Hematology, August 2007; 12(4): 271–287

DNA microarray analysis: Principles and clinical impact

MARCO WILTGEN[1] & GERNOT P. TILZ[2]

[1]*Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria, and* [2]*Clinical Immunology and Jean Dausset Laboratory, Medical University of Graz, Graz, Austria*

(Received 12 January 2007; revised 15 January 2007; accepted 5 February 2007)

Abstract

In recent years, a new technology, allowing the measurements of the expression of thousands of genes simultaneously, has emerged in medicine. This method, called DNA microarray analysis, is today one of the most promising method in functional genomics. Fundamental patterns in gene expression are extracted by several clustering methods like: hierarchical clustering, self organizing maps and support vector machines. Changes in gene expression, as a response to changing environment conditions, diseases, drug treatment or chemotherapy medications, can be detected allowing insights into the dynamic of the genome. Microarrays seem to be an important tool for diagnosis of diseases at a molecular level. Applications are for example the improvement of diagnosis and treatment of cancer and the improvement of the effectiveness of drug treatment. In this introductory paper, we present the principles of DNA microarray experiments, selected clustering methods for gene expression analysis and the impact to clinical research.

Keywords: *DNA microarray, gene expression, hierarchical clustering, self organizing maps, support vector machine, B-cell lymphoma*

Introduction

Nucleic acids (DNA, RNA) are the hereditary components of life and constitute the genom. DNA (deoxyribonucleotid acid) is a double-stranded polymer of four nucleotides: adenine, cytosine, thymine and guanine (Figure 1). The two strands interact together by hydrogen bonds between pairs of nucleotides. The information needed for protein biosyntheses is determined by the genetic code and is inherent through replication. Proteins are used by the cell to read and translate the genomic information into other proteins for performing and controlling cellular processes: metabolism (degradation and biosynthesis of molecules), physiological signalling, energy storage and conversion, formation of cellular structures.

Proteins are synthesized from the genetic code by intermediate of mRNA (messenger RNA). An RNA (ribonucleotid acid) molecule is single-stranded and can pair with DNA. RNA contains the same nucleotides as DNA with the exception that thymine is exchanged with uracil A distinct DNA sequence which codes for particular protein or more precisely for a functional or structural RNA is called a gene. Each gene carries the information needed for one or more proteins performing a specific task in a cell. To retrieve the encoded information in a gene the cells use the process of gene expression. This process consists of two steps (Figure 2). During the first step (called transcription) the DNA sequence of the gene is copied into mRNA. This mRNA serves in the second step (called translation) as template for the protein biosyntheses. A gene is said to be expressed in a cell if its corresponding mRNA is present in the cell. The amount of mRNA at a given time point in the cytoplasm of cell serves as measure for the expression level. Or in other words, the expression level reflects the activity of a specific gene and therefore the amount of the related protein needed by the cell. The complete collection of all transcripts in a cell at a given state of development is called transcryptome. By comparing

Correspondence: M. Wiltgen, Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, A-8036 Graz, Austria. Tel: 43 316 385 3587. Fax: 43 316 385 3590. E-mail: marco.wiltgen@meduni-graz.at

ISSN 1024-5332 print/ISSN 1607-8454 online © 2007 Informa UK Ltd.
DOI: 10.1080/10245330701283967

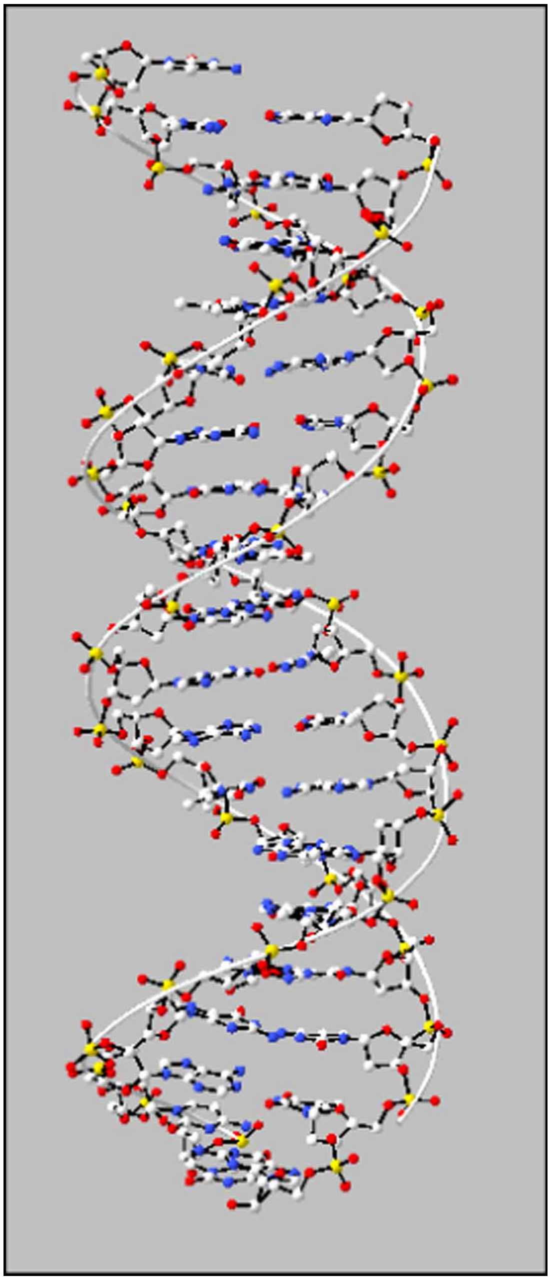


Figure 1. DNA is a double-helical polymer. The two halves of the DNA helix serve as template for replication and contain the information needed for protein syntheses. The polymer chain is composed of four nucleotides: adenine (A), thymine (T), guanine (G) and cytosine (C). The DNA code consists of patterns build up from these nucleotides. The two strands are connected by interacting base pairs (A–T and G–C). The visualization was performed by use of an atomic coordinate file from the PDB database.

the transcriptomes isolated from cells at different time points with a reference transcriptome, changes in the transcriptome levels for every gene in the genome are detected.

Insight into the functional behaviour of the genome can be obtained by determining which genes are induced or repressed in response to a step of the cell cycle, a development phase, or a response to the environment, such as treatments with drugs. Genes with similar expression behaviour under same conditions are likely to have a related function. DNA microarrays allow it to measure simultaneously the level of expression for every gene in a genome. Multiarray plates and microchip assays are used to screen hundred or thousands of gene fragments in one assay. Therefore, DNA microarrays provide insights into the dynamics of a genome or genomic shift in metabolism. This is known as functional genomics. Gene expression analysis is part of a new interdisciplinary discipline between information science and molecular biology, known as bioinformatics [1–4].

For didactic purposes, the principles of the microarray technique is illustrated by the metabolic shift of yeast (*Saccharomyces cerevisiae*) from anaerobic fermentation to aerobic oxidation. First yeast is growing in a media with sufficient glucose which is fermented into ethanol. After all the glucose has been used, the yeast cells switch from fermentation to the decomposition of the ethanol into other products (like glycogen) by use of oxygen. The 6200 genes of the yeast genome are amplified by polymerase chain reaction (PCR: a method in molecular biology to selectively amplify small amounts of DNA of given length and sequence). The PCR results are then purified and put on a glass slide. The spotted DNA is denatured (single-stranded) and linked with covalent bonds to the glass slide (these are the so called probes). First cell where a gene is expressed must be found. This is done by finding the corresponding mRNA. Cytoplasmic concentrations of mRNA are good indicators for gene expression. The DNA sequence is obtained by enzyme reverse transcriptase, which produces a DNA sequence out of the mRNA (complementary DNA or cDNA represents the coding sequence of a gene including flanking regions).

By a systematic arranging on the glass slide, a particular sequence (or a gene) can be identified by the location of the spot on the slide (Figure 3). At different time steps mRNA is isolated from the cell population and converted into cDNA (the so called targets). The nucleotides, which are used for the synthesis of the cDNA include either a green Cyanine fluorescent dye (called Cy3) or a red dye (called Cy5). Therefore the corresponding cDNAs are labelled green or red. Usually, the reference cDNA is labelled with Cy3 fluorescent dye and the test cDNA with Cy5 fluorescent dye. In the case of the yeast population, the cDNA produced during the anaerobic fermentation (the reference sample) is labelled green and the cDNA from the aerobic oxidation (the test sample) is labelled red. The sets of cDNA from two different samples (green and red) are mixed together and incubated with the single-stranded DNA on the microarray. The cDNAs will hybrids with the respective complementary DNA strand (which represent its corresponding gene). After incubation time the

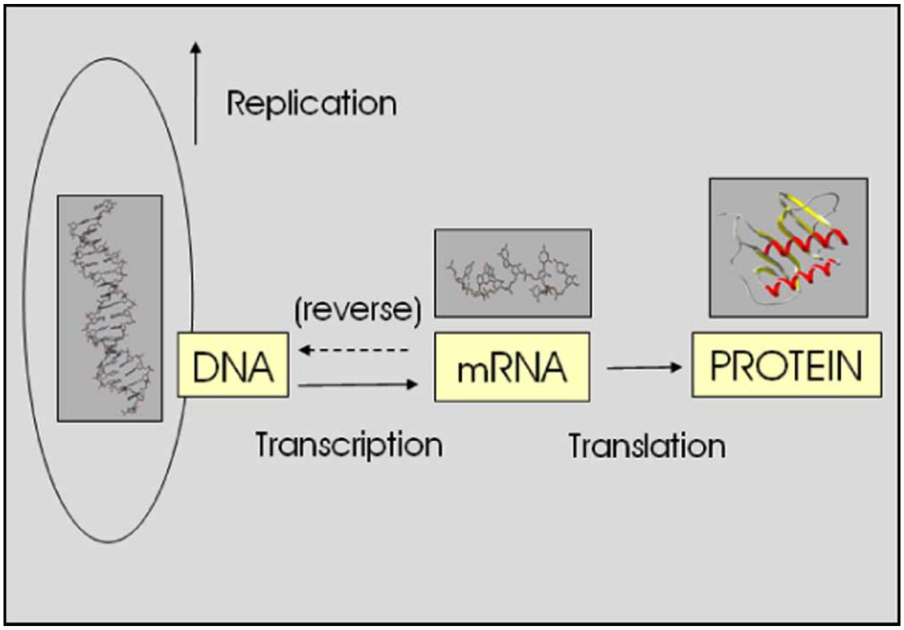


Figure 2. The genetic code in parts of the DNA, the genes, serves as construction plan for proteins. During gene expression a mRNA copy is made from the gene on the DNA. The mRNA is transported from the cell nucleus to the ribosomes in the cell cytoplasm where it serves as template for protein syntheses.

cDNAs that did not bind to any spot are washed off. Then, the location and intensities of the fluorescent dyes are recorded with a scanner. The scanner consists of lasers with different wave length and a sensor. The slide is scanned twice, first with green laser light (532 nm) that excites the Cy3 fluorescent dye and than with red laser light (635 nm) exciting the Cy5 dye. The dyes emit characteristic fluorescent radiation which is measured by the detector. From the scanning process result two digital monochrome images from the microarray, one for the green dyes and one for the red dyes.

The measured fluorescence intensity for any spot is proportional to the amount of mRNA in the probe. Mostly, the ratios (Cy5/Cy3), providing a measure for the relative intensities, are used for the analysis of gene expression. To visualize the relative gene expression the two images are pseudo-coloured and merged to a ratio image of the microarray. A red spot indicates that a gene produces more mRNA in the test probe than in the reference probe (Figure 4). Than, the gene activity is induced in the test probe. A green spot indicates that the gene has a lower activity in the test sample than in the reference sample. The gene activity is repressed in the test sample. A yellow spot (equal intensity of green and red label) indicates that there is no change in the gene activity level in the two samples. (Black represents spots where no cDNA has bound to the single-stranded DNA of the gene). If the test samples are taken at different successive time steps, the behaviour of the genes under different conditions can be studied (Figure 5). The repression and induction of the activities of the genes give insights into the dynamic behaviour of the cell at the molecular level.

Fundamental patterns in gene expression are extracted by clustering methods. These processes organize the genes into biological relevant clusters with similar expression patterns. The analysis of the gene expression data provides three distinct results:

(1) If new detected genes with previously unknown functions are clustering repeatedly with genes of known function, the function of the new genes can be predicted.
(2) Genes in a common cluster often contain conserved promoter (site on the gene for the transcription initiation and direct binding of the RNA polymerase) sequence motivs.
(3) Clustered gene expressions may be components of genomic circuits that work together and perform a single task.

Microarrays are today one of the most promising method in functional genomics. This technique offers us the possibility to determine thousands of expression values in hundreds of different conditions [5–10]. For example, the differences in gene expression profiles in a normal and a cancer cell. A well studied eukaryotic organism is yeast. Yeast as a model organism is so important because yeast cells and human cells have many genes for fundamental biological processes in common. Most of the cell-signalling systems are the same in both cells. Because mitosis is almost identical, yeast plays an important role in cancer research. A lot of microarray data sets are provided by the Stanford Microarray Database [11]. The data base stores raw and normalized data from microarray experiments and the microarray image files (Figure 6). The data are

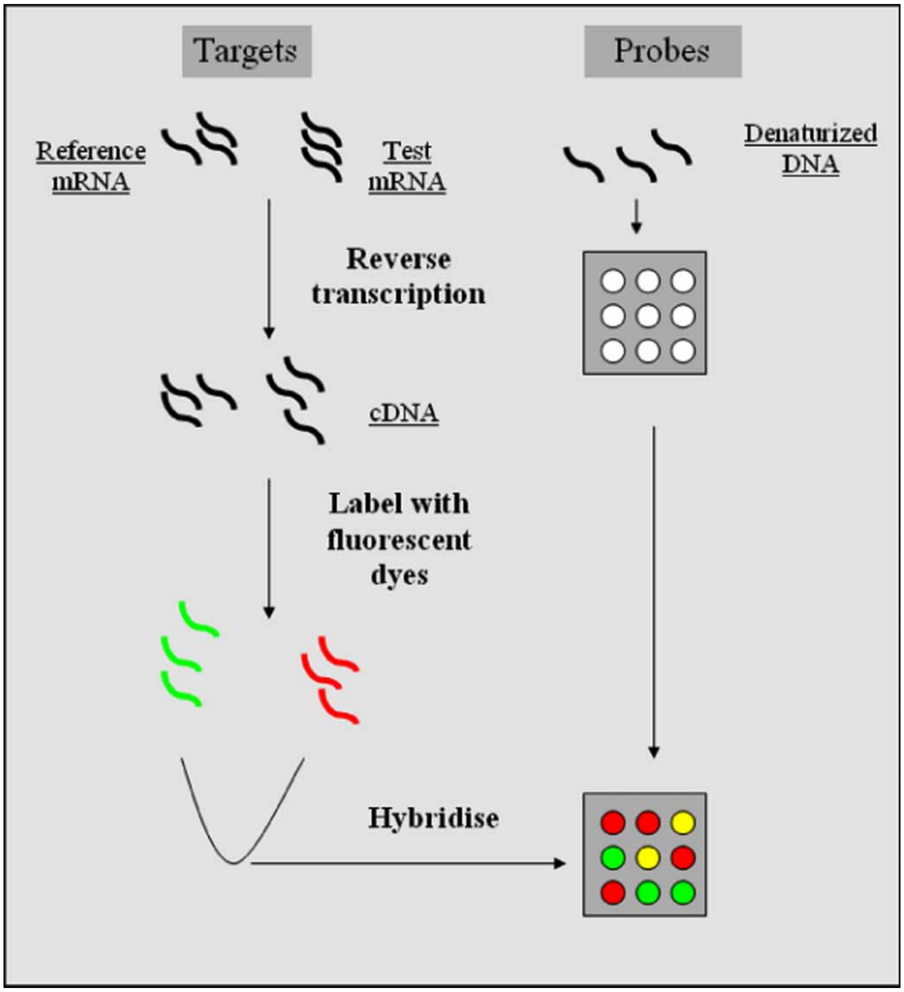


Figure 3. In a microarray experiment, first parts of single-stranded DNA (e.g. genes) are bounded in an array on a glass slide. They serve as probes. During gene expression a certain amount of mRNA occurs in cell cytoplasm. The functional behaviour of cells are studied under different conditions. In the begin reference mRNA is extracted from the cells. Then, for example after a metabolic shift due to changing conditions, mRNA is isolated from the cell population at successive time steps (test samples). The mRNA from both samples (reference and test) are converted into cDNA, labelled with different fluorescent dyes and mixed together. Then they are incubated with the single-stranded DNA on the microarray. The different cDNA hybride with the respective complementary DNA parts and the colours indicate different states of gene expression. Available in colour online.

available via Internet (http://genome-www.stanford.edu/microarray/). Microarray experiments are large scale experiments and needs an extensive study design. Several preliminary steps are needed to enable a reliable analysis and interpretation of the expression levels.

Microarray experiments

Biological and medical conclusions and predictions resulting from microarray data depend from the experimental design of the array and the reliability of the output data. Therefore prior to the gene expression analysis several procedures including the printing technique of the microarray slides, the use of replicates, the data pre-processing and normalization are necessary [12].

Microarray slides

Microrrays consist of glass slides containing DNA spots in a high density. In principle, three different techniques are used to produce microarrays, where two techniques are based on the mechanical deposition of external synthesized DNA probes (PCR products) on the glass slide and with the third method the probes are synthesized directly on the slide:

1. Microspotting: The DNA probes are spotted by a robot onto the glass slide via a micro capillary and linked by covalent bonds to the glass slide.
2. Microspraying: The DNA probes are sprayed (touch free) on the slide by inkjet printing.
3. *In situ* arrays: The DNA probes are synthesized directly on the glass slide with the use of photolithographic techniques.

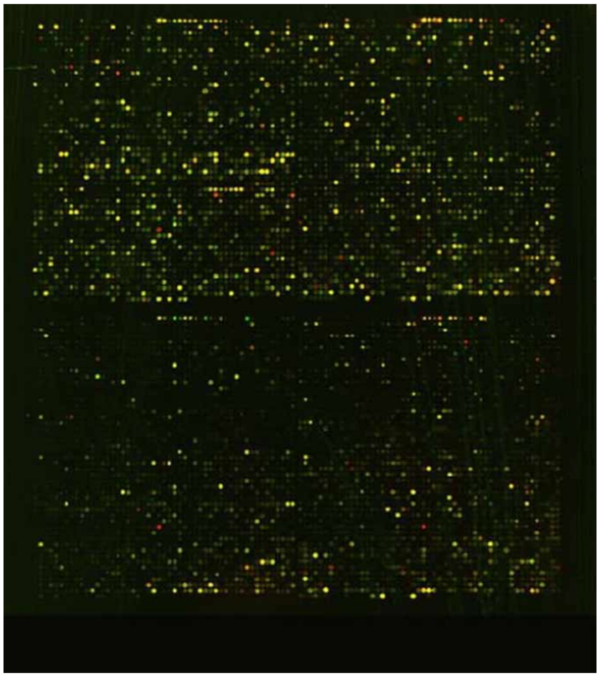


Figure 4. Scan of a micoarray. This part of a larger array illustrates one time point of a DNA microarray experiment where the gene expression in fibroblast cells was measured (fibroblast cells differentiate to different kinds of connective tissue cells and are involved in wound repair). Green dye was used for the cDNA extracted from the reference cells and red dye was used for the fibroblast cells at a given time point after stimulation with serum. Red spots indicate induced genes, green spots indicate repressed genes and yellow spots indicate no change. Available in colour online.

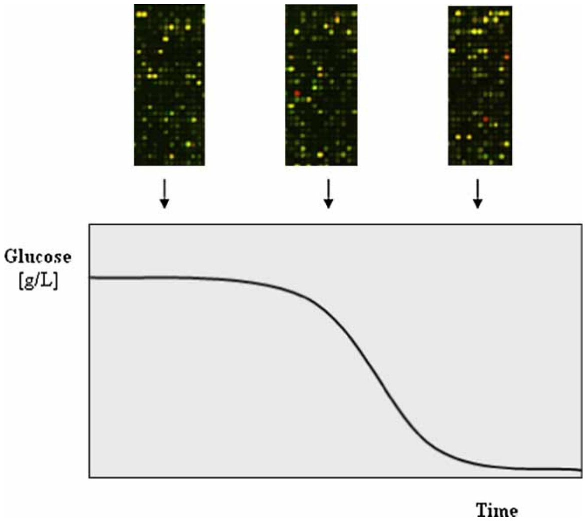


Figure 5. Microarray experiments enable the study of the gene expression of a cell under different conditions. The repression and induction of the activities of the genes during the changing conditions can be followed up. In our example, a metabolic shift of the yeast cells from anaerobic fermentation to aerobic oxidation occurs when the glucose concentration diminish. This is reflected by changing gene activity.

In this paper, we will concentrate our attention on spotted microarrays, which is based on a competitive hybridization of two mRNA samples with two different dye labels. A microarray slide may consists of 4 × 4 sectors, each containing 399 probes. That makes together 6.384 probes per array. Spotted microarrays allow a great flexibility in the choice of the array elements, especially for customized arrays for special investigations.

To get more statistically meaningful interpretation of the results, replication is necessary.

Replicates

To guarantee reliable and reproducible results different variations (random and systematic) that occur in every microarray experiment must be taken into account [13,14]. Replicates allow averaging and reduce variability in the summary statistics. The additional data from the replicates can be analyzed with statistical methods allowing an evaluation of the slide quality and the estimation of the variance between slides. The onset of the replicates is a very important step in the design of microarray experiments. There are two forms: biological and technical replicates.

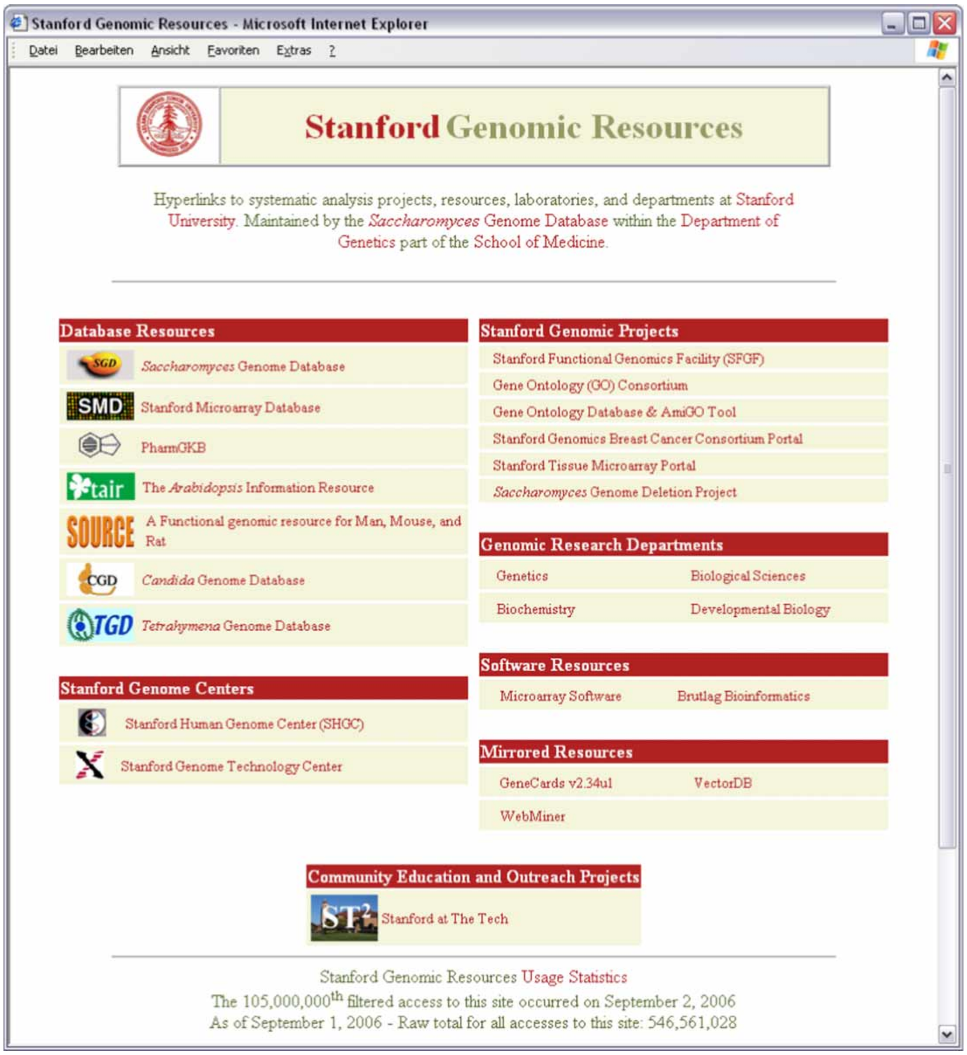
Random variations occur from biological variability inside a population. Biological replicates, where mRNA samples are taken from independent biological sources, allow averaging which reduce the variability and enable more independent experimental results.

Technical replicates are used to reduce variability introduced by measurement errors. Technical replicates can be realized by multiple hybridizations from the same sample. They are used to evaluate the technical artefacts resulting from: scanner settings, reagents, robotic printing process etc.

The technical realization and hybridization is just one step in a pipeline consisting of: pre-processing, normalization, data analysis and data interpretation. The next point is the pre-processing of microarray data.

Pre-processing of microarray data

To eliminate potential sources of error at the beginning a pre-processing of the microarray data is necessary [15]. First, spots with insufficient quality must be removed from the data set. These are spots showing either a very low expression value (which cannot be sufficiently distinguished from background) or saturated spots (which cannot provide a reliable intensity measurement). Second, to improve the accuracy of the measured values a background correction is necessary. The fluorescence of the background material is generally added to the spot intensity during the scanning process. To determine the true intensity value of the spot, the background intensity must be removed. This can be done by



Stanford Genomic Resources

Hyperlinks to systematic analysis projects, resources, laboratories, and departments at Stanford University. Maintained by the *Saccharomyces* Genome Database within the Department of Genetics part of the School of Medicine.

Database Resources

- *Saccharomyces* Genome Database
- Stanford Microarray Database
- PharmGKB
- The *Arabidopsis* Information Resource
- A Functional genomic resource for Man, Mouse, and Rat
- *Candida* Genome Database
- *Tetrahymena* Genome Database

Stanford Genome Centers

- Stanford Human Genome Center (SHGC)
- Stanford Genome Technology Center

Stanford Genomic Projects

- Stanford Functional Genomics Facility (SFGF)
- Gene Ontology (GO) Consortium
- Gene Ontology Database & AmiGO Tool
- Stanford Genomics Breast Cancer Consortium Portal
- Stanford Tissue Microarray Portal
- *Saccharomyces* Genome Deletion Project

Genomic Research Departments

Genetics	Biological Sciences
Biochemistry	Developmental Biology

Software Resources

Microarray Software	Brutlag Bioinformatics

Mirrored Resources

GeneCards v2.34u1	VectorDB
WebMiner	

Community Education and Outreach Projects

Stanford at The Tech

Stanford Genomic Resources Usage Statistics
The 105,000,000th filtered access to this site occurred on September 2, 2006
As of September 1, 2006 - Raw total for all accesses to this site: 546,561,028

Figure 6. Access to microarray data (microarray dataset from Stanford University) is possible via the Stanford Genomic Resources homepage: http://genome-www.stanford.edu. For the following illustrations, we choose an array dataset from fibroblasts response by stimulation with a serum. Fibroblasts are cells playing a role in wound repair.

subtraction of the background intensity from the expression intensity.

The raw data from a microarray after the scanning process consist of pairs of image files (one for each dye). Automatic analysis systems are used for the extraction of the red and green intensities for each spot on the array. First, the location of the spot centres must be detected. Spots usually vary in size and shape. Therefore segmentation is needed to determine which pixel on the array image belongs to the spot (foreground) or to the background. Than for each spot on the array and each dye, beside spot and background intensities, quality measures are extracted. The spot quality is defined by: brightness, uniformity and the area. The quality can be related to the standard deviation of the pixels of the spot, filtering out those spots for which the standard deviation is to big or if the values for mean and median are to different.

But before an analysis of the gene expression levels can be done, several systematic variations of the measured levels must be removed. The removing concerns all the non-biological variations introduced in the measurements. This task refers to the process of normalization, making microarray data comparable.

Normalization

Normalization is necessary for within and between slides comparisons [16]. The need for normalization can be shown with self–self hybridizations where the same mRNA probes are labelled with Cy3 and Cy5 dyes. The scatter plot shows an imbalance of green and red intensity values. The green intensities tend to be higher then the red ones. The dye effect results from different properties of the two dyes such as: lower incorporation rate of Cy5, the quantum yield and the photobleaching. Normalization should be used to correct the dye-imbalance. Normalization is used in different ways: within-slide normalization and between slides normalization.

Within-slide normalization. First, each of the slides must be normalized separately before normalization between slides can be done. Within-slide normalization can be regarded as correction of the dye effect. The simplest normalization method (global normalization) is based on the assumption that the red–green imbalance is constant for all spots of the array. Then Cy3 and Cy5 are related by a constant factor. Correction of the ratios is done by simply subtracting a constant in that way, that the mean of the log-ratios is zero. In the global normalization approach is assumed that the correction parameter is constant across the whole intensity range. The experience shows however that the dye effect is intensity dependent. Locally weighted linear regression can be used to remove such intensity dependent bias by subtraction with (intensity dependent) values of an estimated function.

Another method to normalize within-slides is the paired slide normalization with dye swap. A dye swap pair consists of two slides, where hybridization is done twice with reverse dye assignment. Then, by assuming that the dye effects for the two slides are similar, normalization is done by combining the log-ratios of the two slides.

Between-slides normalization. After within-slide normalization, the single slides contain log-ratios with values centred around zero. Between slides there are often scale differences due to changes in scanner settings or similar influences. Therefore, after within-slide normalization the log-ratios of a series of slides must be scaled to provide comparable values across different biological conditions.

Representation of expression data

Expression matrix. After normalization the data are typically reported as expression ratios. Let us say that N genes are simultaneously probed on a microarray. Further, let us assume that we have a series of M arrays, each representing one experimental condition (sample of mRNA at a given time step). Out from the data on the microarray plates, a matrix is constructed where the genes are arranged in rows and the expression values for each experimental condition are listed in columns (Figure 7). Then the vector of the ith row in the matrix describes the expression values of the ith gene during the M time steps where the samples are taken.

$$X = (x_{ij})$$

If the expression values are represented as simple ratios then the values of the over expressed genes are lying in the range $(1 < x_{ij} < +\infty)$ and the under expressed genes are represented by the range $(0 \leq x_{ij} < 1)$. To overcome this discrepancy with different range lengths the logarithm of the ratios is taken.

$$x_{ij} = \log\left(\frac{(Cy5)_{ij}}{(Cy3)_{ij}}\right)$$

$(Cy5)_{ij}$, Red fluorescence value of gene i in the sample j; $(Cy3)_{ij}$, Green fluorescence value of gene i in the sample j.

Then, the over expressed genes are assigned to positive values and the under expressed genes to negative values, where equal changes are of equal magnitudes (with opposite direction). Gene expression at a constant level is represented by zero.

Visualization. To visualize the primary data of the expression matrix, the numerical ratios are represented by a colour that reflects qualitatively and quantitatively the experimental observation. This enables to explore the data in an intuitive manner. The table for the pseudo-colour ranges from green to red. If a gene was induced at a given time point it is represented by the red colour. The greater the induction, the brighter is the red colour. If a gene is repressed the value is represented by the green range. Values with logarithmic ratios close to zero are coloured black. The relative intensities represent relative expressions for each gene at a given time point (experiment) in the expression matrix, where brighter elements are highly differently expressed.

If the genes are related, the expression progression along different time steps and conditions can be plotted (progression plot). This makes of course only sense after clustering when the progressions of the genes in a single cluster are plotted. If the expression patterns of many genes are plotted simultaneously, the presentation of general and common patterns becomes difficult. Overall, properties can be visualized by plotting the mean and standard deviation of all genes in the cluster.

Gene expression analysis

Data analysis means the identification and clustering of common patterns of gene expression. The patterns allow conclusions about the common behaviour of genes under different conditions (cell cycle, different development phases, or a response to treatments with

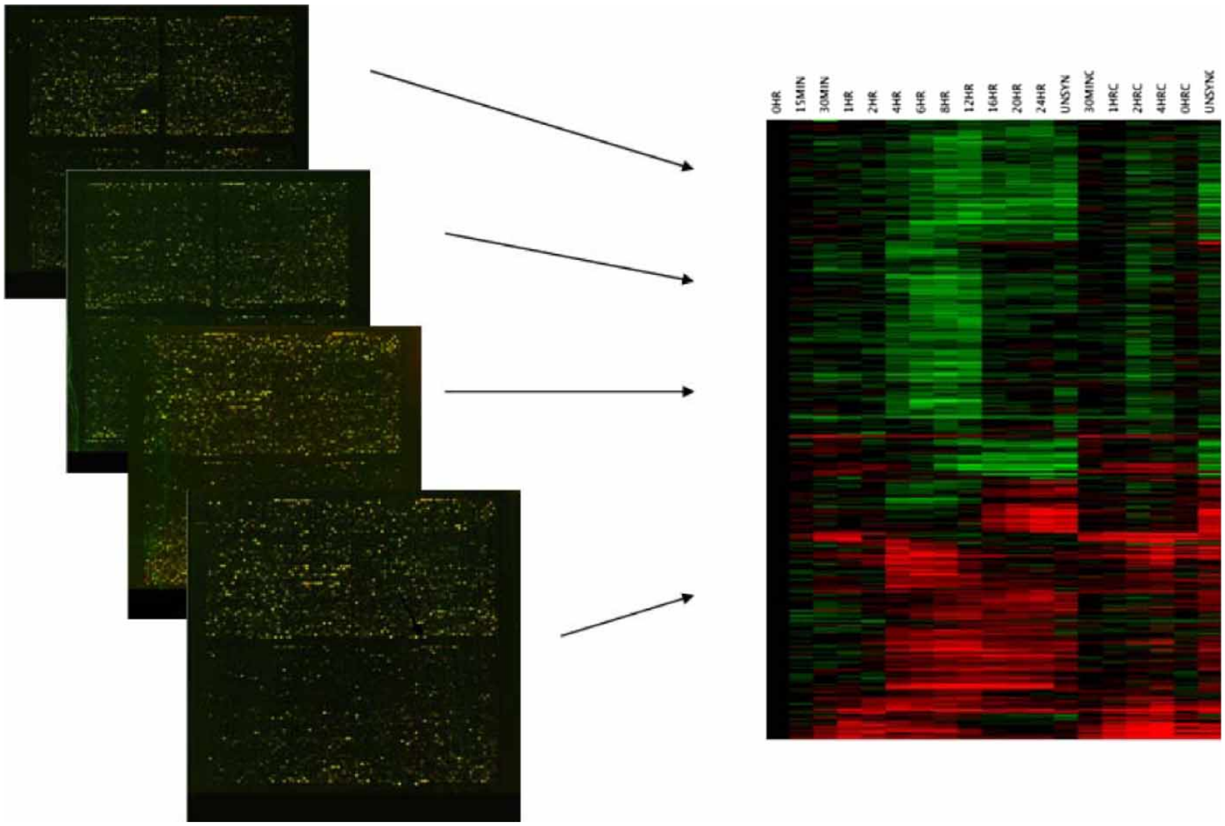


Figure 7. The results from the microarray slides are reported as gene expression matrix. The series of microarrays probe the expression levels of the fibroblast cells in different samples. In the expression matrix, the genes are arranged in the rows and the columns represent the expression values at different time points.

drugs etc.). Goal is the detection of similarities or differences [17]. The expression matrix can be studied by either comparing rows or columns. The rows or columns of the expression matrix can be considered as vectors.

$$y_i = (x_{i1}, \ldots, x_{iM}) \quad i = 1, \ldots, N$$

Similarities between rows possibly result from co-regulation of genes. These genes can be functionally regulated. Comparison of columns (experiments) provides insights into the behaviour of the whole gene set I under different conditions.

$$I = \{y_1, \ldots, y_N\}$$

Detection of differentially expressed genes allows it to study the response of various compounds to the investigated conditions (for example: cell lines of tumour samples).

Data adjustment

Prior to any data analysis some data adjustments are necessary. To enable a better detection of relationships, the data vectors (rows in the data matrix) must be rescaled. One procedure is the subtraction of the mean value (along a row) from each data element.

$$y'_i = y_i - \bar{y}_i \quad \text{with :} \quad \bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} x_{ij}$$

After this rescaling the expression values of each gene reflect the variation from the mean value. This is a kind of standardization where the mean of each row will be zero. This allows it to remove certain types of bias (resulting from multiplying the data elements of all the genes by a fixed value).

$$y'_i = \frac{y_i}{\sigma_i} \quad \text{with :} \quad \sigma_i = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(x_{ij} - \bar{y}_i)^2}$$

Additionally, to amplify weak expression signals and to suppress strong signals, the values of the data vectors are divided by their respective standard deviation.

Similarity and distance measures

To enable a comparison of gene expression patterns, first a measure for the similarity between the expression data must be defined.

$$d : I \times I \rightarrow \mathbb{R}$$

where the coefficients are satisfying the conditions:

$$d_{nn} = 0 \quad \text{and} \quad d_{nm} \geq 0$$

$$d_{nm} = d_{mn}$$

The similarity measure is computed by summing up the distances between the respective vector elements [18]. With such similarity measures, the problem of finding common expression pattern reduces to a pair-wise linear comparison of the data vectors. Representations from a great variety of such measures are the Euclidian distance and the Pearson correlation coefficient.

Euclidian distance:

$$d_{nm} = \|y_n - y_m\| \quad = \sqrt{\sum_{j=1}^{M} (x_{nj} - x_{mj})^2}$$

Pearson correlation coefficient:

$$p_{nm} = \frac{\sum_{j=1}^{M} (x_{nj} - \bar{y}_n)(x_{mj} - \bar{y}_m)}{\sqrt{\sum_{j=1}^{M} (x_{nj} - \bar{y}_n)^2} \sqrt{\sum_{j=1}^{M} (x_{mj} - \bar{y}_m)^2}}$$

This allows the calculation of a distance (or similarity) matrix, which is the input for the clustering algorithms. More sophisticated approaches are Spearman Rank-Order correlation, Kendall's Tau and Mutual Information.

Clustering methods

Clustering is the task of separating a set of data into several subsets due to similarities. The aim is to find a partition P in which the data vectors in the same cluster C_i are similar to each other and dissimilar to the data vectors in the other clusters C_j.

$$P = \{C_1, \ldots, C_g\}$$

where the classes $C_l = \{\ldots, y_k, y_h, \ldots\}$ are satisfying the following conditions:

i) $\bigcup_{k=1}^{g} C_k = I$
ii) $C_i \cap C_j = \varnothing$ with: $i \neq j$

Generally, there exist several possibilities for the partitioning of a set of N data vectors into g disjoint clusters. Therefore from the set $\Gamma^\star$ of all possible partitions, the optimal partition $\hat{P}$ satisfying to the best the quality criteria h:

$$h : \Gamma^\star \rightarrow \mathbb{R} \quad \text{with :} \quad h(\hat{P}) = \max_{P \in \Gamma^\star}(h(P))$$

is selected. The goal is to find clusters that minimize intracluster variability while maximizing intercluster distances. This can, as an example, be realized with the following quality criteria:

$$h(P) = \sum_{k=1}^{g} \sum_{n,m \in C_k} d_{nm} \rightarrow \min_{P \in \Gamma^\star} \quad \text{where :} \quad d_{nm}$$
$$= \|y_n - y_m\|$$

The motivation to find clusters is driven by the assumption that genes showing similar expression patterns have common functions, common regulatory elements or common cellular origin [19,20].

Two clustering strategies are possible: unsupervised and supervised. Unsupervised methods allow the analysis of the gene expression data set without an *a priori* knowledge or input. Supervised methods determine expression patterns that fit a predetermined pattern resulting from a previous training.

For the following illustrations, we choose a microarray dataset from the Stanford Genomic Resources at the Stanford University. The dataset results from fibroblasts response by stimulation with a serum [21]. Fibroblasts are cells playing a role in wound repair. The clustering operations were performed with the Gene Expression Similarity Investigation Suit (GENESIS) software [20]. The program runs on a PC with Microsoft Windows.

Unsupervised methods

As representatives algorithms for gene expression analysis we present hierarchical clustering and self organizing maps. These are the most popular methods in finding trends in gene expression data.

Hierarchical clustering. Hierarchical methods produce nested clusters, where small clusters are nested inside larger ones [22]. The clusters are showing different levels of detail, depending on the actually considered partition P^ν of clusters C_l.

$$P^\nu = \{C_1, \ldots, C_g\}$$

First the distance matrix, containing the distances between all data vectors is calculated. The hierarchical cluster algorithm can be formulated as follow (agglomerative procedure):

Step 1: Every data vector represents at the beginning one cluster:

$$P^0 = \{\{y_1\}, \ldots, \{y_N\}\}$$

Step 2: Find the two clusters $C_k, C_l \in P^{\nu-1}$ with minimum distance defined by the linkage rule.

$$h_\nu = \min_{\substack{C_k, C_l \in P^{\nu-1} \\ k \neq l}} D(C_k, C_l)$$

Step 3: Merge the two clusters

$$C_t = C_k \cup C_l \text{ with } C_t \in P^{\nu}$$

Step 4: Repeat steps 2 and 3 until the total number of clusters is one.

$$P^n = \{I\}$$

A common linkage rule is the single linkage, where the distance between two clusters is defined by the two closest data vectors (nearest neighbour) in the different clusters.

$$D(C_k, C_l) = \min_{\substack{n \in C_k \\ m \in C_l}} (d_{nm})$$

The distance may be defined as the Euclidian distance. The value h_v is called the index of the hierarchy (stands for the hierarchical level) and satisfies the condition:

$$0 = h_0 \leq h_1 \leq h_2 \leq \cdots$$

The hierarchical clustering produces a representation of the data as a dendrogram (tree structure), where similar expression patterns are nested in a hierarchy of sub clusters (Figure 8). Hierarchical clustering is a commonly used procedure for gene expression analysis. A great advantage is that only a few parameter (linkage rule and distance measure) need to be specified. The result is a reordering of the genes, where genes with similar behaviour are close to each other in the tree structure. The different classes must be determined by the user by selecting sub trees from the dendrogram (generally to this purpose information not resulting from classification procedure is necessary). Other linkage rules which are commonly used are: complete linkage (distance between two clusters is defined by the two data vectors with the greatest distance) and average linkage (distance between two clusters is defined by the distance of the mean data vectors).

Self organizing maps. Self organizing maps are partitioning algorithms which project high dimensional data into clusters on a low dimensional regular grid [23,24]. Clusters that are similar to one other appear in adjacent cells of the grid. The self organizing map algorithm is based on competitive learning where the training is completely data driven.

A self organizing map is formed by points located on a regular (1 or 2 dimensional) grid. Each point on the grid is represented by an M-dimensional weight vector m_l where M is the dimension of the input vectors.

$$m_l = (m_{i1}, \ldots, m_{iM})$$

The number of points on the grid represents the number of expected clusters. Prior to the training phase the weight vectors are initialized. This can be done by choosing random sample data vectors from

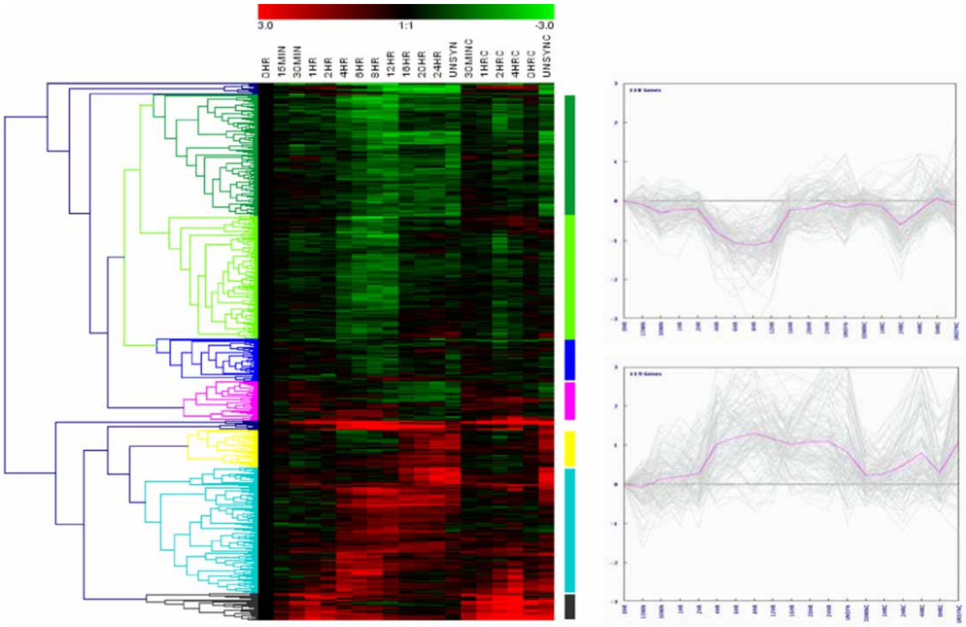


Figure 8. Left side: Dendrogram representation of hierarchical clustering. The dendrogram is a tree like representation in which each cluster is nested into the next cluster (similar clusters are fused into the next level cluster). On the right side, two progression plots of gene expression values inside two different clusters are shown. The progression is shown along different time steps from the beginning of serum stimulation. Top: the genes in the upper common cluster are repeatedly under expressed. Bottom: The genes are over expressed. Additionally, the mean expression value is shown.

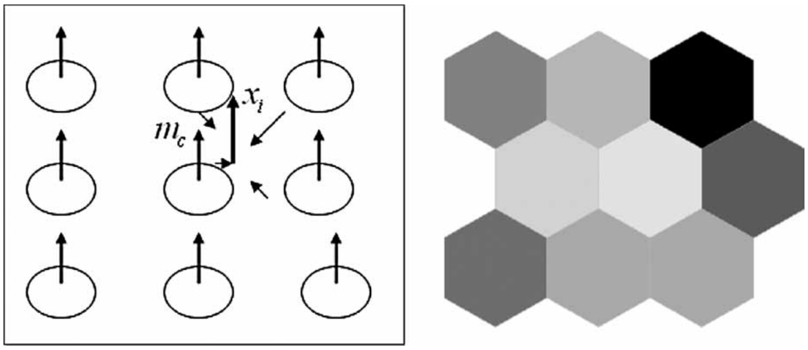


Figure 9. Representation of a self organizing map with 9 clusters. Left: A self organizing map is formed by points located on a regular grid, where each point is represented by a weight vector. The number of points on the grid represents the number of expected clusters. Prior to the training phase the weight vectors are initialized. In each training step, the distance between a randomly chosen sample data vector and all the weight vectors is determined. The point m_c whose weight vector has the minimum distance to the sample vector is selected as best matching unit. Then the best matching unit and its neighbours are updated and move towards the sample data vector. Right: The maximum distance between the weight vector of a unit and its cluster vectors is coded in grey levels. The greatest cluster size is coded in black, the smallest in white.

the training set. In each training step, the distance between a randomly chosen sample data vector and all the weight vectors is determined. The point m_c whose weight vector has the minimum distance to the sample data vector y_i is selected as best matching unit.

$$\|y_i - m_c\| = \min_l(\|y_i - m_l\|)$$

After the determination of the best matching unit, the weight vector of the unit and its topological neighbour are update (Figure 9). This update (at step t) provides that the vectors are moving closer to the sample vector.

$$m_j(t+1) = m_j(t) + h_{cj}(t)[y_i - m_j(t)]$$

The neighbours of the best matching unit are defined by the neighbourhood function. Only the weight vectors in that area around the best matching unit are updated in one step. This enables it that similar clusters are moving step by step together and finally are lying near each other on the map. The neighbourhood function is defined by:

$$h_{ci}(t) = \alpha(t)h(\|r_c - r_j\|, t)$$

where $\alpha(t)$ is the learning function and r_j the location on the grid. The function $h(\|r_c - r_j\|, t)$ can be realized as Gaussian function, where the different neighbours are considered by different weights, depending on their distance from the best matching unit.

The algorithm can be formulated as follow:

Step 1: Choose the number of columns (q_1) and the number of rows (q_2) on the grid. The predefined number of clusters is then $K = q_1q_2$.
Step 2: Define the learning function α and the neighbourhood function $h(d)$.
Step 3: Initialize the weight vectors m_j with $j \in (1, \ldots, q_1) \times (1, \ldots, q_2)$.

Step 4: Loop over the entire data set

a. Loop over every data vector y_i

a.1. Find the weight vector m_c with the minimum distance to y_i

$$\|y_i - m_c\| = \min_l(\|y_i - m_l\|)$$

a.2. Identify the set N_c of neighbour weight vectors m_c

$$N_c = \{m_n | \, \|r_n - r_c\| \leq d\}$$

a.3. Update every neighbour weight vector from N_c:

$$m_n' \leftarrow m_n + \alpha(m_n - y_i)$$

b. Decrease the values of α und d by a predetermined amount.

The algorithm maps the sample vectors into the Voronoi regions where each corresponds to one unit on the grid. Than the data vectors are clustered according to:

$$V_l = \{y_i | \|m_l - y_i\| < \|m_k - y_i\|, l \neq k\}$$

Each Voronoi region is one cluster consisting of similar gene expression data (Figure 9). The clustering structure of a self organizing map can be visualized by displaying distances between vectors. Similar clusters are lying near each other on the map.

Self organizing maps need a previous definition of the grid points and therefore the number of expected clusters. By decreasing the number of grid points, clusters with greater variability of their members are achieved. If no neighbour points of the grid point are considered, that means if there is no linkage between neighbouring points on the grid, than the process reduces to a k-means type of clustering.

Supervised methods

The above discussed clustering methods attempt to classify expression patterns in an unsupervised fashion. Supervised methods needs a previous learning procedure with a teacher signal. As a representatives algorithm for gene expression analysis we present the support vector machine.

Support vector machine. The support vector machine is a binary classification method to discriminate two different data sets [25–27]. Support vector machines use a training set to specify in advance, which data vectors should cluster together. In microarray analysis, a set of genes showing a distinct expression pattern is selected. These genes may be part of a common pathway. Additionally, a second set of genes that show a different expression pattern is selected. These two sets are used as training set and the support vector machine is trained to discriminate between the members and non-members of the given functional set. The two sets are labelled with (+) and (−). In the following, we restrict the discussion to linear separable cases.

In principle, the support vector machine tries to find a hyperplane, which separates the two classes in the training set:

$$wy + b = 0$$

w is the normal vector of the hyperplane and y are the data vectors on the hyperplane with the dot product wy. Additionally, the support vector machine tries to maximize the margin of the hyperplane. The margin of the hyperplane is the sum of the distances of the nearest vectors on both sides (Figure 10). Than the problem is to find a pair of hyperplanes in that way that the two classes of the training set must satisfy the conditions:

$$H_1 : wy_k + b \geq 1$$

for the set of data vectors y_k with label (+)

$$H_2 : wy_l + b \leq -1$$

for the set of data vectors y_l with label (−)

The two hyperplanes are parallel and no data vectors of the training set are laying between the two hyperplanes (for linear non separable cases the constraints on the two hyperplanes can be relaxed, so called soft margin). The data vectors lying on the two hyperplanes:

$$H_1 : wy_s + b = 1 \quad H_2 : wy_s + b = -1$$

are called support vectors. The margin is then determined by this pair of parallel hyperplanes. The perpendicular distances d_1 and d_2 of H_1 and H_2 from the origin come to:

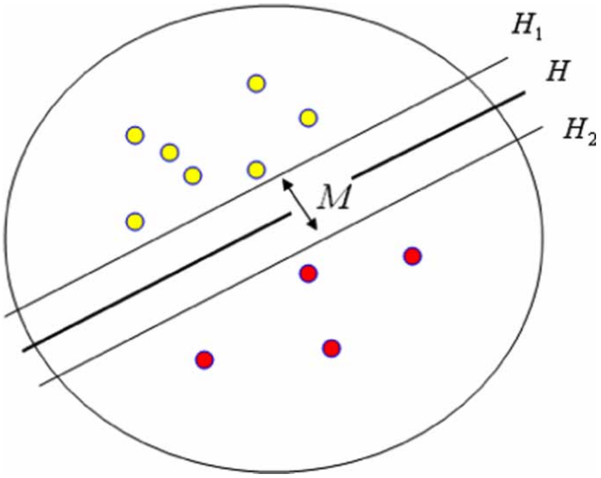


Figure 10. Principle of the support vector machine. The algorithm tries to find a hyperplane H, which separates the two classes of data vectors in the training set. Additionally, the support vector machine maximizes the amount of the margin M of the hyperplane, defined by the sum of the distances of the nearest data vectors on both sides. The data vectors lying on the two parallel hyperplanes (H_1 and H_2) are called support vectors.

$$d_1 = \frac{|1 - b|}{\|w\|} \quad \text{and} \quad d_2 = \frac{|-1 - b|}{\|w\|}$$

Then the amount of the margin is calculated as:

$$M = \frac{2}{\|w\|}$$

The principle of the support vector machine algorithm is then to minimize $\|w\|^2$ subject to the constraint:

$$t_i(wy_i + b) - 1 \geq 0 \text{ with } t_i = 1 \text{ for } H_1 \text{ and } t_i = -1 \text{ for } H_2$$

where the two hyperplane inequality equations are combined into one. If it is not possible to find a hyperplane, separating the two sets, the support vector machine can be generalized to a non linear support vector machine where a hyperplane can be found in a higher dimensional space. The crucial point is that the data vectors appear in the optimization problem as dot product, so it is not necessary to formulate the higher dimensional space explicitly.

After the hyperplane with optimal margin is determined in the training phase, the support vector machine can recognize unknown genes and classify them as members or non-members on hand of their expression pattern.

Post clustering analysis

The clustered microarray data provide the information needed to make testable predictions. Additionally to the gene expression pattern, sequence and annotation information contained in gene databases can be used for further analysis.

The relationships among genes are used to analyse there role in transcription regulation or in metabolic processes. The genome contains, beside coding sequences, regulatory regions that control gene expression. Clustering known genes and unknown genes with similar expression patterns, lead to the predictions of possible functions of the unknown genes. This "guilt by association" method reveals two possibilities: either the genes are involved in the same cellular process or the genes are induced by a common transcription factor. Than the promoter regions of the clustered genes are examined and possible conserved sequences detected. By searching for these sequences on the genome it is possible to identify the transcription factor binding sites responsible for the regulation of these clustered genes.

Beside promoter analysis the gene expression data are used for the mapping onto the chromosomes or biological pathways. Consecutive genes are often similar expressed (due to a common regulatory network) and can be easily identified. High expression and correlation values of certain genes can provide information about abnormal amount of chromosomal material in a cell (aneuploidy). Further genetic networks (for example: Bayesian networks), developed out of the microarray data can help detecting interaction between genes.

Protein biosynthesis

Goal of the gene expression is the biosynthesis of proteins. The amino acid sequences are assembled on the ribosome during the translation process by use of mRNA as a template (Figure 2). Twenty different amino acids are involved as elements in protein sequences, where each is coded by a triple of nucleotides on the mRNA. Once an amino acid sequence is synthesized it folds together to a well defined and for its sequence unique 3D structure. It can be differentiated between the primary structure (the sequence of the residues), the secondary structure (α-helices, β-sheets and loops) and the tertiary structure (folding of the secondary structure elements into a three dimensional structure).

Functional specificity of a protein is linked to its structure. Due to the folding the residues, which are responsible for the protein function, are brought into a precise geometric arrangement. The interactions between proteins within the organism result in metabolism, reproduction and form. Some proteins, like hormones, induce the expression of genes (Figure 11), while others are directly involved in the initiation and transcription processes (Figure 12).

Protein structures are determined experimentally by crystallographic methods and are deposited into the protein database (PDB), an international repository for structure files [28]. At the moment PDB (http://www.rcsb.org/pdb) contains more than 35,000 known protein structures. A PDB data file contains the coordinates of all the atoms of the proteins and is used as input for protein visualization. In this paper, we used the Swiss-Pdb Viewer for the visualization of DNA and proteins [29].

Genetic networks and reverse engineering

The synthesis of proteins is complicated and emphasis different tasks like: transcription controlling, RNA splicing, transport of mRNA, translation controlling, posttranslational modifications and degradation of protein products. Because in DNA microarray experiments only gene expressions (or more precise the production of mRNA) are measured, the modelling of cellular processes reduces to the modelling of transcription processes.

Processes inside cells are described as networks of gene products like mRNA and proteins. A genetic network describes the reciprocal regulation (Figure 11) of the involved genes. The interactions result from the expression of a gene and the activation of other genes by the produced protein (Figure 12). The state of a cell at a given time step is defined by its gene expression level which determines the behaviour of the cell (at the molecular level) at the next time step. In other words, the temporal changes of gene expression levels are modelled.

By reverse engineering a genetic network is estimated from the experimental data. A collection of data may results from the measurements of gene expression levels at different time steps (reflecting for example the temporal development of a cell or its response reaction to the stimulation with drugs). The results at a given time step describe the gene expression level and therefore the state of the cell at this time point. Then, the values of the gene expression levels at the next time step should be determined by the rules and parameters of the network.

The genetic networks are based on the assumption that every gene has a regulative influence on the other genes and this influence can be estimated by different weighting factors. The values of such weighting factors are positive in the case of a stimulating influence, zero if a gene is not influenced by other genes and negative if they are inhibited. Such kinds of genetic networks are described mathematically by differential equations of the following kind:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \sum_j w_{ij}x_j(t) + \sum_k v_{ik}u_k(t) + b_i - \lambda_i x_i(t)$$

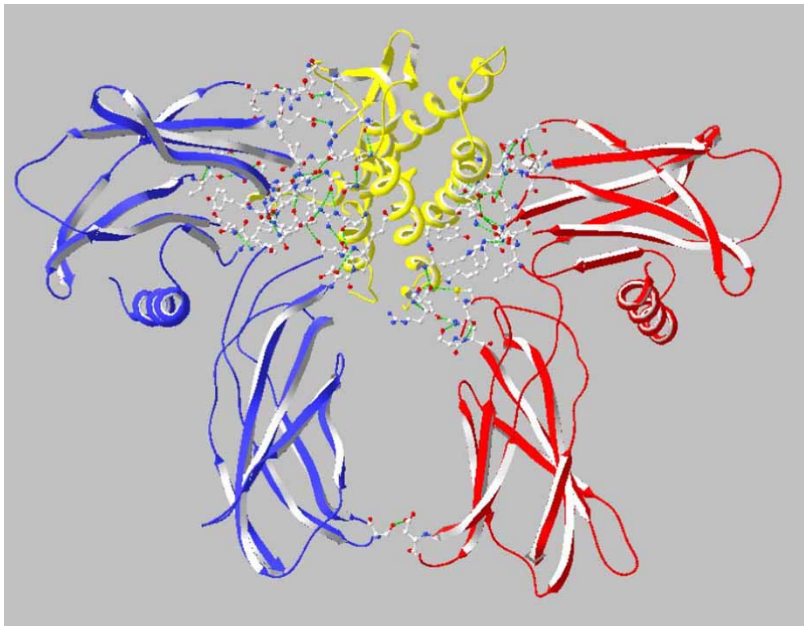


Figure 11. Growing factors are extra cellular proteins with hormonal function, which stimulate the gene expression by activating receptors localized in the cell membrane. Erythropoietin (yellow) is a growing factor (with 4-helices bundle structure) involved in the differentiation of spinal cord stem cells into erythrocytes. The binding of the growing factor causes a dimerisation of the receptor (blue and red). The dimerisation of the extra cellular part (the binding part) leads to a dimerisation of the intra cellular part which provides the really signal. This activates further interaction chains resulting in the initiation of transcription processes and gene expression. The visualization was performed by use of an atomic coordinate file from the PDB database. Resolution: 1.90 angstroms. The interacting residues are visualized in a ball-and-stick representation and the rest of the growing factor and receptors as ribbons. Available in colour online.

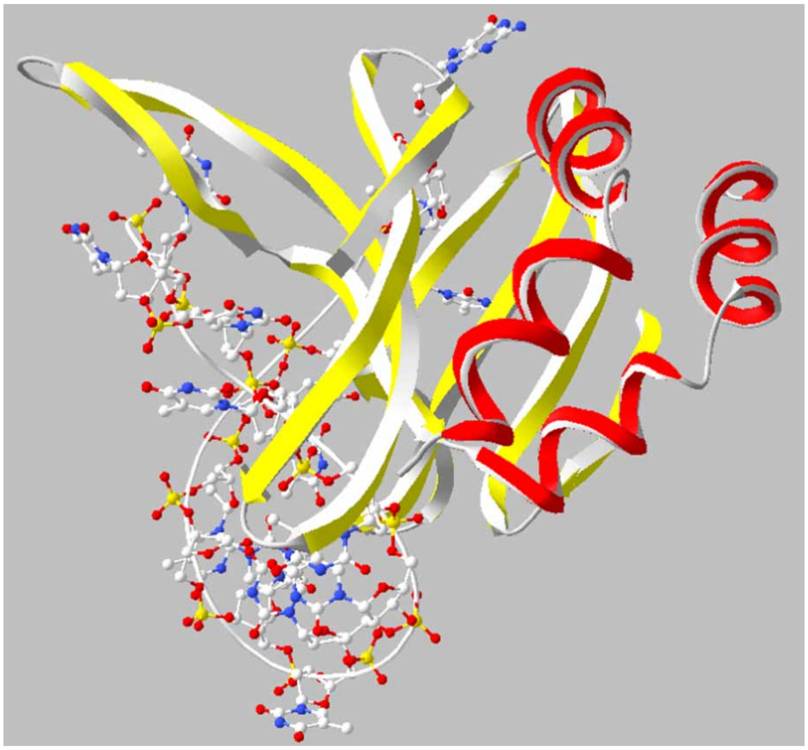


Figure 12. Crystal structure of human transcription cofactor PC4 interacting with a part of a single-stranded DNA. Resolution: 1.74 angstroms. The cofactor is part of a protein complex directly involved in the transcription process which consists of the steps: initiation, elongation and termination. The DNA segment is visualized in a ball-and-stick representation (covalent bonds are represented as sticks between atoms, which are represented as spheres). The protein (transcription cofactor) is represented as a ribbon, showing α-helices and β-sheets as secondary structure elements. The visualization was performed by use of an atomic coordinate file from the PDB database.

The variable $x_i(t)$ describes the gene expression of gene i at time point t. The weighting factor w_{ij} considers the influence of gene j on gene i. The parameter v_{ik} considers the influence of an external stimulus $u_k(t)$ (stimulation with drugs, nutritive substances, hormones etc.) on gene i at time point t. The value b_i describes the basic expression level of gene i and λ_i is the degradation constant of the i-te gene product.

Such equations describe linear models and are used for the modelling of the entire gene interactions in a cell [30]. Beside the described gene network, also Boole networks (based on a binarisation of the expression values) and Bayesian networks (where the gene regulations are described by probability distributions) are used [31,32].

Goal of the genetic networks is to obtain an, as far as possible, realistic model of the cellular processes at a molecular level. This offers the possibility to simulate the interaction of drugs and to model different therapies, which enable the re-establishing of an optimal gene expression status, at the computer. This is of special importance for the clinical and pharmacological research.

Clinical research with DNA microarrays

DNA microarrays are powerful tools to improve the quality of medicine. Microarrays seem to be an important tool for diagnosis of diseases [33–37]. Of special interest is how gene expression is changed by various diseases. Clinical application of DNA microarrays emphases cancer microarrays and the improving of health care (in common specific cancer chemotherapies the efficacy can as low as 25%). They allow it to compare the effectiveness of a drug treatment for different types of cancer and to explore the genomic responses to drug treatments (pharmacogenomics). Compared are gene expression profiles in healthy state, disease state and after drug administration. Main goals are: identification of disease associates genes and expression profiles, identification of variations in genes that affect individual response to drugs. Medical benefits are improved diagnosis and treatment. (Common drug treatments are effective only for 50–75% patients, [38].)

Improve diagnosis and treatment of cancer

A well known study with micrarrays is the diagnosis of diffuse large B-cell lymphoma, which is a very aggressive malignancy of B-cells. Certain kinds of lymphoma cancer can be differentiated by their specific gene expression. The exact diagnosis of the type of lymphoma is essential for a successful treatment. A subset of genes, the so called signatures genes, was able to classify the samples of normal and malignant lymphocytes based on cell types. By use of the signatures genes the diffuse large B-cell lymphoma samples are reclassified into two distinct clusters with two different clinical outcomes. Patients whose cells appeared to be germinal centre like had a much higher survival rate than those with activated B-cell like lymphoma [39]. In another study, it was shown that (among other procedures) by gene expression profiling the tumor-infiltrating immune cells in large cohorts of colorectal cancer can be characterized, which helps for a better prediction of patient survival. The obtained data support the hypothesis that the adaptive immune response influences the behaviour of human tumors [40].

Improve the quality and utilization of medications

Microarrays are used to better understand how medications work and how their effectiveness can be improved. The mechanism of a drug can be more completely understood by studying how it affects the genome *in vivo* [41,42]. Additionally, the genomic response to a particular drug can reveal why some compounds produce unwanted side effects. To predict the success of chemotherapy agents two data sets are combined [43]. The first data set results from microarray analysis of different tumor cell lines. The second is a drug activity profile that measured the effectiveness of chemotherapy medications on each of the tumour cell lines. The combination of both data sets (the so called clustered image map) enables it to predict which medication might be the best for the cancer of an individual person or why some unwanted site effects occur.

Conclusion

Microarrays are powerful tools to analyze genomes *in vivo*. They allow the analysis of the function of genes and their products at a molecular level. Changes in gene expression, as a response to changing environment conditions, can be detected allowing insights into the dynamic of the genome. New genes can be detected and the biological annotation to the genes can be refined. Substantial progress has been made toward the use of microarray to improve the treatment of cancer. Development and understanding of medications are enhanced when microarrays are incorporated into the process. Establishing robust paths from genomic expression data to improved health care is one of the challenges in the future.

References

[1] Mount DW. Bioinformatics: Sequence and genome analysis. Cold Spring Harbor, New York: Cold Spring Harbor laboratory Press; 2001.

[2] Lesk AM. Introduction to bioinformatics. Oxford: Oxford University Press; 2002.

[3] Campbell AM, Heyer LJ. Discovering genomics, proteomics and bioinformatics. San Francisco: Benjamin Cummings; 2002.

[4] Gibas C, Jambeck P. Developing bioinformatics computer skills. Sebastopol: O'Reilly; 2001.

[5] Schena M, Shalon D, Davis RW, Brown PO. Quantitave monitoring of gene expression patterns with a complementary DNA microarray. Science 20;1995; 270(5235):467–470.

[6] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 24 1997;278(5338):680–686.

[7] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 15 2000;405(6788):827–836.

[8] Young RA. Biomedical discovery with DNA arrays. Cell 2000; 102(1):9–15.

[9] Zhang MQ. Large-scale gene expression data analysis: A new challenge to computational biologists. Genome Res 1999; 9(8):681–688.

[10] Schimek MG, Wiltgen M. Microarray gene expression analysis basics for biometricians. Colloquium Biometryczne 2004; 34a:141–159.

[11] Sherlock G, Hernandez-Boussard T, et al. The Stanford Microarray database. Nucleic Acids Res 2001;29: 152–155.

[12] Tseng GC, Oh M, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray filtering: Quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001;29:2549–2557.

[13] Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002;32:490–495.

[14] Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002;3:579–588.

[15] Kooperberg C, Fazzio TG, Delrow JJ, Tsukiyama T. Background correction for spotted DNA microarrays. J Comput Biol 2002;9:55–66.

[16] Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002;32:496–501.

[17] Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology and drug development: Progress and potential. Biochem Pharmacol 2001;62:1311–1336.

[18] Claverie JM. Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 1999;8(10):1821–1832.

[19] Eisen MB, Paul TS, et al. Cluster analysis and display of genome-wide expression patterns. PNAS 1998;95: 14863–14868.

[20] Sturn A, Quackenbush J, Trajanoski Z. Genesis: Cluster analysis of microarray data. Bioinformatics 2002;18: 207–208.

[21] Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Jr., Boguski MS, Lashkari D, Shalon D, Brown PO. The transcriptional program in response to human fibroblasts to serum. Science 1999;1; 283:83–87.

[22] Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R. Large-scale temporal gene expression of central nervous system development. Proc Natl acad Sci USA 1998;6;95(1):334–339.

[23] Kohonen T, Hynninen J, Kangas J, Jaaksonen J. SOM_PAK. The Self-Organizing Map Program Package. Manual Version 3.1. Helsinki University of Technology. Laboratory of Computer and Information Science 1995.

[24] Tamayo P, Slonim D, Merisov J, Zhu Q, Kitareewan S, Dmitrovsky E, Landers ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;16;96(6):2907–2912.

[25] Brown MPS, Grundy WN, Lin D, Christianini N, Sugnet C, Ares MJr., Haussler D. Support vector machine classification of microarray gene expression data. UCSC-CRL-99-9. Santa Cruz: Department of Computer Science, University of California; 1999.

[26] Brown MPS, Grundy WN, Lin D, Christianini N, Sugnet C, Furey TS, Ares M, Jr., Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;4;97(1): 262–267.

[27] Gaasterland T, Bekiranov S. Making the most of microarray data. Nat Genet 2000;24(3):204–206.

[28] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.

[29] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modelling. Electrophoresis 1997;18:2714–2723.

[30] D'Haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modelling of mRNA expression levels during CNS development and injury. In: Altman R, editor. Proceedings of the Pacific Symposium on Biocomputing. Singapore: World Scientific; 1999. p 41–52.

[31] Somogyi R, Fuhrman S, Wen X. Genetic network inference in computational models and applications to large-scale gene expression data. In: Bower JM, Bolouri H, editors. Computational modelling of genetic and biochemical networks. Cambridge, MA: MIT Press; 2001. p 119–157.

[32] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyse expression data. J Comp Biol 2000;7: 601–620.

[33] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 96(12):6745–6750.

[34] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeck M, Merisov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 15 1999;286(5439): 531–537.

[35] Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999;3; 96(16):9212–9217.

[36] Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet 2000;1:48–56.

[37] Cooper CS. Applications of microarray technology in breast cancer research. Breast Cancer Res 2001;3: 158–175.

[38] Spear B, Heath-Chiozzi M, Huff J. Clinical application of pharmacogenetics. Trends Mol Med 2001;7:201–204.

[39] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JL, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. Distinct types of diffuse large B-cell lymhona identified by gene expression profiling. Nature 2000;3;403(6769): 503–511.

[40] Galon J, Ccostes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C, Tosolini M, Camus M, Berger A, Wind P, Zizindohoue F, Bruneval P, Cugnenc PH, Trajanoski Z, Fridman WH, Pages F. Type, density and location of immune

cells within human colorectal tumors predict clinical outcome. Science 2006;313:1960–1964.

[41] Pagliarulo V, Datar RH. Role of genetic and expression profiling in pharmacogenomics: The changing face of patient management. Curr Issues Mol Biol 2002;4: 101–110.

[42] Chicurel ME, Dalma-Weiszhaus DD. Microarrays in pharmacogenomics—advantages and future promise. Pharmacogenomics 2002;3:589–601.

[43] Scherf U, Ross D, et al. A gene expression database for molecular pharmacology of cancer. Nature Genetics 2000;24: 236–244.