are evaluated. Shifting the S-scores by their mean and rescaling by the standard deviation yield the statistically meaningful Z-scores.

For classification of structures, the FSSP uses the Z-scores for all pairs in a representative subset of the PDB. A fold tree is generated by applying an average-linkage hierarchical clustering algorithm[25] to this all-against-all

Grant sponsor: Minerva Foundation; Grant sponsor: Germany-Israel Science Foundation; Grant sponsor: US-Israel Science Foundation (BSF).

*Correspondence to: Eytan Domany, Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: fedomany@wicc.weizmann.ac.il.

Received 23 February 2001; Accepted 13 July 2001

**TABLE I. Abbreviations and Definitions**

| Abbreviation | Definition |
|---|---|
| 3Dee | Database of protein domain definitions |
| ASTRAL | The ASTRAL compendium for sequence and structure analysis |
| CATH | Protein structure classification |
| CO | Classification by optimization |
| DALI | Protein structure comparison by alignment of distance matrices |
| DHS | Dictionary of homologous superfamilies |
| FSSP | Fold classification based on structure-structure alignment of proteins |
| HOMSTRAD | Homologous structure alignment database |
| MMDB | Molecular modeling database |
| PDB | Protein data bank |
| SCOP | Structural classification of proteins |
| SSAP | Structure comparison algorithm |

Z-score matrix. An alternate classification based on a more common four-level hierarchy is also available.[24]

### CATH Database

Orengo and coworkers use a combination of automatic and manual procedures to create a hierarchical classification of domains (CATH).[16] They arrange domains in a four-level hierarchy of families according to the protein class (C), architecture (A), topology (T), and homologous superfamily (H). The class level describes the secondary structures found in the domain[26] and is created automatically. There are four class types: mainly-α, mainly-β, α-β, and proteins with few secondary structures (FSS). The architecture level, on the other hand, is assigned manually (using human judgment) and describes the shape created by the relative orientation of the secondary structure units. The shape families are chosen according to a commonly used structure classification (e.g., barrel, sandwich, roll, etc.). The topology level groups together all structures with similar sequential connectivity between their secondary structure elements. Structures with high structural and functional similarity are put in the same fourth-level family, called *homologous superfamily*. Both the topology and homologous superfamily levels are assigned by thresholding a calculated structural similarity measure (SSAP) at two different levels, respectively.[27,28] The CATH database has been recently linked to the Dictionary of Homologous Superfamilies (DHS) database,[29] which allows further analysis of structural and functional features of evolutionary related proteins. There is a growing need for annotating proteins classified in structural databases because structural genomic initiatives are providing a large number of new proteins whose function might be gathered by distant homology informations.

### SCOP Database

The Structural Classification of Proteins (SCOP)[17] database is organized hierarchically. The lower two levels (family and superfamily) describe near and distant evolutionary relationship, the third (fold) describes structural similarity, and the top level (class) describes the secondary structure content.[26] SCOP is linked to the ASTRAL compendium,[30] which provides a series of tools for further analysis of the classified structures, mainly through the use of their sequence. At variance with FSSP and CATH, SCOP is constructed manually, by visual inspection and comparison of not only structures but also sequences and functions.

### Automated Assignment of SCOP and CATH Classifications

In this work we present a method, Classification by Optimization (CO), to predict without human intervention the SCOP fold level and the CATH topology level from the FSSP pairwise structure similarity score. A protein for which the Z-score is available is classified into a SCOP fold and into a CATH topology by the CO method, an optimization procedure that finds the assignment of minimal cost, where the cost is defined in terms of Z-scores (see Materials and Methods). The query for the classification of any such protein can be submitted to the web site.[31]

## RESULTS

### Consistency of the FSSP, CATH, and SCOP Classifications

We found that the FSSP and CATH databases are consistent.[21] In this section we show that SCOP is also consistent with these to a large extent (see also Ref. 22). In the rest of this work we use this fact to derive an automated procedure to assign the CATH and SCOP classifications starting from the FSSP Z-scores (which are updated weekly) in a fully automated fashion to include new releases in the PDB.[1] Here we further discuss the consistency of the three classification schemes by introducing concepts and quantities that are later used in the prediction of the CATH and SCOP classifications.

We first illustrate the correlation between the FSSP similarity score and the CATH classification. A simple and visually appealing way to study this problem is shown in Figure 1. The element $Z_{ij}$ of the Z-score matrix [Fig. 1(a)] represents the score for superimposing structure $i$ with structure $j$ of the set PFrCs (a subset of the proteins in FSSP and CATH, see Table III and Materials and Methods) using the DALI algorithm.[23,24] In Figure 1(a) only the pairs with $Z > 2$ are shown; therefore, the matrix is sparse and the proteins are ordered in a random fashion. Figure 1(b) is produced by reordering the rows and columns of the original Z-score matrix [Fig. 1(a)]. The reordering is performed according to the CATH classification in the following way: for each of the proteins in this set we have the CATH classifications at all levels. First, we order the proteins by their class; within the class, by the architecture; within it by the topology, and so on. This reordering generates a permutation of the columns and rows of the Z matrix. The solid black grid in Figure 1(b) separates the proteins according to their CATH class, and a thin grid is placed at the boundaries between architectures.

Figure 1(b) shows the underlying order behind the apparent randomness of Figure 1(a) and reveals the extent

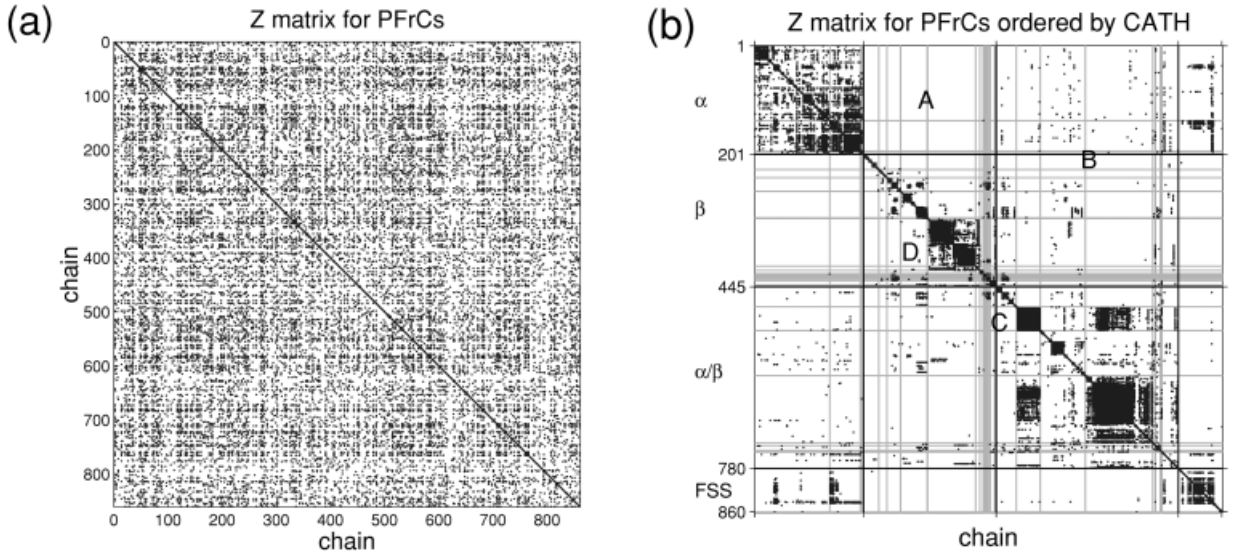



Fig. 1. **a:** Z-score matrix between all pairs of proteins in the PFrCs set. A black dot represents $Z > 2.0$. **b:** Same Z-score matrix with rows and columns rearranged by using the CATH classification (see text). Part (b) shows the underlying order behind the apparent randomness of part (a) and illustrates the extent to which the FSSP Z-scores reflect the CATH classification. The regions A, B, C, and D are discussed in the text.

to which the FSSP Z-scores reflect the CATH classification.

Several interesting observations can be made. First, consider the Class level of CATH. As can be seen in Figure 1(b), there are no matrix elements with $Z > 2.0$ in region A that connect proteins of the mainly-α class to the mainly-β class. At variance with this, some proteins from both of these classes have large Z-scores with proteins from the α-β class (region B). This is reasonable, because of the way similarity is defined by FSSP; a mainly-α protein can have a high Z-score with an α-β protein because of high similarity with the α part. Second, in the Architecture level, we observe that there are architecture families that are highly connected within themselves, e.g., α-β barrels (482–525: region C), whereas for others the intrafamily connections are more sparse. The similarities within the mainly-β sandwich family (318–406: region D) have two relatively distinct subgroups, which suggest an inner structure corresponding to the lower levels in the CATH hierarchy. Checking the topology level (the third CATH level) for this architecture, one indeed finds two large topology subfamilies, the immunoglobulin-like proteins (324–366: upper left part of region D) and the Jelly-Rolls (373–402: lower right part of region D), which correspond precisely to the two strongly connected subgroups that appear in Figure 1(b).

We found that the CATH classification at the level of topology is reflected in the Z-matrix. This is to be expected because the Z-score measures the structural similarity of two aligned proteins while preserving their connectivity. Overall, this analysis shows that the Z-matrix is correlated with the CATH classification. In a similar way it is possible to show that the Z-score is correlated with the SCOP classification. The results are available at the web site.[31]

These findings suggest that Z-scores can be used to predict the CATH and SCOP classifications of yet unclassified proteins. In what follows, we demonstrate that this indeed can be done. We also estimate the success rate of our predictions and provide a web site[31] that can be used to retrieve our predictions for the CATH topology and the SCOP fold for new entries in FSSP.

We also verified that the CATH and SCOP classifications are to a large extent mutually compatible. An immediate consequence of this is that it is possible to construct a "translation table," $\hat{T}$, from the proteins that have already both a CATH and a SCOP classification. In this way, given a CATH entry, one can obtain the corresponding SCOP classification (see Fig. 2). Row $i$ of the table refers to a particular CATH topology and column $j$ to a particular SCOP fold. The element $\hat{T}_{ij}$ of the table is the measured fraction of times that a protein has a CATH topology $i$ and a SCOP fold $j$. This number is calculated by enumerating all the 10,197 single-domain proteins with known CATH and SCOP classifications (PCsSs), and it is an estimate of $T_{ij}$, the joint probability distribution for a protein to have CATH topology $i$ and SCOP fold $j$. If the CATH and SCOP classifications had been independent, every element $T_{ij}$ could have been expressed as a product of $C_i$, the fraction of proteins that belong to CATH topology $i$, and $S_j$, the fraction that belongs to SCOP fold $j$, that is, $T_{ij} = C_i * S_j$. Randomly placing 10,197 proteins using such a probability distribution yields $4780 \pm 40$ nonzero elements in the matrix. In the other extreme case, if there had been a full correspondence between the SCOP and CATH classifications, the table would have had a single nonzero element in each row and column (in each CATH topology row the nonzero element would have been in that SCOP fold column that corresponds to it). In this case, the proteins in PCsSs would have been distributed among 284

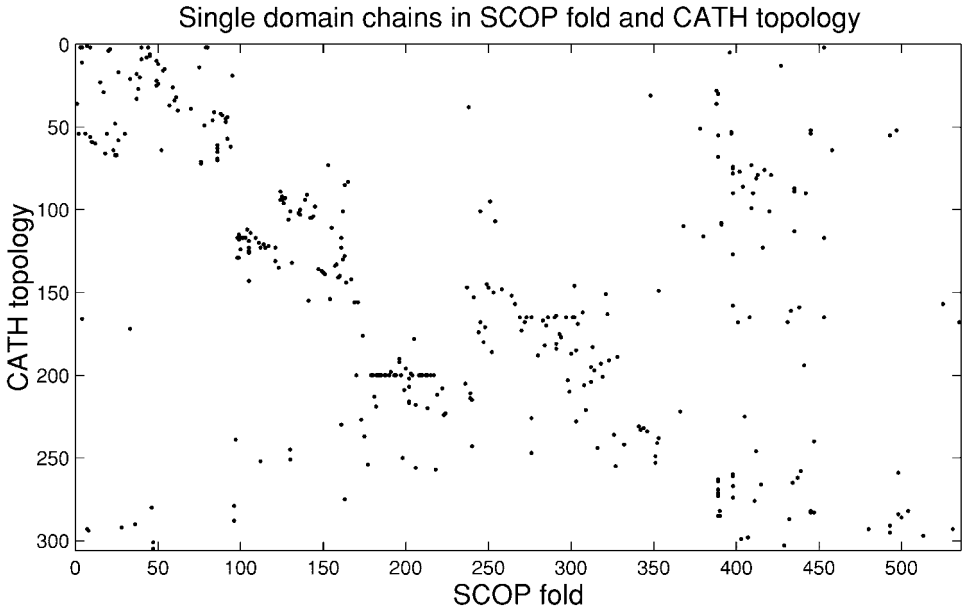



Fig. 2. Translation table from the CATH topology to the SCOP fold and vice versa. Nonzero entries of $\hat{T}_{ij}$ appear as black dots. $\hat{T}_{ij}$ is proportional to the number of proteins of CATH topology *i* that have a SCOP fold *j* in PCsSs.

nonzero elements (the number of distinct CATH topologies in PCsSs).

We found 369 nonzero elements in $\hat{T}$, meaning that the CATH and SCOP classifications are highly dependent. Still, the correspondence is not entirely one-to-one; in general, more than one SCOP fold corresponds to a given CATH topology. The number of such folds is, however, typically small. Such a translation table may be used to predict the SCOP classification of a structure already classified in CATH or at least to significantly restrict the number of possibilities and vice versa. For example, the assignment of the CATH topology to a protein with known SCOP fold can be done by selecting the CATH topology with the largest value in the translation table for that particular SCOP fold. Such an assignment is correct in 93% of the cases. The corresponding assignment of the SCOP fold from the CATH topology is correct in 82% of the cases. Although this is possibly useful information, in this work we do not assign classifications in this way.

## SUMMARY OF THE COCLASSIFICATION PERFORMANCE

Every time the FSSP Z-scores are updated (once a week) the CO classification can be applied to all the proteins that appear in the new FSSP release but are not yet classified in CATH or in SCOP. The possible outcomes of the classification procedure are as follows:

1. Correct classification: the predicted classification will agree with the future release of the databases.
2. Rejection: the program is unable to classify the structure.
3. Ambiguous classification: a classification is returned (both for CATH and SCOP), but a later release provides a different classification.

The frequencies of these outcomes greatly depends on the statistics of the set of proteins to be classified. More specifically, rejected proteins are of two types: proteins that do not have high Z-scores with any other proteins ("islands"; see Materials and Methods) and clusters of proteins that are similar among themselves but do not have high Z-scores with other proteins outside their cluster ("superislands"). The fraction of islands and superislands is a feature of the particular set of proteins to be classified. The occurrence of a superisland suggests that a new classification type (a new topology for CATH and and new fold for SCOP) might be needed. The work of maintaining CATH and SCOP can be thus focused on the classification of a representative from each of these superislands.

For the set PFCs, the fraction of islands and superislands is 5%. We used this set to provide an upper bound for the performance of the CO method (see below); however, for the set PF$\bar{\text{C}}$ the fraction of rejections goes up to 22%. If rejections are not counted, we classify correctly 98% of the PFCs proteins. On the other hand, we could test our predictions also against the new CATH release v2.0. Of 1582 proteins that were assigned to previously existing CATH topologies, CO has classified correctly 80%. The difference in success rates between PFCs and PF$\bar{\text{C}}$ is due to the different way in which the test set is nested in the larger set of structures with known classification. In the first case, the test set consisted of 20% of the members of PFCs, selected at random; the remaining 80% were used to "predict" the classification of the test set. In the second case, the members of CATH v1.7 were used to predict the classification of the new proteins that were added when

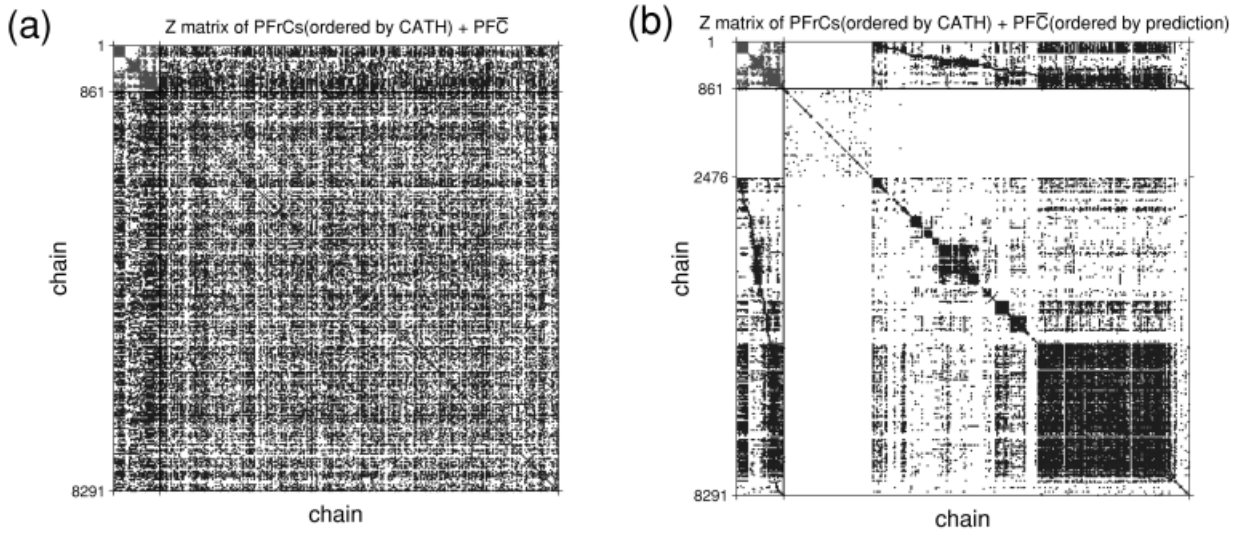



Fig. 3. **a:** Z-score matrix between all pairs of proteins in the combined PFrCs + PFC̄ sets. The submatrix in the upper left corner is the reordered Z-score matrix of the set PFrCs, which was already shown in Figure 1(b). The rest of the matrix presents the Z-scores for the proteins in the set PFC̄. **b:** The same matrix as in (a) with the rows and columns relative to the proteins in PFC̄ reordered according to our assignment of their CATH topology. With the CO method, the original order in the submatrix PFrCs is propagated to the entire matrix.

CATH v2.0 was released. These new structures are not distributed uniformly at random among the members of CATH v1.7.

Ambiguous classifications are due to two different mechanisms. The first stems from a well-known problem with the way the FSSP similarity index is calculated (the "Russian doll effect"; see below). The second kind of "mistake" is actually not a wrong classification; rather, it happens when the newly classified structure lies within the ambiguous "twilight zone" between two closely related topologies (for CATH) or folds (for SCOP), as demonstrated in detail below.

### Automated Assignment of CATH Classification From FSSP

In this section we describe the procedure that we used to predict CATH topology level from the FSSP scores. We identified a set of 7431 proteins (PFC̄; see Materials and Methods) that appear in FSSP but were not yet processed by CATH 1.7. Our goal is to predict the CATH topology of these 7431 proteins by using (a) the Z-scores between all proteins in PF (see Materials and Methods) and (b) the known classifications of the set PFrCs (see Materials and Methods).

Predicting topologies is a classification problem that we treated with pattern recognition tools. We tested several prediction algorithms using cross-validation to estimate their performance.[21] Every one of the algorithms that were tested can be viewed as a two-stage process. In the first stage, a new similarity measure is produced from the original Z-scores. This is done either by a direct rescaling of the original Z-scores or by using the results of various hierarchical clustering methods to produce new similarity measures. The second stage consists of using these similarities as the input to some classification method, yielding predictions for the classes and architectures. In this work we present only results obtained by one particular method (CO), which uses the original Z-score as a similarity measure (see Materials and Methods). A complete list of the results obtained by using other methods can be found in Ref. 21, which is available on the web site.[31]

Our final assignments for the set PFC̄ using the CO method are listed in the web site. A more illustrative way to present these results is shown in Figure 3. In Figure 3(a) we present the Z-score matrix for the combined set PFrCs + PFC̄. The submatrix in the upper left corner is the reordered Z-score matrix of the set PFrCs, which was already shown in Figure 1(b). The rest of the matrix in Figure 3(a) presents the Z-scores of PFrCs with the set PFC̄ (randomly ordered) and the Z-scores of PFC̄ among themselves. In Figure 3(b) we reordered the rows and columns whose index was >860, corresponding to proteins in PFC̄. Although in the matrix of Figure 3(a) these proteins appear in a random order, in Figure 3(b) they appear in the order imposed by our prediction of their CATH topology. One can see that the original order in the submatrix PFrCs is propagated by our assignment procedure to the set PFC̄. For example, focus on the small black square at the upper left corner of the matrix. This small black square represents the high Z-scores among the mainly-α class of proteins in PFrCs. In the corresponding top rows of the full matrix we see high Z-scores between these structures and some proteins from PFC̄. In particular, the small group with indices near 2476 are "close" to these mainly-α structures and hence are also classified as such. On the other hand, there is a large group of structures from PFC̄ (between 861 and 2476), which do not have high Z-scores with any of the proteins in PFrCs or with any of the other structures in PFC̄ with index >2476. Hence,

we are unable to classify this group of structures on the basis of their FSSP scores.

Figure 3(b) illustrates the central idea of this work. We perform a task that is intermediate between clustering and classification. We take proteins of known classification and we use them as fixed a priori values in a clustering procedure.

The overall success rate of our prediction estimated by cross-validation was 93%. To understand the significance of these success rates, we derived a statistical (see Materials and Methods) upper bound for this kind of prediction. This upper bound is 95% (see Materials and Methods), hence the figure of 93/95 = 98% given above.*

We estimated the accuracy of the prediction by using the following procedure. First, the set PFCs was randomly "diluted"; that is, we randomly chose a certain fraction of the proteins in PFCs and placed them in a test set, pretending that we did not know their classification. The FSSP scores of the entire set were then used to classify the test set. For each protein from the test set, we either return a predicted classification or reject the protein (i.e., we declare that we are unable to classify it). The quality of any classification algorithm (see Materials and Methods) is measured by its success rate (fraction of correctly classified proteins, out of the test set) and by the purity (success rate out of the nonrejected proteins). For the CO method, the results were 93% for the success rate and 98% for the purity (using a dilution of 20%). More extensive tests at other dilutions and for other methods are of classification are discussed in Ref. 21 and available at the web site.[31]

We also tested directly the reliability of the CO assignments by using the CATH version 2.0 (PC2). In PC2, 1640 single-domain proteins that are present in $PF\bar{C}$ were assigned to one of the topologies that existed in v1.7. Fifty-eight of these we "rejected." In 1266 cases of the remaining 1582 (80%), our prediction agrees with the one given in CATH v2.0. Almost all the cases in which we misassigned a domain can be explained in a simple way. These cases are discussed in detail in a following section.

The CO method can also be used to predict directly the C level and the A level of CATH. We found that when the C and A levels were predicted as a byproduct of predicting the T level, the resulting C and A were consistent with those predicted directly.

### Automated Assignment of SCOP Classification From FSSP

We used the CO method to predict the SCOP fold for a set of 3451 proteins ($PF\bar{S}$) that belong in PF but not yet in PS. The results are available on the web site.[31] The estimated success rate (by cross-validation) was 93%. As in the case of CATH, this number increased when we discarded proteins in the "twilight zone" (see the next section).

### Twilight Zone for Protein Classification

The attempt to assign a new protein to a known fold might lead to frustration because at times one is undecided about two or more possibilities. To assess that two proteins have similar structures, a similarity score is needed. FSSP uses the Z-score, CATH uses the SSAP score, and SCOP uses a subjective evaluation, which is also a kind of score. The problem arises when the protein to be classified has high scores with two proteins already classified, but to different topologies. In this article, these proteins are called *borders* (see Materials and Methods). Being a border protein depends on the similarity score. We showed, however, that FSSP, CATH, and SCOP are to a large extent consistent classifications. Therefore, we suggest that there are "intrinsically" ambiguous cases—cases that are unavoidable in structure comparison. We refer to these ambiguous regions in structure space as the "twilight zone" in analogy with the case of protein sequence comparison where proteins with sequence similarity below 30% cannot be reliably assigned to the same fold. We illustrate this concept by a typical case, shown in Figure 4. This is a border protein. Protein 1dhn (the central one) is the one to be classified (in fact, it is a three-layer sandwich according to CATH). It has a Z-score of 9.3 with protein 1a8rA (on the left), which is a three-layer sandwich topology and a Z-score of 8.7 with protein 1b66A (on the right), which is a two-layer sandwich topology. This example illustrates how structural information alone might not provide a clear-cut criterion for classification of this protein. The incidence of the twilight zone is shown in Figure 5. In Figure 5(a) we present the histogram of the number of protein pairs that have different CATH topologies as a function of their Z-score. This number is a rapidly decaying function of Z. On the contrary, the number of pairs with the same CATH topology is a slowly decaying function of Z. For $Z > 3$, the probability of having the same CATH topology becomes greater than that of having different topologies. For $Z > 7.5$, the probability to have the same topology is 97.5%. In Figure 5(b) we show the corresponding figure for SCOP. The number of folds in SCOP is larger than the number of topologies in CATH; therefore, there is more ambiguity. However, also in this case for $Z > 7.5$, the probability to have the same topology is 93.5%. Taken together, these results indicate that the twilight zone for structure comparison can be bound by $Z \leq 7$.

There are other cases in which the classification of a particular protein is inconsistent with that of all its neighbors. For example, proteins that we called *colonies* (see Materials and Methods) are such that none of their neighbors are of their own kind. This means that the FSSP scores imply that these proteins are similar only to proteins of different classes and architectures. Identifying these proteins can also focus the attention to possible misclassification or to drawbacks of the Z-score. For example, 1 of the 49 colonies (at the architecture level) that

* One must keep in mind that the estimated success rate is calculated for all proteins; both FSSP representatives (≈10% of the proteins) and nonrepresentatives. Because the presence of homologous proteins can create a bias in these estimates, we also tested the success rate of predicting the CATH topology only for the FSSP representatives, which yielded 63%, to be compared with the corresponding upper bound of 74%.

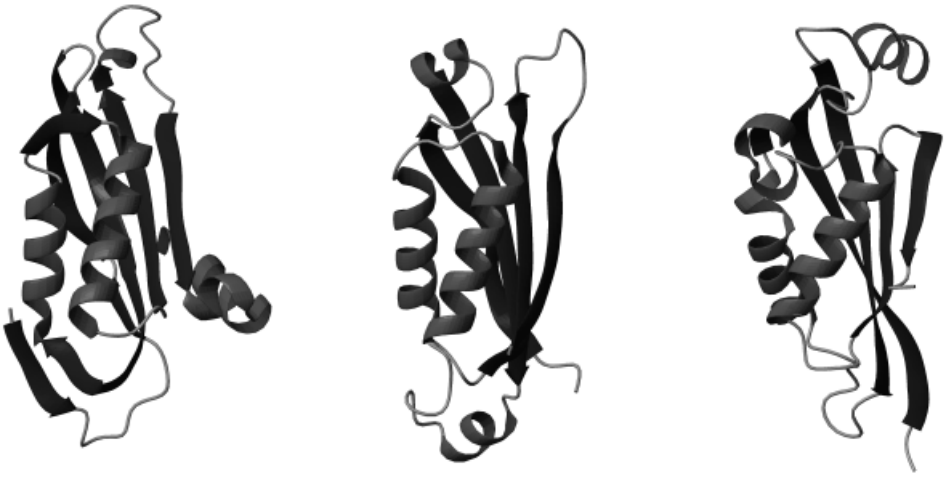


Fig. 4. **Center:** Protein 1dhn, which has a CATH α β three-layer (ββα) sandwich Aspartyl-glucosaminidase chain B (3.50.11) topology. **Left:** Protein 1a8rA, which has also a CATH α β three-layer (ββα) sandwich Aspartylglucosaminidase chain B (3.50.11) topology and has Z-score of 9.3 with protein 1dhn. **Right:** Protein 1b66A, which has a CATH α β two-layer sandwich Tetrahydropterin Synthase, subunit A (3.30.479) topology and has Z-score of 8.7 with protein 1dhn. This example illustrates how structural information alone might be insufficient to provide a clear-cut criterion for the classification of this protein.

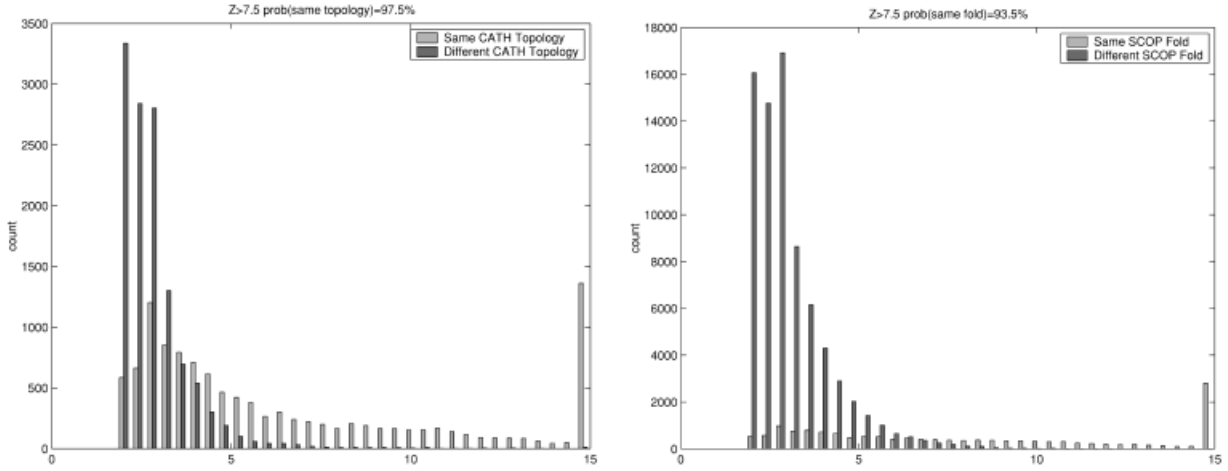



Fig. 5. Twilight zone for protein structure classification. **a:** The number of protein pairs of with a given FSSP Z-score that have different CATH folds is a rapidly decaying function of $Z$. On the contrary, the number of proteins pairs with the same CATH fold is decaying slowly. For $Z < 5$ there is a non-negligible probability to have different folds. We call this threshold the "twilight zone for structure classification." **b:** The corresponding histogram for SCOP folds. The number of SCOP folds is larger than the number of CATH topologies; hence the twilight zone is $Z \simeq 7$.

we found in CATH is the PDB entry 1rboC, which is classified as a α-β two-layer sandwich. It has 15 neighbors in PC, 14 of which are classified as mainly-β sandwiches.

We summarize the results about the assignments of the CATH architecture for proteins that already have a CATH classification (PFCs) in a "confusion table" (see Table II). The first column lists the "correct" classification (as given in CATH v1.7 for the test set); the second column gives the assignments by CO (correct, incorrect, or reject), and the third column lists the corresponding percentages. A full list of the inconsistent proteins is available on the web site.[31]

Another problem is that there are some large Z-scores between proteins of different architectures. Such large Z-scores arise when a protein of one particular architecture has a similar structure to a part of a protein of a different architecture. Swindells et al.[32] call the phenomenon of structures within structures, the "*Russian doll" effect.* Such cases are common between architectures of long proteins that contain substructures corresponding to architectures of shorter proteins; for example, there are many two-layer sandwich proteins that resemble a part of three-layer sandwich proteins. Such relationships can occur at the class level [e.g., α-β proteins that contain mainly-α or mainly-β proteins (1rboC, 1hgeA)]. They can also occur at the architecture level within the same class [e.g., α-β complex architecture contains α-β two-layer sandwich (1regX)]. Other inconsistencies occur when proteins fit two architecture definitions.

**TABLE II. Summary of a "Confusion Table"**

| Original classification | Assigned classification | Cases (%) |
|---|---|---|
| Mainly alpha | | |
| 1.10 Orthogonal bundle | 1.10 Orthogonal bundle | 96.3 |
| | reject | 3.3 |
| 1.20 Up-down bundle | 1.20 Up-down bundle | 97.7 |
| | 4.10 Irregular | 1.2 |
| | 1.10 Orthogonal bundle | 0.7 |
| 1.25 Horseshoe | 1.25 Horseshoe | 100.0 |
| 1.50 Alpha-alpha barrel | 1.50 Alpha-alpha barrel | 100.0 |
| Mainly beta | | |
| 2.10 Ribbon | 2.10 Ribbon | 93.9 |
| | reject | 5.7 |
| 2.20 Single sheet | 2.20 Single sheet | 97.2 |
| | reject | 2.3 |
| 2.30 Roll | 2.30 Roll | 97.2 |
| | reject | 2.1 |
| | 3.10 Roll | 0.7 |
| 2.40 Barrel | 2.40 Barrel | 91.0 |
| | reject | 8.8 |
| 2.50 Clam | 2.50 Clam | 94.4 |
| | 2.40 Barrel | 5.6 |
| 2.60 Sandwich | 2.60 Sandwich | 86.1 |
| | reject | 13.9 |
| 2.70 Distorted sandwich | 2.70 Distorted sandwich | 96.1 |
| | 2.60 Sandwich | 3.9 |
| 2.80 Trefoil | 2.80 Trefoil | 100.0 |
| 2.90 Orthogonal prism | 2.90 Orthogonal prism | 100.0 |
| 2.100 Aligned prism | 2.100 Aligned prism | 100.0 |
| 2.102 3-layer sandwich | 2.102 3-layer sandwich | 78.6 |
| | 2.30 Roll | 21.4 |
| 2.110 4 Propellor | 2.110 4 Propellor | 100.0 |
| 2.120 6 Propellor | 2.120 6 Propellor | 96.1 |
| | reject | 3.9 |
| 2.130 7 Propellor | 2.130 7 Propellor | 100.0 |
| 2.140 8 Propellor | 2.140 8 Propellor | 85.3 |
| | reject | 14.7 |
| 2.160 3 Solenoid | 2.160 3 Solenoid | 100.0 |
| 2.170 Complex | 2.170 Complex | 83.3 |
| | 2.60 Sandwich | 8.6 |
| | reject | 8.0 |
| Mixed alpha-beta | | |
| 3.10 Roll | 3.10 Roll | 99.9 |
| 3.20 Barrel | 3.20 Barrel | 100.0 |
| 3.30 2-layer sandwich | 3.30 2-layer sandwich | 93.5 |
| | reject | 6.0 |
| 3.40 3-layer(aba) sandwich | 3.40 3-layer(aba) sandwich | 96.1 |
| | reject | 3.8 |
| 3.50 3-layer(bba) sandwich | 3.50 3-layer(bba) sandwich | 72.1 |
| | reject | 27.2 |
| | 3.30 2-layer sandwich | 0.7 |
| 3.60 4-layer sandwich | 3.60 4-layer sandwich | 99.7 |
| 3.70 Box | 3.70 Box | 100.0 |
| 3.75 5-stranded propeller | 3.75 5-stranded propeller | 100.0 |
| 3.80 Horseshoe | 3.80 Horseshoe | 100.0 |
| 3.90 Complex | 3.90 Complex | 97.9 |
| | reject | 0.7 |
| Few secondary structures | | |
| 4.10 Irregular | 4.10 Irregular | 90.8 |
| | reject | 8.3 |
| | 1.20 Up-down bundle | 0.8 |

This table summarizes the results about the assignments of the CATH architecture for proteins that have already a CATH classification. Only cases that occur >0.5% are listed. These figures were calculated by using 100 cross-validation runs at 20% dilution.

**TABLE III. The Search Result When Submitting "1cuoA" to the Web Site http://www.weizmann.ac.il/physics/complex/compphys/f2cs/**

| | CATH v1.7 | | | | CATH v2.0 | | | | CATH prediction | | | SCOP 1.53 | | | SCOP prediction | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chain id | # | C | A | T | # | C | A | T | C | A | T | # | C | F | C | F |
| 1cuoa | −1 | | | | 1 | 2 | 60 | 40 | 2 | 60 | 40 | −1 | | | 2 | 5 |

This protein was classified by neither CATH v1.7 nor SCOP 1.53, which are the basis of our predictions. We predicted it to belong to CATH topology 2.60.40 and SCOP fold 2.5. Later it was indeed classified by CATH v2.0 as 2.60.40. The −1 in both CATH v1.7 and SCOP 1.53 represents that it was not classified by them.

### Class Prediction Using the Web Site

To retrieve our prediction for the CATH topology or SCOP fold of a protein, one can use the web site[31] by entering the protein chain identifier in the search box and submitting the query. If the protein appears in our database, then a table will be returned containing both the known and the predicted SCOP and CATH classifications. For example, the submission of the chain identifier "1cuoA" returns Table III. This protein was classified by neither CATH v1.7 nor SCOP 1.53, which are the basis of our predictions. We predicted it to belong to CATH topology 2.60.40 and SCOP fold 2.5. Later, the release CATH v2.0 identified 1cuoA as 2.60.40.

## CONCLUSIONS

The rapidly increasing number of experimentally derived protein structures requires a continuous updating of the existing structure classification databases. Each group adopts different classification criteria at the level of sequence, of structure, and of function similarities. A comparison between different classification schemes can help to understand the optimal interplay between different levels, it can reveal possible misclassification, and it can ultimately offer a fully automated updating procedure. Manual steps can be automated in an ever-increasing way by using the tools made available by other databases.

In this work we showed that it is possible to automatically predict the CATH topology and the SCOP fold from the FSSP Z-scores. It is possible to submit a protein of unknown CATH or SCOP classifications but known FSSP Z-scores to the web site[31] to obtain its CATH and SCOP classifications. Because the FSSP database is updated weekly, our procedure offers the possibility to update also CATH and SCOP with the same frequency (at least down to the topology and fold level, respectively). We introduced a classification method that clusters together structures of known and unknown classification according to their Z-scores. When proteins outside the twilight zone for structure comparison are considered, our method is highly reliable. We suggest that, to classify proteins within the twilight zone, other classification criteria, based on sequence and function similarity, must be adopted.

The advent of genome projects is multiplying the efforts in the field of protein classification. In the past, the aim was to find the structure of the particular protein that was interesting at a given time. Now the hope is to find a large representative set of structures that can encompass most of the existing folds, possibly all of them.[3] In such a large-scale project, human intervention, which is precious in setting the principles of classification, should be gradually replaced by automated procedures.

## ACKNOWLEDGMENTS

We thank Liisa Holm for making the raw FSSP data available to us and for useful discussions during the initial stages of this project. This work is based on a thesis for the M.Sc. degree submitted by G.G. to Tel-Aviv University (1998). We also thank Noam Shental for discussions. M.V. is supported by an European Molecular Biology Organization (EMBO) long-term fellowship; he also thanks the Einstein Center for Theoretical Physics for partial support of his stay at the Weizmann Institute. D.S. thanks the Weizmann Institute of Science for hospitality while part of this work was carried out.

## MATERIALS AND METHODS

### Databases and Protein Sets

Because the CATH and SCOP databases classify domains and FSSP deals with chains, we considered only chains that form a single domain; therefore, these proteins appear as a single entry in the three databases. Several groups have developed methods to identify protein domains.[20,23,33–35] In this work, we used the Dali Domain Dictionary[24] to identify single-domain proteins.

We used the following databases. The CATH release 1.7, which contains 15,802 protein chains, among which 10,906 are classified as single domain. This latter set is called PCs. We also used the CATH release 2.0, which contains 20,780 protein chains, among which 14,389 are single domain (PC2s). The SCOP release 1.53, which contains 20,021 protein chains, among which 15,375 are single domain (PSs). The FSSP release from 14 January 2001, which contains 22,660 protein chains (PF). The FSSP proteins are grouped into 2,494 homology classes so that within a class the sequence similarity is >25%. One protein per class is selected as representative, and we call PFr the set of all representatives. All the protein sets and their sizes are listed in Table IV.

### Classification by Optimization (CO) Method

The classification scheme that we used is based on the minimization of a particular cost function, defined as follows (for the case of the prediction of CATH topology; a similar definition holds for SCOP folds). Each protein is

**TABLE IV. Protein Sets and Their Sizes**

| Name | Description | Size of set |
|---|---|---|
| PF | All chains in FSSP (14 Jan, 2001) | 22,660 |
| PFr | Representative chains in FSSP (14 Jan, 2001) | 2,494 |
| PC | Chains in CATH v1.7 | 15,802 |
| PCs | Single-domain chains in CATH v1.7 | 10,906 |
| PC2 | Chains in CATH v2.0 | 20,780 |
| PC2s | Single-domain chains in CATH v2.0 | 14,389 |
| PS | Chains in SCOP 1.53 | 20,021 |
| PSs | Single-domain chains in SCOP 1.53 | 15,375 |
| PCsSs | Single-domain chains in SCOP 1.53 and CATH v1.7 | 10,197 |
| PFrCs | Single-domain chains in CATH that are representatives FSSP (PFr ∩ PCs) | 860 |
| PFrSs | Single-domain chains in SCOP that are representatives FSSP (PFr ∩ PSs) | 1,626 |
| PFCs | Chains in FSSP and single domain in CATH v1.7 | 10,541 |
| PFSs | Chains in FSSP and single domain in SCOP 1.53 | 14,716 |
| PF$\bar{C}$ | Chains in FSSP and not in CATH v1.7 | 7,431 |
| PF$\bar{S}$ | Chains in FSSP and not in SCOP 1.53 | 3,451 |

assigned an integer number $c_i$, describing its topology (1–305). We assign to proteins with known classification the value of $c(i)$ determined by their CATH classification. To the yet unclassified proteins we assign initially random values from 1 to 305. A cost is calculated for each configuration $C = \{c_i\}$ of topologies, which penalizes the assignment of different topologies to any pair of proteins. The value of this penalty is chosen to be the similarity measure $Z_{ij}$ between proteins $i$ and $j$; the higher the similarity $Z_{ij}$, the more costly it is to place proteins $i$ and $j$ in different topologies. The cost function is defined as the sum of penalties for all protein pairs $\langle i,j\rangle$,

$$E(C) = \sum_{\langle i,j\rangle} Z_{ij}[1 - \delta(c_i, c_j)]. \quad (1)$$

The classification problem is stated as finding the minimal cost configuration of the unclassified proteins, while keeping the topologies (i.e., the $c_i$ values) of the classified proteins fixed. This problem corresponds to finding the ground state of a random field Potts ferromagnet.

We search for a classification $C$ of minimal cost by an iterative greedy algorithm described in detail elsewhere.[21] The algorithm identifies at which iteration, if any, it performed a heuristic decision. For low fractions of unknown topologies, the algorithm usually reaches the global minimum of the cost function.

### Bounds on the Success Rate of the Prediction

In this section we establish a statistical upper bound for the prediction success rate relevant to a family of prediction algorithms.

The Z-matrix can be reinterpreted as a weighted graph; each vertex in the graph represents a protein and the weights on the edges connecting two vertices are the corresponding Z-scores. Edges with $Z < 2.0$ are absent from the graph. Following this representation, we define two proteins as neighbors if they are connected by an edge. By analyzing the connectivity properties of set PC we make inferences about our predictive power.

One can characterize the FSSP-based neighborhood of a protein according to the CATH classification of itself and its neighbors. Every protein must belong to one of four categories:

"Island": The protein has no neighbors.

"Colony": It has no neighbors of its own kind.

"Border": It has neighbors of its own kind as well as of other kinds.

"Interior": The protein has only neighbors of its own kind.

Using these definitions we can arrange the proteins of PC in groups according to their neighborhood category at the class, architecture, and topology levels. The distribution of the proteins among these groups can be used to calculate an upper bound for the CO method, if we assume that the set of unclassified proteins has the same distribution as the classified ones. For example, islands cannot be classified and are therefore rejected. Colonies are bound to be misclassified because none of their neighbors give a clue on their type. Because the fraction of proteins in each category was estimated on the basis of a sample, it can be interpreted only as a statistical upper bound.

We consider the set PFCs to obtain a first type of upper bound for the success rate of the CO method. This set (see Table IV) is formed by 10,541 proteins, among which 5% are islands, a negligible fraction (0.2%) are colonies, 6% are borders, and 88% are interiors. Therefore, the upper bound that we found is about 95% for predicting the topology level in CATH.

The actual prediction performed in this work is done on the set PF$\bar{C}$, which is formed by the 7431 proteins that are in FSSP (14 January 2001) but not in CATH1.7 (see Table IV). Within PF$\bar{C}$ there is a subset of 1617 (about 22%) proteins that are either islands or superislands, that is, they are connected only with other proteins in the subset and therefore they have no connection to proteins with known classification. Thus, the upper bound for this second type of prediction is about 78%.

### Evaluating a Classification Prediction Algorithm

Because an algorithm can output either a predicted classification or a "rejection," if it does not have any prediction, one has to estimate two probabilities: $P_{\text{success}}$ and $P_{\text{reject}}$. Robust estimation of these parameters is produced by cross-validation, a procedure that consists in averaging over many ($T$) randomly sampled test trials. In each trial, the set is divided into two subsets; one is used for training the algorithm and the other set, of $N_{\text{test}}$ proteins, is used to test the algorithm by comparing its prediction to the true classification. The probability estimates are given by

$$\hat{P}_{\text{success}} = 1/T \sum_{t=1}^{T} \frac{N_{\text{success}}}{N_{\text{test}}} \tag{2}$$

$$\hat{P}_{\text{non-reject}} = 1 - \hat{P}_{\text{reject}} = 1/T \sum_{t=1}^{T} \frac{N_{\text{test}} - N_{\text{reject}}}{N_{\text{test}}} \tag{3}$$

Another figure of merit, the purity $P_{\text{pure}}$, is the probability of correctly classifying nonrejected proteins. It is estimated by

$$\hat{P}_{\text{pure}} = \frac{\hat{P}_{\text{success}}}{1 - \hat{P}_{\text{reject}}} \tag{4}$$

## REFERENCES


1. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
2. Thornton JM, Orengo CA, Todd AE, Pearl FMG. Protein folds, functions and evolution. J Mol Biol 1999;293:333–342.
3. Šali A. 100,000 protein structures for the biologist. Nat Struct Biol 1998;5:1029–1032.
4. Martí-Renom MA, Ashley AC, Fiser A, Sanchez R, Melo F, Šali A. Comparative protein structure modeling og genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
5. Heger A, Holm L. Towards a covering set of protein family profiles. Prog Biophys Mol Biol 2000;73:321–337.
6. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
7. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
8. Fisher D, Rice D, Bowie JU, Eisenberg D. Assigning amino acid sequences to 3-dimensional protein folds. FASEB J 1996;10:126–136.
9. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proc Natl Acad USA 1997;94:11911–11916.
10. Murzin AG. Structural classification of proteins: new superfamilies. Curr Opin Struct Biol 1996;6:386–394.
11. Blundell TL, Mizuguchi K. Structural genomics: an overview. Prog Biophys Mol Biol 2000;73:289–295.
12. Finkelstein AV. Protein structure: What is possible to predict now? Curr Opin Struct Biol 1997;7:60–71.
13. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37(Suppl 3):171–176.
14. Bernstein F, Koetzle T, Williams G, Meyer EJ, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
15. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
16. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
17. Conte LL, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
18. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database for protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
19. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
20. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci 1995;4:872–884.
21. Getz G. Clustering and classification of protein structures. M.Sc. Thesis, Tel-Aviv University, 1998.
22. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112.
23. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609.
24. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–57.
25. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall; 1988.
26. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
27. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
28. Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. Proteins 1992;14:139–167.
29. Bray JE, Todd AE, Pearl FMG, Thornton JM, Orengo CA. The CATH Dictionary of Homologous Superfamilies: a consensus approach to analyze distant structural homologues. Protein Eng 2000;13:153–165.
30. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000; 28:254–256.
31. http://www.weizmann.ac.il/physics/complex/compphys/f2cs/index.html.
32. Swindells MB, Orengo CA, Jones DT, Hutchinson EG, Thornton JM. Contemporary approaches to protein structure classification. Bioessays 1998;20:884–891.
33. Islam SA, Luo J, Sternberg MJE. Identification and analysis of domains in proteins. Protein Eng 1995;8:513–525.
34. Swindells MB. A procedure for detecting structural domains in proteins. Protein Sci 1995;4:103–112.
35. Sowdahamini R, Rufino SD, Blundell TL. Nuclear dynamics and electronic transition in a photosynthetic reaction center. J Am Chem Soc 1997;119:3948–3958.

# Publication 12:

## FSSP to SCOP and CATH (F2CS) Prediction Server

Authors: G. Getz, A. Starovolsky and E. Domany
Published in: *submitted*.

# FSSP to SCOP and CATH (F2CS) Prediction Server

Gad Getz[1], Alina Starovolsky[2] and Eytan Domany[1]
[1]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel
[2]Computer Science Department, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

October 23, 2003

**ABSTRACT**
**Summary:** The F2CS server provides access to the software, F2CS2.00, that implements an automated prediction method of SCOP and CATH classifications of proteins, based on their FSSP Z-scores (Getz *et al.*, 2002),
**Availability:** Free, at
http://www.weizmann.ac.il/physics/complex/compphys/f2cs/.
**Contact:** eytan.domany@weizmann.ac.il
**Supplementary information:** The site contains links to additional figures and tables.

Since during evolution protein structures are much more conserved than sequences and even functions (Holm & Sander, 1996), proteins are usually classified first by their structural similarity. Newly solved structures of proteins are regularly stored in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977). Many research groups study the diversity of protein structures and maintain web-accessible hierarchical classifications of them. Three widely used databases are FSSP (Holm & Sander, 1997), CATH (Orengo *et al.*, 1997) and SCOP (Conte *et al.*, 2000); although each has its own way to compare and classify proteins, the resulting classification schemes are, largely, consistent with each other (Getz *et al.*, 2002; Getz, 1998; Hadley & Jones, 1999).

The major difference between these three classification schemes, relevant to this work, is their degree of automation. FSSP is based on a fully automated structure comparison algorithm, DALI (Holm & Sander, 1994; Dietmann *et al.*, 2001), that calculates a structural similarity measure (represented in terms of Z-scores) between pairs of structures of protein chains taken from the PDB. FSSP first selects a subset of representative structures from the PDB and then applies the DALI algorithm to calculate the Z scores for all pairs of representatives. Next, they calculate the Z scores between each representative and the PDB structures it represents. Being fully automated, FSSP can be updated fairly often. FSSP was recently extended by a new database, called Dali (Holm, 2003), which contains all-against-all Z-scores between chains and domains of a larger representative set, PDB90 (Hubbard *et al.*, 1999), in which no two chains are more than 90% sequence identical. In contrast, CATH and SCOP use manual classification at certain levels of their hierarchy, which slows down the classification process and makes it more subjective and error-prone.

CATH arranges protein domains in a four-level hierarchy according to their **C**lass (secondary structure composition), **A**rchitecture (shape formed by the secondary structures), **T**opology (connectivity order of the secondary structures) and **H**omologous superfamily (structural and functional similarity). Classification of Architecture is done by visual inspection; hence CATH is partially manual.

The top level (**C**lass) of the SCOP database also describes the secondary structure content of a protein domain. The next level (**F**old) groups together structurally similar domains. The lower two levels (superfamily and family) describe near and distant evolutionary relationships (Levitt & Chothia, 1976). "Fold" largely corresponds to CATH's topology level (Getz *et al.*, 2002). SCOP is constructed manually, based on visual examination and comparison of structures, sequences and functions.

We present here a web-based server, available at http://www.weizmann.ac.il/physics/complex/compphys/f2cs/, whose aim is to predict, without human intervention, using a protein's FSSP (or DALI) Z-scores, it's full SCOP and CATH classifications. This can help classify proteins of known structure that were not yet processed by SCOP or CATH, and call attention to yet unseen structural classes.

If a protein appears in FSSP, the server returns our prediction. If it is not in FSSP, the user can submit the new structure to the DALI server, insert the resulting Z-scores into our server and obtain its predicted classification. In both cases F2CS outputs a table showing the prediction, along with its confidence level.

## THE SERVER

The current predictions are based on the latest versions

| Chain id | CATH v2.5 | | | | **CATH Prediction** | | | SCOP 1.63 | | | **SCOP Prediction** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | C | A | T | **C** | **A** | **T** | # | C | F | **C** | **F** |
| 1dowb | -1 | | | | **1** | **20** | **5** | -1 | | | 8 | 1 |
| Success% | | | | | 100 | 99 | 100 | | | | 97 | 100 |

**Table 1:** Results obtained by submitting "1dowb" to the F2CS server. This protein was classified by neither CATH v2.5 nor SCOP 1.63 (indicated by -1 in the "number of domains" columns). We predict the following classifications: 1.20.5 for CATH and 8.1 for SCOP, both at 100% confidence level.

of the databases; FSSP (Jun 16, 2002 update), combined with the Dali database (preliminary version, May 2003); CATH version 2.5 (Jul 2003) and SCOP 1.63 release (May 2003). The FSSP database contains 27182 chains, 2860 out of which are representatives. We superimposed on these the Z-scores from the Dali database, which were calculated for 6433 PDB90 chains; we refer to the combined database as FSSP/DD. Only significant Z-scores are reported ($\geq 2$) and used; all other Z-scores are assumed to be zero.

The server implements our method (Getz *et al.*, 2002), *Classification by Optimization* (CO), an optimization procedure that searches for that class assignment of proteins (that were not yet processed by CATH or SCOP), which attains a minimal cost. The cost of an assignment is the sum of Z-scores between all pairs of proteins that were not assigned to the same class. This is a "partially supervised" algorithm, since it utilizes for its prediction the labels of the proteins with known classification and also the Z scores among the training and predicted sets. We can not classify "isolated" proteins, which are not connected by a path of neighboring chains (*i.e.* $Z \geq 2$) to a chain of known classification.

We generate a prediction database of chains which appear in FSSP/DD but not in SCOP or CATH by applying our algorithm for each classification scheme. The FSSP/DD version we are using contains 4014 chains which do not appear in CATH v2.5 (for 3170 of which we supply a prediction) and 511 which are not in SCOP 1.63 (403 of which have a prediction). Since CATH and SCOP handle protein domains whereas FSSP/DD entries are protein chains (consisting of one or more domains[1]), we use as a training set the single domain chains that are of known classification. Note that SCOP and CATH do not always agree on their separation of proteins into domains.

Our prediction's success rate was estimated using a blind test in which we hid the assignments of 3605 proteins from CATH and 4570 proteins from SCOP and tested our predictions against the known classifications. The success rate was tested for each class separately.

With every new release of the databases, F2CS can be updated; the newly released CATH/SCOP classifications are added to the training set, while predictions are made for proteins contained in a new FSSP/DD release which are not yet classified by CATH or SCOP.

## USAGE

In order to retrieve our prediction for CATH's class, architecture and topology or SCOP's class and fold of a protein, enter the protein chain's identifier in the search box and submit the query. If the protein appears in our database, a table will be returned containing both the known and the predicted SCOP and CATH classifications. For example, submission of the chain identifier "1dowb", which was classified neither by CATH v2.5 nor SCOP v1.63, returns Table 1. We predict CATH classification 1.20.5 and SCOP 8.1, both near 100% confidence level. The "Success%" link points to a table with the exact numbers by which the success rates were estimated.

In case the queried protein is not in our database, the user can obtain its predicted classification by following these two steps: (a) submit the protein's PDB file to the DALI server (the engine behind FSSP) which calculates its structural similarity to the FSSP representatives and returns a list of the representatives and Z-scores for which $Z \geq 2$. (b) Paste DALI's reply in the appropriate query box in our server.

[1]We do not classify the few cases, when a single domain contains several different chains or a combination of their parts.


## ACKNOWLEDGEMENTS

We thank L. Holm for directing us to her new Dali database, and M. Vendruscolo for his advice and active involvement in the initial stages of this project, which was partially supported by the German-Israel Science Foundation (GIF). G.G. is supported by the Sir Charles Clore Doctoral Scholarship.


## REFERENCES


Bernstein, F., Koetzle, T., Williams, G., Meyer, E. J., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.,* **112**, 535–542.

Conte, L. L., Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G. & Chothia, C. (2000) SCOP: a structural classification of proteins database . *Nucleic Acids Res.,* **28**, 257–259.

Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. & Holm, L. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3 . *Nucleic Acids Res.,* **29**, 55–57.

Getz, G. (1998) *Clustering and Classification of Protein Structures.* M. Sc. Thesis, Tel-Aviv University.

Getz, G., Vendruscolo, M., Sachs, D. & Domany, E. (2002) Automated assignment of scop and cath protein structure classifications from fssp scores. *Proteins,* **46**, 405–415.

Hadley, C. & Jones, D. T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure,* **7**, 1099–1112.

Holm, L. & Sander, C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.,* **22**, 3600–3609.

Holm, L. & Sander, C. (1996) Mapping the protein universe. *Science,* **273**, 595–602.

Holm, L. & Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds . *Nucleic Acids Res.,* **25**, 231–234.

Holm, L. (2003) Dali database at `http://www.bioinfo.biocenter.helsinki.fi:8080/dali`, private communication.

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., Chothia, C. (1999) SCOP: a structural classification of protein database . *Nucleic Acids Res.,* **27**, 254–256.

Levitt, M. & Chothia, C. (1976) Structural patterns in globular proteins. *Nature,* **261**, 552–558.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997) CATH - a hierarchic classification of protein domain structures. *Structure,* **5**, 1093–1108.

# Chapter 5

# Summary and Conclusions

The newly born field of functional genomics generates vast amounts of data collected using different experimental techniques. Most of these techniques are developed towards high-throughput experiments which can measure thousands to hundreds of thousands of numbers simultaneously. Moreover, automation of these experiments make them easier to perform in many labs all over the globe. On top of that, the slow but steady agreement on standards for storing and dissemination of the data, and the ease of access to databases over the internet, make this data available for all researches. This data are, however, very noisy, which poses a challenge for data mining and statistical analysis tools to succeed in extracting valuable biological information from it. Functional genomics and the development of computational biology methods have already had considerable impact on both biology and computer science. In biology, and especially in cancer research, new types and sub-types of diseases are identified, genes are assigned new functions, cellular networks are being enhanced and new key-player genes are discovered. On the computational side, new data mining and machine learning tools are being developed and adapted to handle the specific tasks. For example, string matching algorithms were developed for genomic searches; biclustering algorithms that identify informative submatrices in a large matrix were invented for gene expression analysis and three dimensional local structure matching are used to compare protein structures.

The work described in this thesis can be divided into two parts: (i) development and implementation of data mining and analysis methods and (ii) application of these methods to various biological problems. The developed analysis methods include unsupervised, semi-supervised and supervised techniques. Most of the applications dealt with gene expression data. Other types of data which were analyzed include antigen reactivity data and protein structures.

For unsupervised analysis we introduced a biclustering method which we named coupled two-way clustering (CTWC), and applied mainly to gene expression data. The aim of CTWC is to identify small sets of genes which can provide a significant statement regarding a certain subset of the samples. The main two problems with this task are: (i) the exponential number of possible sub-matrices which prevents one from performing exhaustive

searches and (ii) controlling the rate of false discoveries. The first problem drove us to using an efficient heuristic search method, whereas the second requires careful assignment of the proper $p$-values for the generated statements and correct handling of multiple hypotheses. We use a density-based clustering method, super paramagnetic clustering (SPC) [74], to identify clusters of genes and samples. A stability measure is used to select statistically significant clusters (which can be translated to $p$-values using permutation tests). We do not handle multiple hypotheses as part of CTWC, but once a clear separation is found, one can apply supervised methods to search for additional genes (or samples) that perform such a partition. For such a supervised analysis, we apply the FDR method [44] to control the rate of false positives.

We also presented a method for semi-supervised analysis, which uses the same $q$-state Potts model as SPC, in which prior knowledge is incorporated as fixed spins. This model enlarges the phases in which correct classification is obtained and by that improves over unsupervised analysis. The work on this model is still at its early stages and many questions regarding it are still open. In this work we introduced the concept of typical cuts to the semi-supervised community, which until now focused on searching for minimal cuts. Semi-supervised analysis is a very young field and has not been utilized much for the analysis biological data. Since knowledge regarding biological systems is continuously increasing and one would like to integrate this knowledge in the analysis of new data, I anticipate that semi-supervised methods will play an important role in computational biology.

Regarding supervised analysis, we suggested a method which can identify genes which are related to survival, based on comparing survival curves of two patient groups. The analysis of survival data in conjunction with gene expression data is extremely important in cancer and drug research. Such analyses can generate better taxonomy of the diseases and allow for individual therapies and better diagnosis. The use of such analyses in drug development can partition patients according to their response to certain therapies. This is an example of the potential in combining information from various sources, which is discussed below.

There are still many challenges and open questions regarding the analysis of gene expression data. The following list provides possible direction for further research (organized according to the experiment cycle), related to the methods presented in the thesis:

**Sample selection and preparation**

- The choice of samples to be examined can directly influence the possible statistical statements generated by the analysis. For example, in order to generate diagnostic tools and study survival of a certain cancer, one has to select patients who were treated in a similar manner; otherwise, differences in survival that are caused by the different therapies can be mistakenly attributed to different gene expression profiles. The challenge is to obtain a large number of tumor samples with proper records regarding their handling and patient treatment.

- The problem of mixed cell-types is also crucial in the analysis of tumor samples. Each tumor sample consists of cells of different types and, therefore, a gene expression measurement of mRNA extracted from the entire sample yields a weighted average of the mRNA concentrations of the different cell-types. To tackle this problem, several labs have recently started using microdissection technology to separate the different cell-types under a microscope. This improves the purity of the sample but still various cell-types contaminate each sample at different rates. This problem raises both technological and computational challenges; the development of more precise microdissection methods that may utilize different cell markers to identify the desired cells and the improvement of microarray technologies to enable accurate measurements of mRNA extracted from a small number of cell. On the computational side, algorithms for *in silico* purification of the measurements, by subtracting the effect of the contaminating cells, are needed.

**Experimental design and preprocessing**

- Better technology and statistical models for the measured data and its noise are necessary. Variations in experimental results are still large and contain uncontrolled systematic errors. Experiments performed in the same lab by different people, or at different times, display consistent differences [140] - these differences need to be controlled by more precise and standard experimental protocols or removed by proper scaling methods which are based on elaborate statistical models. Some advances in this field are being made [13,15–17] but much has yet to be done. This is an important step prior to any analysis.

- Continuing the previous item, there is a lack in methods to quantitatively compare results obtained by different microarray technologies. Such a method would enable us to join data originating from different labs and by that enlarge the data sets available for analysis. Large data sets are needed to extract weak but statistically significant signals from the data, and since microarray experiments are expensive, the number of experiments performed in a single lab is usually below the statistical requirements [6, 38].

**Cluster analysis and biclustering**

- Cluster analysis is widely used in gene expression analysis. A still open question is how to assign p-values for the identified clusters. Commonly, a null hypothesis of a uniform distribution is used and the p-value is calculated by measuring the probably to obtain certain features of the cluster from randomly permutated data. This procedure, however, does not test for alternative hypotheses. For example, one can identify a dense group of genes as a cluster for which the p-value is significantly low. This, however, does not mean that this cluster is not part of a larger group of

genes, which has a much lower p-value and should have been chosen as the cluster. Correctly assigning p-values is extremely important for controlling the false discovery rate and for combining with other tests in further analysis.

- An additional question regarding clustering which is important for experimental design deals with the number of samples that are needed in order to extract significant clusters. As in simpler statistical tests, one would like to know the number of samples, $N_s$, needed in order to identify, at a given $p$-value, clusters which have a density difference of at least $\Delta\rho$ from their environment, *i.e.* know the function $N_s(\Delta\rho, p)$.

- Regarding coupled two-way clustering (CTWC) [108], a score for a generated statement would improve the search for significant biclusters in the data. In contrast to the currently used scores (see Sec. 2.2.6), biclusters that form dense low-dimensional manifolds, which can be identified by CTWC, should score highly.

- A different possible extension of CTWC is to use combinations of several clusters in the search for significant partitions. Currently, CTWC uses a single gene cluster at a time. This restricts the search which, on the one hand, helps avoiding the exponential number of biclusters, but, on the other hand, may fail to detect more elaborate signals in the data. For example, in the analysis of survival and gene expression in breast cancer (see [38] and references therein), it is demonstrated that expression levels of a single cluster of genes, representing a single cellular process, does not suffice to reach high quality partitions of the samples according their survival. Using several clusters at a time is, of course, more time consuming and needs more delicate statistical analysis.

- Both unsupervised and supervised methods may produce less false positives if they incorporate a statistical model for the noise. In the current practice, the log-transform is used to transform the experimental noise to be approximately intensity independent and Gaussian distributed. This approximation is valid only for intermediate expression levels; low expression levels have additional noise components which are, in general, larger, whereas the noise at high expression levels is relatively smaller. Incorporation of a more sophisticated model for the noise as part of the analysis may produce better distance measures between genes and samples.

### Combination of various data sources

A major challenge which holds a great promise, in my view, is the analysis of data generated from different sources. The assumption that the gene expression levels represent the "state" of the cell is far from being correct. Many processes are performed at the protein level and include mechanisms as protein modifications, localization and degradation. Additional cellular regulation is performed by methylation of the DNA which prevents RNA

production. DNA modifications, called in general *epigenetics*, may prove to be as important as transcription factors in regulation of cellular processes. Experimental techniques to detect these various cellular details are being developed, and some are advancing towards becoming suitable for large scale experiments. Data from such experiments complement gene expression data and provide a much better view of the examined cells.

Other data, which are already being analyzed in conjunction with gene expression [130, 133, 169], are extracted from genomic sequences. Promoter regions are analyzed to identify patterns of transcription factor binding sites [96]. These can be used to unravel regulatory networks and infer biological function of genes. Analysis of homolog and ortholog genes by comparing genomes of different species can help identify conserved and unconserved DNA regions which, in many cases, imply functional and non-functional DNA segments, respectively. Additional information is the genomic location of the genes, *i.e.* on which chromosome and at which segment they are located. This information can also help infer gene-function or associate a gene cluster with a specific location which may be damaged or altered, which is important in the analysis of tumor samples.

Genetic analysis of different individuals provides additional data regarding genes. Millions of single nucleotide polymorphisms (SNPs) are already known and can help partition patients into groups according to their genetic markers. These groups can be compared to clusters obtained from gene expression analysis and by that increase the confidence of the identified group and link genetic markers with specific molecular profiles. Additional data regarding genes, which is central in cancer research, originates from mutation analysis of genes (*e.g.* P53). Analysis of the effects of mutations in specific genes on gene expression can help identify key-player genes in cancer and other diseases.

Other types of data regarding the samples are characteristics of the patients from which these samples were extracted (as in tumor analysis). These include survival data, choice of therapy, age group, gender, race. One can also characterize a patient according to his/her immune-system response, *e.g.* antigen reactivity [102] (see Publication (6)), to various antigens.

All these data sets can be arranged in matrices with common axes. For example, information regarding the genes can be organized in rows as the gene expression data and, likewise, information regarding the samples/patients can be organized in columns. These multiple matrices (which can be of more than two dimensions) should be analyzed simultaneously. A weak signal in one data set can be validated or invalidated by signals in other data sets. A combined $p$-value that takes into account the data from all such data sets can compensate for high $p$-values in each data set alone.

### Gene networks

The main challenge of functional genomics, which is to decipher the gene functions and regulatory networks, is still open. Advances in reverse engineering of the cell mechanisms are constantly being made [133, 169, 170] but still are far from being completed. In order to analyze the phenotypic effect of genes and infer causal relationships between their ex-

pression levels, a new technology was developed which replaces the classical and tedious method of knockouts. This technology is based on RNAi (RNA interference), a natural mechanism for sequence-specific, post transcriptional gene silencing triggered by double-stranded RNA. Synthetically generated short interfering RNAs, which are part of the gene targeted for silencing, are inserted into the cell and by various mechanisms [171] cause the silencing of the specific gene. This enables performance of high-throughput experiments in which many genes or combinations of genes are silenced. Measuring gene expression for each of these "knockout" perturbations can help reverse engineer their function and relationship with the other genes.

I think that shortening the experiment-analysis cycle, by creating integrated teams of experimentalists and experts in data analysis can help tackle these challenges and increase the rate of biological discoveries.

# Bibliography

[1] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] J. C. Venter et al. The Sequence of the Human Genome. *Science*, 291:1304–1351, 2001.

[3] E.S. Lander. Array of hope. *Nature Genetics*, 21:3–4, 1999.

[4] P. Hieter and M. Boguski. Functional genomics: it's all how you read it. *Science*, 278:601–602, 1997.

[5] F. H. C. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[6] P. Sebastiani, E. Gussoni, I. S. Kohane, and M. F. Ramoni. Statistical Challenges in Functional Genomics. *Statistical Science*, 18:33–70, 2003.

[7] T. R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[8] A. A. Alizadeh, M. B. Eisen MB, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet H, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[9] C. H. Chung, P. S. Bernard, and C. M. Perou. Molecular portraits and the family tree of cancer. *Nat. Genet.*, 32:533–540, 2002.

[10] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *International Conference in Computer Vision (ICCV)*, pages 377–384, 1999.

[11] X. Zhu, Z. Ghahramani, and J. Laferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, 2003.

[12] M. K. Kerr and M. Martin adn G. A. Churchill. Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, 7:819–837, 2000.

[13] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 31:e15, 2003.

[14] Microarray suite 4, affymetrix inc.

[15] Microarray suite 5, affymetrix inc.

[16] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. USA*, 98:31–36, 2001.

[17] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, 2:research0032:1–11, 2001.

[18] Dna-chip analyzer.

[19] Quantarray microarray analysis software manual. perkin-elmer inc.

[20] M. B. Eisen. ScanAlyze: Software and Documentation. Free Methods in Segmentation of cDNA Microarray Images.

[21] G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.*, 32:490–495, 2002.

[22] B. P. Durbin, J.S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:S105–S110, 2002.

[23] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.

[24] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York, USA, 1973.

[25] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice–Hall, Englewood Cliffs, 1988.

[26] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

[27] V. N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons Inc., New York, 1998.

[28] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[29] A. Albrecht, S. A. Vinterbo, and L. Ohno-Machado. An Epicurean learning approach to gene-expression data classification. *Artif. Intell. Med.*, 28:75–87, 2003.

[30] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17:S215–S224, 2001.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[32] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. A. Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. USA*, 97:262–267, 2000.

[33] G. Valentini, M. Muselli, and F. Ruffino. Bagged ensembles of svms for gene expression data analysis. In *Proceeding of the International Joint Conference on Neural Networks - IJCNN 2003*, 2003.

[34] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., New York, USA, 2001.

[35] L. Breiman. Bagging predictors. *Annals of Statistics*, 24:123–140, 1996.

[36] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19:1090–1099, 2003.

[37] Y. Freund and R. E. Shapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55:119–139, 1997.

[38] I. Kela, G. Getz, L. Ein-Dor, and E. Domany. Is there a Unique Gene-Expression Signature of Survival in Breast Cancer? In preparation, 2004.

[39] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles . *Journal of Computational Biology*, 7:559–584, 2000.

[40] A. Ben-Dor, N. Friedman, and Z. Yakhini. Overabundance Analysis and Class Discovery in Gene Expression Data. *submitted*, 2003.

[41] R. R. Sokal and J. F. Rohlf. *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman and Co., New York, 3rd edition, 1995.

[42] D. R. Cox and D. Oakes, editors. *Analysis of Survival Data*. Chapmen and Hall, London, UK, 1984.

[43] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. USA*, 100:8418–8423, 2003.

[44] Y. Benjamini and Y. D. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B.*, 57:289–300, 1995.

[45] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29:1165–1188, 2001.

[46] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19:368–375, 2003.

[47] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. USA*, 98:5116–5121, 2001.

[48] J. D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. Technical report, Department of Statistics, Stanford University, 2001.

[49] J. D. Storey and R. Tibshirani. Estimating false discovery rate under dependence, with applications to dna microarrays. Technical report, Department of Statistics, Stanford University, 2001.

[50] J. D. Storey. A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B*, 64:479–498, 2002.

[51] J. D. Storey and R. Tibshirani. Sam thresholding and false discovery rates for detecting differential gene expression in dna microarrays. In *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York, 2003.

[52] L. J. van 't veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[53] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, USA, 1993.

[54] R. Simon, M. D. Radmachera, K. Dubbin, and L. M. McShane. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *J. Natl. Cancer Inst.*, 95:14–18, 2003.

[55] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. USA*, 97:10101–10106, 2000.

[56] M. Bittnerand, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix M. Rad-macher, R. Simon, Z. Yakhinik, A. Ben-Dork, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[57] J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer. Gene ex- pression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, 58:5009–5013, 1998.

[58] E. W. Weisstein. *CRC Concise Encyclopedia of Mathematics, Second Edition.* CRC Press, 2002.

[59] C. E. Shanon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423, 623–656, 1948.

[60] T. M. Cover and J. A. Thomas, editors. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[61] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–726, 2000.

[62] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. USA*, 95:14863–14868, 1998.

[63] P. T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[64] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. USA*, 98:10869–10874, 2001.

[65] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[66] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, New York, 3rd edition, 2001.

[67] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute and Computer Science Division, University of Berkely, 1998.

[68] T. P. Minka. Expectation-maximization as lower bound maximization. *Online article*, 1998.

[69] K. Rose and G.C. Fox E. Gurewitz. A deterministic annealing approach to clustering. *Phys. Rev. Lett.*, 11:589–594, 1990.

[70] J. Schneider. First-order phase transitions in clustering. *Phys. Rev. E*, 57:2449–2451, 1998.

[71] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. USA*, 96:6745–6750, 1999.

[72]

[73] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of Computer Vision and Pattern Recognition*, 1997.

[74] M. Blatt, S. Wiseman, and E. Domany. Superparamagnetic clustering of data. *Phys. Rev. Lett.*, 76:3251–3254, 1996.

[75] O. Barad. Advanced clustering algorithm for gene expression analysis using statistical physics methods. Master's thesis, Weizmann Institute of Science, Rehovot, Israel, 76100, 2003.

[76] N. Shental, A. Zomet, T. Hertz, and Y. Weiss. Learning and inferring image segmentations with the gbp typical cut algorithm. In *Proceedings of International Conference on Computer Vision*, 2003.

[77] M. Blatt, S. Wiseman, and E. Domany. Data clustering using a model granular magnet. *Neural. Comput.*, 9:1805–1842, 1997.

[78] F. W. Wu. The Potts model. *Rev. Mod. Phys.*, 54:235, 1982.

[79] E. Levine and E. Domany. Resampling Method for Unsupervised Estimation of Cluster Validit. *Neural Comp.*, 13:2573–2593, 2001.

[80] C. M. Perou, S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen, D. T. Ross, A. Pergamenschikov, C. F. Williams, S. X. Zhu, J. C. Lee, D. Lashkari, D. Shalon, P. O. Brown, and D. Botstein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. USA*, 96:9212–9217, 1999.

[81] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, and P. O. Brown amd D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[82] T. H. Cormen, C. L. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 1990.

[83] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[84] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[85] D. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, 2000.

[86] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *europhys.*, 19:451–458, 1992.

[87] B. A. Berg and T. Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267:249–253, 1991.

[88] B. A. Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, 68:9–12, 1992.

[89] B. A. Berg and T. Celik. A New Approach to Spin Glass Simulations. *Phys. Rev. Lett.*, 69:2292–2295, 1992.

[90] Y. Iba. Extended ensemble monte carlo. *International Journal of Modern Physics C*, 12:623–656, 2001.

[91] M. Safran, I. Solomon, O. Shmueli, M. Lapidot, S. Shen-Orr, A. Adato, U. Ben-Dor, N. Esterman, N. Rosen, I. Peter, T. Olender, V. Chalifa-Caspi, and D. Lancet. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, 18:1542–1543, 2002.

[92] V. Chalifa-Caspi, I. Yanai, R. Ophir, M. Shmoish, H. Benjamin-Rodrig, N. Rosen, P. Kats, M. Safran, O. Shmueli, and D. Lancet. GeneAnnot: Annotation of high-density oligunocleotide arrays and their linking with GeneCards. In *ISBM*, 2003.

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

[94] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, 31:19–20, 2002.

[95] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids. Res.*, 23:4878–4884, 1995.

[96] L. Hertzberg, O. Zuk, G. Getz, and E. Domany. Finding Motifs in Promoter Regions. submitted, 2004.

[97] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, 296:1205–1214, 2000.

[98] D .G. Ginzinger. Gene quantification using real-time quantitative pcr: An emerging technology hits the mainstream. *Experimental Hematology*, 30:503–512, 2002.

[99] C. A. Iacobuzio-Donahue and R. H. Hruban. Expression profiling of pancreatic ductal adenocarcinoma. In M. Landanyi and W. L. Gerald, editors, *Expression Profiling of Human Tumors*, pages 257–275. Humana Press, New Jersey, USA, 2003.

[100] D. Volk and M. G. Stepanov. Resampling methods for document clustering. cond-mat/0109006, 2001.

[101] D. Mirzayof. Clustering Analysis of Biochemichal Chips. Master's thesis, Weizmann Institute of Science, 2001.

[102] F. Quintana, G. Getz, G. Hed, E. Domany, and I.R. Cohen. Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity. *European journal of immunology*, 2003. Submitted.

[103] G. Hed, A. K. Hartmann, D. Stauffer, and E. Domany. Spin domains generate hierarchical ground state structure in J = +/- 1 spin glasses. *Phys. Rev. Lett.*, 86:3148, 2001.

[104] E. Domany. Cluster Analysis of Gene Expression Data. *J. Stat. Phys.*, 110:1117–1139, 2003.

[105] G. Getz and E. Domany. Coupled Two-Way Clustering Server. *Bioinformatics*, 19:1153–1154, 2003.

[106] I. Tsafrir, D. Tsafrir, and E. Domany. Sorting points into neighborhoods (SPIN). to be published.

[107] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[108] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. USA*, 97:12079–12084, 2000.

[109] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsemeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30:41–47, 2002.

[110] D. Greenbaum, R. Jansen, and M. Gerstein. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics*, 18:585–596, 2002.

[111] G. Getz. Clustering and Classification of Protein Structures. Master's thesis, Tel-Aviv University, 1998.

[112] H. Agrawal. Extreme Self-Organization in Networks Constructed from Gene Expression Data. *Phys. Rev. Lett.*, 89:268702, 2002.

[113] T. Rozovskaia, O. Ravid-Amir, S. Tillib, G. Getz, E. Feinstein, H. Agrawal, A. Nagler, E. Rappeport, I. Issaeva, Y. Matsuo, U. R. Kees, T. Lapidot, F. Lo Coco, R. Foa, A. Mazo, T. Nakamura, C. M. Croce, G. Cimino, E. Domany, and E. Canaani. Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements. *Submitted*, 2003.

[114] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic, 1996.

[115] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513–523, 1988.

[116] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41:391–407, 1990.

[117] S. T. Dumais, T. K. Landauer, and M. L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Proceedings of the ACM SIGIR '96 Workshop on Cross-Linguistic Information Retrieval*, 1996.

[118] T. G. Kolda and D. P. O'Leary. A semidiscrete matrix decomposition for latent semantic indexing information retrieval. *ACM Trans. Inf. Syst.*, 16(4):322–346, 1998.

[119] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data. *J. Amer. Stat. Assoc.*, 58:415–434, 1963.

[120] J. A. Hartigan. Direct clustering of a data matrix. *J. Amer. Stat. Assoc.*, 67:123–129, 1972.

[121] D. J. Watts. *Small worlds. The dynamics of networks between order and randomness.* Princeton University Press, Princeton, NJ, 1999.

[122] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12:61–86, 2002.

[123] Y. Cheng and G. M. Church. Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:93–103, 2000.

[124] A. Califano, G. Stolovitsky, and Y. Tu. Analysis of Gene Expression Microarrays for Phenotype Classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:75–85, 2000.

[125] T. Hastie, R. Tibshirani, M. Eisen, P. Brown, D. Ross, U. Scherf, J. Weinstein, A. Alizadeh, L. Staudt, and D. Botstein. 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, 1:research0003.10003.21, 2000.

[126] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. In *Proceedings of the sixth annual international conference on Computational biology*, pages 49–57. ACM Press, 2002.

[127] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pages 89–98, 2003.

[128] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[129] S. Busygin, G. Jacobsen, and E. Krämer. Double Conjugated Clustering Applied to Leukemia Microarray Data. In *Second SIAM ICDM, Workshop on clustering high dimensional data*, 2002.

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

[131] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E*, 67:9–12, 2003.

[132] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1989.

[133] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. 2004. To be published.

[134] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.

[135] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, 1994.

[136] R. R. Wilcox. *Fundamentals of Modern Statistical Methods*. Springer, New York, 2001.

[137] W. Feller. *An introduction to Probability Theory and Its Applications, Vol. I*. John Wiley & Sons, 3rd edition, 1970.

[138] K. Keyomarsi, S. L. Tucker, T. A. Buchholz, M. Callister, Y. Ding Y, G. N. Hortobagyi, I. Bedrosian, C. Knickerbocker, W. Toyofuku, M. Lowe, T. W. Herliczek, and S. S. Bacus. Cyclin E and survival in patients with breast cancer. *N. Engl. J. Med.*, 347:1566–1575, 2002.

[139] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[140] G. Getz, E. Levine, E. Domany, and M. Q. Zhang. Super-paramagnetic clustering of yeast gene expression profiles. *Physica A*, 279:457–464, 2000.

[141] K. Karuppiah, A. Ninette, G. Rechavi, J. Jakob-Hirsch, I. Kela, N. Kaminski, G. Getz, E. Domany, and D. Givol. Dna microarrays identification of primary and secondary target genes regulated by p53. *Oncogene*, 20:2225–2234, 2001.

[142] I. Kela. Clustering of gene expression data. Master's thesis, Weizmann Institute of Science, 2002.

[143] G. Getz, H. Gal, I. Kela, D. A. Notterman, and E. Domany. Coupled Two-Way Clustering Analysis of Breast Cancer and Colon Cancer Gene Expression Data. *Bioinformatics*, 2003. In press.

[144] S. Godard, G. Getz, M. Delorenzi, P. Farmer, H. Kobayashi, M. Nozaki, A.-C. Diserens, M.-F. Hamou, P.-Y. Dietrich, L. Regli, R. C. Janzer, P. Bucher, R. Stupp, N. de Tribolet, E. Domany, and M. E. Hegi. Classification of human astrocytic gliomas on the basis of gene expression: A correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Submitted*, 2003.

[145] J.-E. Dazard, H. Gal, N. Amariglio, G. Rechavi, E. Domany, and D. Givol. Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer. *Oncogene. In press*, 2003.

[146] M. Ladanyi and Gerald W. L. *Expression Profiling of Human Tumors: Diagnostic and Research Applications*. Humana Press, Totowa, New Jersey, 2003.

[147] Affymetrix$^{TM}$manuals.

[148] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, 29:137–140, 2001.

[149] M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, T. Hubbard, A. Kasprzyk, D. Keefe, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and E. Birney. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, 31:38–42, 2003.

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

[151] I. R. Cohen. The cognitive paradigm and the immunological homunculus. *Immunol. Today*, 13:490–494, 1992.

[152] Gad Elizur. Antigen chips: development and analysis. an application to autoimmune diseases. Master's thesis, Weizmann Institute of Science, 2004.

[153] W. S. El-Deiry, S. E. Kern, J. A. Pietenpol, K. W. Kinzler, and B. Vogelstein. Definition of a consensus binding site for p53. *Nat. Genet.*, 1:45–49, 1992.

[154] Yeast cell-cycle data can be obtained from http://cellcycle-www.standford.edu.

[155] M.Q. Zhang. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, 23:233–250, 1999.

[156] F. Cardoso. Microarray technology and its effect on breast cancer (re)classification and prediction of outcome. *Breast Cancer Res.*, 5:303–304, 2003.

[157] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, 33:49–54, 2003.

[158] A. Borg, M. Ferno, and C. Peterson. Predicting the future of breast cancer. *Nat. Med.*, 9:16–18, 2003.

[159] C. Venclovas, A. Zemla, K. Fidelis, and J. Moult. Comparison of performance in successive CASP Experiments. *Proteins: Structure, Function, and Genetics*, Suppl. 5:163–170, 2001.

[160] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368, 2003.

[161] F. Bernstein, T. Koetzle, G. Williams, E. Jr. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[162] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–602, 1996.

[163] J. M. Thornton, C. A. Orengo, A. E. Todd, and F. M. G. Pearl. Protein folds, functions and evolution . *J. Mol. Biol.*, 293:333–342, 1999.

[164] L. Lo Conte, B. Ailey, T. J. P. Hubbard, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: a structural classification of proteins database . *Nucleic Acids Res.*, 28:257–259, 2000.

[165] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[166] L. Holm and C. Sander. Dali/FSSP classification of three-dimensional protein folds . *Nucleic Acids Res.*, 25:231–234, 1997.

[167] L. Holm. Dali database at `http://www.bioinfo.biocenter.helsinki.fi:8080/dali`. private communication, 2003.

[168] L. Holm and C. Sander. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20:478–480, 1995.

[169] Y. Pilpel. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, 29:153–159, 2001.

[170] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat. Genet.*, 31:64–68, 2002.

[171] G. J. Hannon. RNA interference. *Nature*, 418:244–251, 2002.