

# Gene Expression

Viann Chan      Nick Hontzeas      Vincent Park

December 2000

## 1 Introduction

Gene expression analysis is of great importance in the biological sciences. It provides insight into many functions of a cell since changes in the physiology of an organism are accompanied by changes in the pattern of gene expression. Thus one can deduce the functional and toxicological characteristics of a compound based on the effects this compound will have on gene expression. Many methods have been developed for detecting differences in gene expression levels and identifying the differentially expressed genes. Some of these methods include representational difference analysis (RDA), suppression subtractive hybridization, serial analysis of gene expression (SAGE), differential display PCR (DD-PCR) and cDNA microarrays. This review section will briefly cover DD-PCR (one of the most popular gene expression techniques) and the now emerging cDNA microarray technology.

## 2 DD-PCR

Through the arbitrary amplification and comparison of different mRNA sources, DD-PCR allows identification of differentially expressed genes in various in vitro and in vivo systems. This method does not depend on availability of cDNA clones or prior knowledge of sequence yet it lacks the qualitative component of Northern blots. Many genes have been identified using this technique including genes stimulated by ethylene in *Arabidopsis*. However DD-PCR has many limitations. It generates short expressed sequence tags (ESTs) (100 – 400bp), which represent mainly the 3' UTR of mRNAs. Identification of functional coding sequences requires the screening of cDNA li-

braries and the sequencing of full-length cDNAs. In addition only a limited number of different conditions can be compared and screening is based on differences in mRNA length and not identity [18, Liang and Pardee 1992].

### 3 DNA microarray

In order to study gene expression in DNA microarray technology, mRNA is hybridized to a high-density array of immobilized target sequences, each corresponding to a specific gene. The mRNAs being sampled are labeled with fluorescent dyes and are then hybridized to the array where each mRNA will quantitatively hybridize to its complementary target sequence. The expression level of a particular gene can be visualized by observing the fluorescence at each spot of the array. If two, or more, differently labeled mRNAs are used then one can quantitatively compare gene expression in both samples [12, Harrington et al. 2000].

Two main methods exist to manufacture DNA microarrays also referred to as DNA chips. In the first method oligonucleotides are directly synthesized on a solid surface. A photomask is used to illuminate specific areas of a glass surface that is made by incorporating linker molecules carrying a photo-labile protective group. The surface is then incubated with a solution containing a photoprotected nucleotide that only associates with the light-activated areas. A second photomask is used to deprotect other areas on the surface and then another type of nucleotide is attached to these areas. Thus a defined set of oligonucleotides is synthesized on the surface. By using this method DNA chips with very high densities ( $> 250000$  oligo *spots/cm*<sup>2</sup>) can be made and also large number of identical arrays. However this type of a DNA chip is extremely expensive and has no flexibility in design [12, Harrington et al. 2000].

The second method referred to as DNA micro-dispensing. The advantage it has over the previous method is that it can be done in a regular lab, providing more flexibility. Basically small quantities of DNA solution are put into a solid surface. This is done by the use of micro-dispensing robots. There are of two kinds of robots:

**passive dispensers** apply DNA with a pin that touches the solid surface

**ink-jet devices** that do not touch the surface

The density of the chip depends on the capability of the dispensing device. This method allows the constant update of the array and DNA can be deposited in a desired format. The DNA can be put into different kinds of surface on which it binds. When glass is used, a positively charged layer is used to bind negatively charged DNA. On the other hand DNA that have 5' modifications can be covalently bound to a glass surface that carries reactive groups. There are other materials being considered for DNA attachment that aim at better signal to noise ratios, increase in binding capabilities as well as reproducibility and reusability [12, Harrington et al. 2000].

Each microarray is designed based on the particular research question that one wants to answer. As discussed earlier DNA chips have the advantage that a large set of genes can be examined in parallel. If all the genes of an organism are known sequences corresponding to all the open reading frames (ORFs) can be put on the array thus allowing for the simultaneous expression and analysis of all mRNAs. For example in *Arabidopsis* the complete genome has been (99.9%) sequenced. Thus some groups have already started looking at gene expression differences in *Arabidopsis* when treated with different hormones such as ethylene, jasmonic acid, abscisic acid and even with a pathogen [25, Schenk et al. 2000]. They simultaneously looked at the expression profiles of 2375 genes when treated with these hormones/pathogen. Changes in mRNA expression considered significant were only those that data analysis showed to be in excess of 2.5-fold.

Analyses of the data revealed that 705 ESTs on the microarray showed significant differential expression in response to the various treatments. However it was not clear how reliable some of the data was since the expression pattern of two defense genes of *Arabidopsis*, PDF1.1 and PDF2.1 were shown to be upregulated in the microarray after ethylene treatment. Other groups, however, by using DD-PCR have shown that ethylene upregulates PDF1.2 but has no effect on PDF1.1. Thus this particular microarray result is false. This was attributed to the fact that microarrays are very poor in differentiating between closely related sequences. There are numerous other examples where microarrays have been used. Examples include the profiling of all the open reading frames of *Saccharomyces cerevisiae* [13, Hauser et al. 1998], the ex-

amination of gene expression in human epithelial cells when challenged with the pathogen *Pseudomonas aeruginosa* [15, Ichikawa et al. 2000], and even just monitoring gene expression of *E. coli* grown in different media [27, Tao et al. 1999].

In order to visualize the expression levels of mRNAs in microarrays, the mRNAs are fluorescently labelled prior to hybridization. Some of the most common dyes used are *Cy3* and *Cy5* as well as fluorescein and rhodamine. The fluorophore should have a narrow excitation and emission peak, a high level of photon-emission and resistance to photobleaching. *Cy3* is less sensitive to photobleaching than *Cy5* and gives less background fluorescence on a glass surface than *Cy5* [22, Nelson et al. 2000]. Currently companies are developing dyes with better characteristics to use with better scanners. For example Umedik Inc. as US company and BioMedical Photometrics Inc. based in Waterloo Canada are collaborating to develop fluorescence based microarray readers based on confocal laser optical techniques. It is argued that by the addition of rapid laser reading this will make DNA-chips cheaper and more time effective.

Other companies such as Nanogen Inc. are developing chips that integrate microelectronics and molecular biology through the use of semiconductor microchips. These so called nanochips are said to offer high accuracy, rapid reaction times, simultaneous tests on a single sample, and a wide applicability to a variety of charged molecules. Nanogen in conjunction with Egea Inc. have combined software, robotics and new biochemical methodology to create long, computer-designed DNA molecules over 10000 bp. Similarly Clinical Micro Sensors has used organic molecules to form electronic circuits, a process known as bioelectronics. This chip called the eSensor is a small circuit board containing electrodes attached to small fragments of DNA or RNA. It can detect any form of DNA or RNA based on complementary base pair recognition. After bonding takes place a signal is generated at the electrode and translated through signal processing technology to identify and quantify each target sequence. The company claims that this biochip can be used to detect viruses or bacteria or SNPs and that it can simultaneously test multiple DNA or RNA fragments, thus making it useful for a broad range of genomic techniques including genetic counseling, agriculture and food safety [6, Dutton 2000].

Gene chips are especially attractive to pharmaceutical companies since they can provide a very powerful tool for the development of drugs specific to individual genetic variation. This emerging field is referred to as pharmacogenomics. The main question that pharmacogenomics attempts to answer is that why do different patients with the exact same disease symptoms respond differently to the same drug. In order to answer this question pharmacogenomics attempts to combine the fields of functional genomics with that of molecular pharmacology with the use of DNA microarrays and specific DNA chips. For example DNA-chips can be used for the detection of mutations in specific genes as diagnostic "markers" of the onset of a particular disease. The Affymetrix Genechip is a good example of this. They produce a chip called the *p53* geneChip that is designed to detect SNPs of the *p53* proto-oncogene. In addition other gene chips have been designed to detect differences in gene expression levels in cells that are diseased versus those that are healthy. Another interesting application could be in the field of population genetics. Currently HiberGen, Irelands first genomics company will focus probing the genetic homogeneity of the Irish population both in the discovery of new disease related genes and for pharmacogenomic programs that aim to identify gene variants related to drug response [23, Persidis 2000].

One can deduce from all the above that a great challenge for the industry is the standardization of DNA-chips. The assays and the equipment should eventually be standardized so that the data could be easily integrated into existing equipment. Recently standards for storing and exchanging genomics and proteomics are being developed by the BioPathways Consortium, which recently held its first meeting at the intelligent systems for molecular biology meeting in San Diego. They are attempting to standardize the information and the format in which its stored so that researchers have a realistic hope of selecting the right search terms the first time, searching multiple databases simultaneously and using the information as easily as they access and use other types of information from the web. The BioPathways consortium currently is surveying the industry to determine exactly what information is available and how it is stored.

Also evident is the continuous incorporation of different technologies into the development of DNA-chips. For example flame hydrolysis deposition (FHD) of glasses is widely used in the telecommunications industry and is now being applied for the development of new chips. Moreover in another application

Luminex Corp. is collaborating with Invitrogen in a three-year project to develop a virtual array technology capable of performing 10000 to 100000 assays simultaneously in a single tube and analyzing them simultaneously. This technology will expand on existing technology by Luminex, which employs colour-coded microspheres, lasers and digital signal processors to generate real time, quantitated data for multiple distinct tests that can be performed simultaneously in a single tube. This highlights a current industry trend toward "information-density" screening in combination with high-throughput screening [17, Kerr 2000].

The array technology is also being used to develop protein arrays. Until recently it was not possible to analyze proteins using high-density array and automated approach as described for the DNA arrays. This problem has now been addressed. Ligand-coated surfaces are now available to capture and analyze specifically labeled proteins. Largely based on DNA microarray chip technologies protein arrays have been recently refined. These developments include the generation of low-density protein arrays on filter membranes, or the use of either photolithography of silane or gold monolayers, combining microwells with microsphere sensors, or ink-jetting onto polystyrene film. In protein chip technology surface plasmon resonance has contributed greatly to interaction studies in a chip format. The SELDI ProteinChip technology combines affinity capture and high resolution mass analysis with novel and efficient means for the characterization of proteins and their interaction partners and has applications such as epitope mapping and receptor-ligand or protein-drug interaction studies [4, Cahill 2000].

## 4 Data Analysis

Microarrays typically produce massive data points. For example a small scale experiment with five samples and two replicates will produce approximately 100000 data points. Thus one can see that proper computer strategies are needed to interpret and manage this plethora of data. The three steps involved in data interpretation are: data normalization, data filtering and pattern identification. In order to effectively compare expression levels the data must first be normalized. Subsequently the data is reduced by not considering genes expressed below a defined threshold, as for example in the *Arabidopsis* experiment earlier where no genes below the threshold of 2.5

were considered. Finally biological function is assigned to expression profiles by finding patterns in the data. Methods used for the interpretation of the data range from lists of decreased to increased gene expression based on user defined threshold to clustering and visualization programs such as hierarchical clustering and k-means clustering.

Data mining strategies are also used. They are divided into two categories: differential gene expression and coordinated gene expression. In differential gene expression pair-wise comparison data are looked at such as for example "non-ethylene treated plants to ethylene treated plants". Coordinated gene expression analysis is done by assessing the expression levels of a large number of genes over a period of time or through a series of experimental conditions. The data are then normalized across samples distinguishing biological change from random noise [12, Harrington et al. 2000].

## 4.1 Supervised and Unsupervised Methods

The primary difference between supervised and unsupervised methods is that unsupervised methods try to analyze data without a teacher signal; that is, these methods have no prior knowledge of true functional classes and typically use similarity or distance measures to distinguish between groups. Supervised methods specify which data should cluster together through some sort of training [1, Brazma 2000].

An example of an unsupervised method is hierarchical clustering. This method starts with each vector being its own cluster, and iteratively joins the two closest clusters together. The distances are then recalculated (distance measures may be specific to the given problem) [1, Brazma 2000].

An example of a supervised method is the Fisher's Linear Discriminant (FLD). The goal of this method is to find a direction, such that when the data is projected along this direction, the distinction between two functional groups are maximized [<http://www.data4s.com/biolk.html>, leukemia case study].

## 5 Unsupervised Methods

Increasing number of the statistical methodologies have been applied in finding the functional genomic clusters in the gene expression data. They can be roughly divided into three categories: simple criteria matching, those that use Euclidean distance, and comprehensive pair-wise comparisons.

### 5.1 Simple Criteria Matching Methods

The first category is the simplest way to use the gene expression data. After the treatment, each gene expression level is measured and is sorted according to the fold-difference. Genes that demonstrate a fold-change greater than a given threshold are the considered “clustered” with the intervention [3, Butte and Kohane 2000]. This is simple to use but is not complex enough to retrieve the correlations among gene expressions.

### 5.2 Distance and Similarity Measures Methods

#### 5.2.1 *K*-means

The *K*-means clustering algorithm was introduced by J.B. MacQueen in 1967 [21, MacQueen 1967]. It is an unsupervised method which partitions the data set into *K* clusters. The number of clusters (*K*) and the number of iterations must be specified before running the algorithm. Clusters are formed by grouping data points which are close together or “similar”. Euclidean distance is commonly used as the distance measure. Each gene is represented as a point. The vector for each point comprises of the gene expression level at each measured time interval. At each iteration, the mean vector for each cluster is computed and points are reassigned to the cluster with the closest mean vector. Iterations are repeated until either the clusters converge (the mean vector for each cluster does not change), or the maximum number of iterations has been reached.

Let  $X$  be the total set of  $N$  vectors (or points) in an  $n$ -dimensional vector space. Let  $X = S = S_1 \cup S_2 \cup \dots \cup S_k : S_i \neq \phi$ , and  $S_i \cap S_j = \phi$  for  $1 \leq i, j \leq K$ , where  $i \neq j$ . Let  $N_i$  be the number of vectors in cluster  $S_i$  such that  $\sum_{i=1}^K N_i = N$ .

Let  $x, y \in X$ , define the distance measure between  $x$  and  $y$  to be the Euclidean Distance on the given vector space:

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (1)$$

Let  $m_i$  be the mean of cluster  $S_i$ .

$$m_i = \left(\frac{1}{N_i}\right) \left(\sum_{x_k \in S_i} x_k\right) \quad (2)$$

Algorithm

Initialization:

Arbitrarily form  $K$  clusters and run for  $c$  iterations.

Iteration:

Calculate the mean  $m_i$  for each cluster  $S_i$ .

For each vector,  $x_k \in S_j$  calculate  $d(x_k, m_i)$  for each  $S_i$ .

Reassign the vector  $x_k$  to the cluster with the closest mean.

ie.  $S'_j = S_j \setminus \{x_k\}$  and  $S'_i = S_i \cup \{x_k\}$  such that  $d(x_k, S_i) = \min_{l=1}^K (d(x_k, S_l))$

This algorithm runs in  $O(N * n * K * c)$ . It was applied expression profiles collected from 3000 yeast genes at 15 time intervals during the cell cycle into 30 clusters [28, Tavazoie et al. 1999]. Half of the clusters had significant patterns.

The disadvantage of using this method is that there are  $K^{\frac{N}{K}}$  ways of initializing the  $K$  clusters. Choosing an optimal cluster is  $NP$ -complete [16, Johnson 1982]. Another problem with this method is that for certain experiments, the number of clusters expected is unknown. Also, there is no clear way of choosing the exact number of iterations the algorithm should be run.

### 5.2.2 Self-Organizing Maps

Self-organizing maps (SOM) are a variation on the  $K$ -means method. The genes are represented as the points in the multi-dimensional space. Coordinates of the points are the expression levels at various time points. A grid

of centroids is allowed to drift towards the collections of points in the multi-dimensional space. The resulting centroids reflect that the clusters of related genes have a relatively smaller Euclidean distance in the multi-dimensional space [3, Butte and Kohane]. At the end, the phylogenetic-type tree are constructed; the branch length of the tree is proportional to the Euclidean distance between genes, and the coordinates of nodes represent the expression levels at various time points.

### 5.3 Comprehensive Pair-wise Comparison Methods

The last methodology is to comprehensively compare all genes against each other using a metric. It also generates the phylogenetic-type trees with branch lengths proportional to the correlation coefficients between vectors in coordinates representing the expression levels at various time points, but it uses vectors for each genes. There have been also many attempts made to find the new and better methodology which is deviated from the above categories.

#### 5.3.1 Coupled Two-Way Clustering

The Coupled Two-Way Clustering analysis(CTWC) is an example of a method which uses pair-wise comparisons. The CTWC is to identify the relevant subsets of the data and to find the correlations among them that were masked and hidden when full dataset was used. This method assumes that as the clusters are broken down into the smaller subset, the more informations among genes are available; however, this is not always the case. The iterative clustering process is used to find the pairs of genes and samples that produce the stable and significant partitions. The CTWC can be implement with any existing clustering algorithm.

Usually, the Superparamagnetic clustering algorithm(SPC) is used because it is robust against the noise and capable of controlling the resolution of the performed clusters at different given temperatures. The test data are organized in an expression level matrix  $A$  where a row represents a single gene and each column corresponds to a particular sample. The pairs of subsets of genes and samples are found by taking all possible submatrices of the original data and applying the standard (uncoupled) two-way clustering procedure to every one of them. Unfortunately this is impossible to implement since the

number of the submatrices grows exponentially with the size of the problem. In practice, the only submatrices for the stable clusters found by other clustering algorithm such as SPC are used to reduce the complexity. Then the resulting pairs are used to identify genes that partition the samples according to known classification, to discover new partitions, to sensitively identify subpartitions, and to reveal conditional correlations among genes [10, Getz et al. 2000].

### 5.3.2 Mutual Information Relevance Networks

There is also a technique that computes comprehensive pair-wise mutual information for all genes in the microarray [3, Butte and Kohane 2000]. The Mutual Information(MI) means the degree of one gene non-randomly associating with another. Consequently the high MI implies that the two are closely related biologically. The MI is calculated using the entropy of genes. All genes are compared against each other to give the complete MI for every possible pairs of genes. Once all the MIs are know, the threshold MI is set so that all the pairs of genes above the threshold are linked together to construct the Relevance Networks(RN). RN is the graph representation to illustrate the functional closeness among genes instead of using a phylogenetic-type tree strucutre. Despite the complexity of comparing each pair increases exponential with the size of the problem, the resulting MIs are complete.

## 6 Supervised Methods

### 6.1 Support Vector Machines

The Support Vector Machines (SVMs) is a supervised method that uses training data to identify groups with similar gene function and outliers. SVMs were invented by Vladimir Vapnik and his co-workers and introduced in his paper at the Annual Conference on Computational Learning Theory in 1992. (Many of the theoretical components had already been developed, but had not been put together.)

This method is used to determine the characteristic of a functional class. A set of genes in a functional class are labelled postively (ie. +1). A set of genes not a the functional class are labelled negatively (ie. -1). These two sets are combined to form the training set.

The SVM uses the training set to learn to distinguish between members and non-members of the functional class based on expression data. Once the SVM has been trained, it can be applied on new genes or re-applied to the training data (to search for misclassified genes).

SVMs were used to analyze 2467 genes in the yeast *Saccharomyces cerevisiae* and trained to recognize 5 functional classes from 79 different experiments. Among the learning techniques applied (SVM, LFD, Parzen Windows, and two variations on decision trees), the SVM method outperformed all the other methods that were analyzed. [2, Brown 2000].

Let  $X_i$  be the log ratio of the expression level  $E_i$  for gene  $X$  in experiment  $i$  to the expression level  $R_i$  of gene  $X$  in the reference state, normalized so that  $\vec{X} = (X_1, \dots, X_{79})$  has Euclidean length 1:

$$X_i = \frac{\log(E_i/R_i)}{\sqrt{\sum_{j=1}^{79} \log^2(E_j/R_j)}} \quad (3)$$

Let output  $y = +1, -1$  be labels for  $X_i$ .  $X_i$  is positive is induced by the reference state, and is negative if it is not induced by the reference state. Each  $\vec{X}$  represents a point in an m-dimensional space. The goal is to form a hyperplane in the hyperspace to separate the training data into the two groups. However, when the data is nonseparable, it may be necessary to map the data to a higher dimensional space called a feature space. This can be computationally intensive and may result in overfitting of data.

SVMs avoid overfitting by choosing a maximum margin hyperplane. This is a hyperplane which maximizes the minimum distance from the hyperplane to the closest training point. The training point is known as a support vector. By defining a kernel function that acts as the dot product in the feature space, this space does not need to be explicitly represented and the vectors in the feature space do not need to be computed. The kernel function must be defined, continuous, and positive. [29, Vapnik 1998].

The maximum margin hyperplane can be represented as a linear combination of training points (support vectors) that are closest to the hyperplane [20, Lu and Yang 1999]. The location of the hyperplane in the feature space depends

on weights on the training data. Points which are far from the hyperplane have zero weights, points (support vectors) close to the boundary between the two functional groups have non-zero weights. Removing support vectors will change the location of the hyperplane.

If there are misclassified data points (ie. points which are wrongly identified to be in the functional group), the SVM may not be able to find a hyperplane to separate the data. In order to overcome this, a soft margin can be used. A soft margin can be obtained by either adding some constant factor to the kernel function output (when input vectors are identical), or to set an upper bound on the training weights.

## 7 Future Perspective

Gene expression is a growing interdisciplinary field. It overlaps in many areas of science, medicine, computer science - software and hardware, statistics, engineering, etc. New applications of gene expression are identified daily.

Data analysis will become increasingly important as more information is collected. This stockpiling of information has already presented many challenges in areas of data-mining and data-retrieval, demands on hardware and software, the need for standardization of methods, accessibility of supercomputers (for small independent labs), and analysis of large datasets.

The methods discussed in this paper only survey a few of the algorithms used in practice. More research into data analysis is needed to better methods in dealing with large datasets and to find new methods to interpret data in more biologically intuitive ways. As the Human Genome project completes, microarrays will be used on a large scale to search and identify genes involved in complex diseases. This will be looking for a needle in a haystack (where the haystack is growing exponentially).

## 8 Online Resources

### 8.1 Gene Expression

<http://www.oci.utoronto.ca/services/microarray> - Ontario Cancer Institute

<http://cmgm.stanford.edu/pbrown/mguide> - How to build a “homemade” system

<http://industry.ebi.ac.uk/~alan/MicroArray/> - Gene expression links

<http://www.nature.com/ng/microarray/abstracts.html> - The Microarray Meeting: Technology, Application and Analysis (Nature Genetics)

## References

- [1] Brazma A., Vilo J., (2000) “Gene expression data analysis.” *Federation of European Biochemical Societies: Letters*. 480(1), pp. 17-24.
- [2] Brown M.P.S., et al., (2000) “Knowledge-based analysis of microarray gene expression data by using support vector machines.” *Proceedings of the National Academy of Sciences USA*. 97(1), pp. 262-267.
- [3] Butte A.J., Kohane I.S., (2000) “Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements.” *Pacific Symposium Biocomputing*. pp. 418-429.
- [4] Cahill D.J., (2000) “Protein arrays: A high-throughput solution for proteomics research?” *Proteomics: A Trends Guide*. pp.47-51.
- [5] Dangond F., (2000) “Chips around the world: Proceedings from the *Nature Genetics* Microarray Meeting.” *Physiological Genomics*. 2(2), pp. 53-58.
- [6] Dutton G., (2000) “Bioinformatics standards.” *Genetic Engineering News.*, 20(17).
- [7] Eisen M.B., et al. (1998) “Cluster analysis and display of genome-wide expression patterns.” *Proceedings of the National Academy of Sciences USA*. 95(25), pp. 14863-14868.
- [8] Gerstein M., Jansen R., (2000) “The current excitement in bioinformatics-analysis of whole-genome expression data: How does it

relate to protein structure and function.?" *Current Opinion in Structural Biology*. 10(5), pp. 574-584.

- [9] Getz G., et al., (2000) "Super-paramagnetic clustering of yeast gene expression profiles." *Physica A*. 279, 457-464.
- [10] Getz G., Levin E., Domany E., (2000) "Coupled two-way clustering analysis of gene microarray data." *Proceedings of the National Academy of Sciences USA*. 97(22), pp. 12079-12084.
- [11] Ghosh D., (2000) "High throughput and global approaches to gene expression." *Combinatorial Chemistry & High Throughput Screening*. 3(5), pp. 411-20.
- [12] Harrington C.A., et al., (2000) "Monitoring gene expression using DNA microarrays." *Current Opinion in Microbiology*. 3, pp. 285-291.
- [13] Hauser N.C., et al., (1998) "Transcriptional profiling on all open reading frames of *Saccharomyces cerevisiae*." *Yeast*. 14, pp. 1209-1221.
- [14] Heyer L.J., et al., (1999) "Exploring expression data: Identification and analysis of coexpressed genes." *Genome Research*. 9(11), pp. 1106-1115.
- [15] Ichikawa J.K., et al., (2000) "Interaction of *Pseudomonas aeruginosa* with epithelial cells: Identification of differentially regulated genes by expression microarray analysis of human cDNAs." *Proceedings of the National Academy of Science USA*. 97(17), pp. 9659-9664.
- [16] Johnson D.S., (1982) "The NP-Completeness Column: An ongoing guide." *Journal of Algorithms* 3. pp. 182-195.
- [17] Kerr E. (2000) "New tools launched at SBS." *Genetic Engineering News*. 20(17).
- [18] Liang P., Pardee A.B. (1992) "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction." *Science*. 257, pp. 967-971.
- [19] Lockhart D.J., Winzeler E.A., (2000) "Genomics, gene expression and DNA arrays." *Nature*. 405, pp. 827-835.

- [20] Lu G., Yang P.K., (1999) "Gene Classification with Support Vector Machines." <http://www.cse.ucsc.edu/~pyang/pyang/biofinal.html>
- [21] MacQueen J.B., (1967) "Some Methods for Classification and Analysis of Multivariate Observation." *Proceedings 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability*. 1, pp. 281-297.
- [22] Nelson R.W., et al., (2000) "Biosensor chip mass spectroscopy: A chip-based proteomics approach." *Electrophoresis*. 21, pp. 1155-1163.
- [23] Persidis A. (2000) "Pharmacogenomics." *Nature Biotechnology*. 18(supp), pp. 40-42. (reprinted from (1998) "The business of pharmacogenomics." *Nature Biotechnology*. 16(2), pp. 209-210.)
- [24] Rifkin R., Pontil M., Verri A. (1999) "A Note on Support Vector Machine Degeneracy." *Lecture Notes in Artificial Intelligence 1720: Proceedings, 10th International Conference on Algorithmic Learning Theory*. pp. 252-263.
- [25] Schenk P.M., et al. (2000) "Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis." *Proceedings of the National Academy of Science USA*. 97(21), pp. 11655-11660.
- [26] Tamayo P., et al., (1999) "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." *Proceedings of the National Academy of Sciences USA*. 96(6), pp. 2907-2912.
- [27] Tao H., et al., (1999) "Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media." *Journal of Bacteriology*. 181(20), pp. 6425-6440.
- [28] Tavazoie S., et al., (1999) "Systematic determination of genetic network architecture." *Nature Genetics*. 22(3), pp. 281-285.
- [29] Vapnik V., (1998) *Statistical Learning Theory*. Wiley.
- [30] Woolf P.J., Wang Y., (2000) "A fuzzy logic approach to analyzing gene expression data." *Physiological Genomics*. 3(1), pp. 9-15.