

# Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction

Javier Tamames<sup>a</sup>, Dominic Clark<sup>a</sup>, Javier Herrero<sup>b</sup>, Joaquín Dopazo<sup>b</sup>,  
Christian Blaschke<sup>c</sup>, José M. Fernández<sup>c</sup>, Juan C. Oliveros<sup>c</sup>,  
Alfonso Valencia<sup>c,\*</sup>

<sup>a</sup> *ALMA Bioinformatics, S.L., Spain*

<sup>b</sup> *Centro Nacional de Investigaciones Oncológicas (CNIO), Spain*

<sup>c</sup> *Protein Design Group, National Center for Biotechnology (CNB-CSIC), Cantoblanco, Madrid E-28049, Spain*

Received 18 July 2001; received in revised form 21 March 2002; accepted 27 March 2002

## Abstract

Expression arrays facilitate the monitoring of changes in the expression patterns of large collections of genes. The analysis of expression array data has become a computationally-intensive task that requires the development of bioinformatics technology for a number of key stages in the process, such as image analysis, database storage, gene clustering and information extraction. Here, we review the current trends in each of these areas, with particular emphasis on the development of the related technology being carried out within our groups. © 2002 Elsevier Science B.V. All rights reserved.

*Keywords:* Text analysis; Protein function; Expression arrays; Medline

## 1. Introduction

In the past few years, the introduction of expression array technology has facilitated the analysis of the expression levels of large sets of genes. By exploring the transcription profiles of genes, expression array data provide new and very

relevant information on the biology of organisms. The task is currently aided by the availability of commercial arrays that contain accurate representations of the genomes of several organisms, including the human genome (<http://www.affymetrix.com>).

A wide variety of information can be obtained by using expression arrays. In exploring the genomic content of organisms, DNA arrays have been used for: discovering regulatory elements for genes that display common patterns of gene expression (Cohen et al., 2000; Leemans et al., 2001); protein function prediction for related genes (Brown et al., 2000; Cummings and Relman,

*Abbreviations:* GEISHA, gene expression information system for human analysis; SOTA, self-organising tree algorithm.

\* Corresponding author. Tel.: +34-91-585-4570; fax: +34-91-585-4506

*E-mail address:* [valencia@cnb.uam.es](mailto:valencia@cnb.uam.es) (A. Valencia).

2000); genotyping (Behr et al., 1999); and genomic analysis of non-sequenced organisms (Hayward et al., 2000; Akman and Aksoy, 2001).

DNA arrays are also very valuable tools for studying the organisation of biological processes. Genetic networks, metabolic pathways and the temporal development of biological processes have all been addressed for a diverse set of organisms, including humans (deRisi et al., 1997; Cho et al., 1998; Ferea and Brown, 1999; Iyer et al., 1999; Richmond et al., 1999; Tavazoie et al., 1999; Ideker et al., 2001).

On the clinical side, drug discovery is a field that particularly benefits from the use of DNA array technology (Debouck and Goodfellow, 1999). It has been successfully applied to drug target identification (Kozian and Kirschbaum, 1999) and drug development (Gray et al., 1998) and validation (Marton et al., 1998a,b; Wilson et al., 1999). Pharmacogenomics is now expected to develop principally via the use of DNA array information (Evans and Relling, 1999; Scherf et al., 2000). Diagnostic research is also expected to benefit, with preliminary and very promising results having been produced for the diagnosis of some types of cancer (Golub et al., 1999; Scherf et al., 2000). Pathogenicity mechanisms and disease progression can also be followed readily (Cumings and Relman, 2000; Geiss et al., 2000).

DNA array technology is therefore a powerful tool. Nevertheless, the analysis of DNA array data relies heavily on the availability of computational methods for:

- Array design. In most experimental scenarios, researchers are interested in creating their own arrays. In general, this will involve using sets of genes with known features: genes of known function; genes that exhibit particular expression patterns discovered in previous experiments; genes that have sequence or functional relationships with genes of interest; genes that are sequence variants of genes of interest, or biological controls.
- Image analysis. Even if most commercial robots for scanning arrays include their own image analysis software, alternative commercial software is also available for this, as well as a multitude of open source packages. Quantification of the data is a key step on which the overall analysis is wholly dependent. It is also directly related to the peculiarities of fabrication of the arrays, a problem that is even more pronounced in the case of nylon-based arrays (macroarrays).
- Storage and organisation of experimental results. The potential for carrying out thousands of experiments using thousands of genes creates an obvious need for database structures that are able to store the results of these experiments. Only with the availability of well-designed databases will it be possible to carry out complex queries of data that relate to various different experiments.
- Comparison of expression profiles to determine groups exhibiting similar behaviour. The final results of an expression array experiment consist of large collections of expression profiles that give information on the levels of expression of each gene under various different conditions. The structure of this data can be very complex, in line with the various possible experimental conditions being analysed (e.g. different patients with different doses of a drug at different response times). Currently most applications simply consider all conditions as being equal, however this still leaves key questions to be answered, such as how to define the distances between expression patterns and how to form groups of genes with similar expression patterns based on the chosen distance measure.
- Functional interpretation. This involves interpretation of the biological meaning of the various groups of genes that have been produced (usually known as gene clusters). The results of the experiments and the data analysis are groups of genes with similar expression patterns and these groups will include genes for which the level of knowledge on their function is very variable. The key biological questions are then: what is the relationship between these genes? Why do they have similar expression patterns? Is there some real biological meaning behind the clustering? Finding answers to these questions is normally carried out by inspecting in detail the information

available in the literature and databases. This can be a huge task involving a number of scientific and technical challenges and is usually never completely resolved, as alternative explanations are often possible.

In carrying out functional genomics projects, research institutions and private companies have addressed these aspects of expression array analysis via the use of proprietary development of technology and local implementations. In the following sections we review the general availability of solutions for the problem areas mentioned above and provide a focus on our own development of technology in these areas.

## 2. Array design

Correct identification of the critical aspects of DNA array experiments allows implementation of appropriate bioinformatics solutions. One of the first obstacles encountered in analysis of the data regards the design of the array that is to be used. Except for certain special cases (Loftus et al., 1999; Rockett et al., 2001), ‘brute force’ approaches are typically used that consist of putting as many genes as possible onto an array. At best, some form of pre-screening is carried out using various arrays with all possible genes present and then a selection is made from these a posteriori. Whichever solution is chosen in the end, it is clear that pre-selecting genes based on prior knowledge of their function, the tissue in which they are expressed etc., should always improve the effectiveness of the experiment.

In general, this selection process involves identifying genes on the basis of some of their properties: genes of known function or located at a given position on the chromosome, etc. There are various different databases and public data repositories containing information on the sequences that can be used for the selection, such as UniGene (<http://www.ncbi.nlm.nih.gov/unigene>), TIGR gene indices (TIG in <http://www.tigr.org/tdb/tgi.shtml>) and DoTS (<http://www.allgenes.org>), etc. These resources provide high-quality annotation for the cDNA contained therein, but since the

number of sequences is extremely high (more than 13 million sequences in the last GenBank release for 2001). Integrated software that allows user-friendly querying, array project management and clone tracking can therefore be very useful.

Pre-selection of genes is possible using a system based on a list of genes that includes information from various sources, allowing efficient retrieval of genes with the desired characteristics. Such a system would allow the user to manage the huge volume of information in a comprehensive way. A prototype design for a system such as this is summarised in Fig. 1. This consists of a relational database storing information on genes. The database uses UNIGENE (<http://www.ncbi.nlm.nih.gov/unigene/>) as an index, onto which all the information identified as useful for array design (such as chromosomal location, tissue in which each gene is expressed, description of function, etc.) is mapped. Private or clinical information can also be included. Since for many genes there is no explicit information available, a second source of information is based on homology queries with other genes or on searches for functional motifs. Information from other sources may also be reflected, such as experimental information on any incidents occurring during the quality control phases in previous experiments.

Researchers may then formulate queries using keywords to obtain lists of genes that correspond to their particular interest. These lists, called projects, can be split, merged and broken down into those genes for which clones exist in-house and those for which clones do not exist. This provides a new functionality for the mass of information stored in the various different databases, whose combined use is otherwise virtually impossible.

## 3. Image analysis

There are several public and commercial solutions for the problem of analysis of the images produced in DNA array experiments. All are capable of working with the standard TIFF images obtained in fluorescence-based experiments and, in

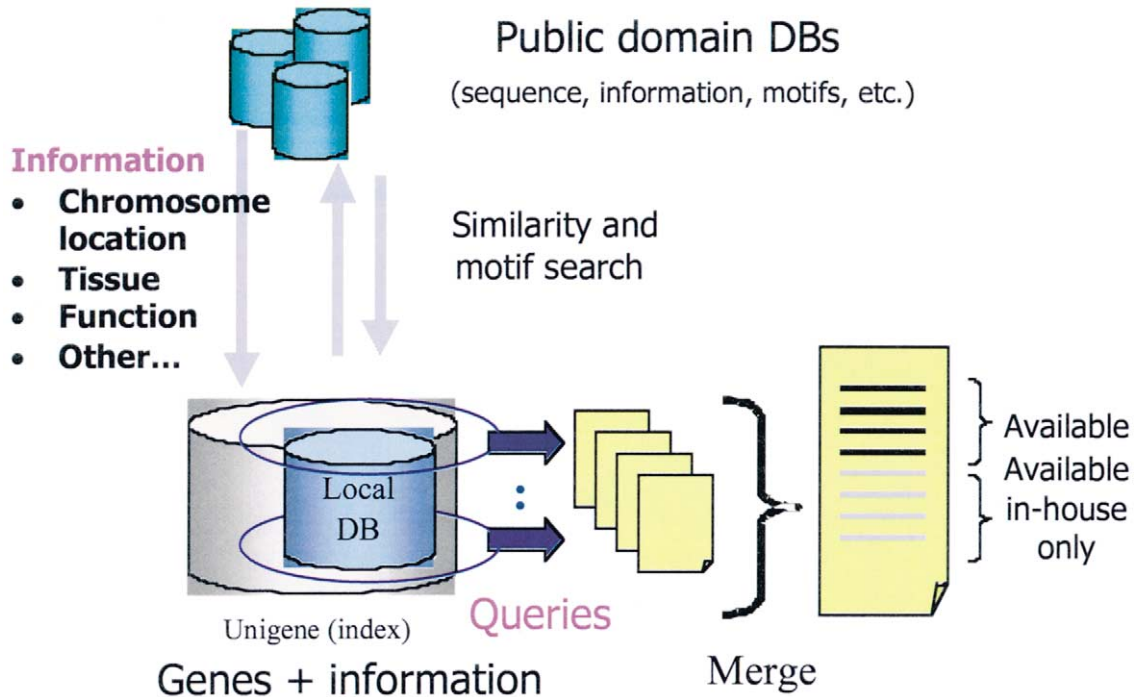


Fig. 1. Schema of the organization of the proposed array design system. Data gathered from various sources are integrated within a relational database in which genes are indexed according to UNIGENE. Queries can be performed on this database allowing the retrieval of interesting genes. Local databases (clone databases) provide information on whether or not these genes are available in a local collection.

some cases, the GEL images from radioactive macroarray experiments.

Most of the systems for image analysis are based on adjusting a grid with the correct number of rows and columns to fit the spots, followed by quantification of the different signals, including a more or less robust measurement of the background associated with each one. The quality of the programs can be measured according to the effectiveness with which each program solves the main steps in the process of image analysis. It is very difficult to be objective here and the best way to choose the optimal package for a given laboratory or facility is to test the available demo versions and check if they are useful or not. Instead of ranking the various different options mentioned, we will instead describe the critical steps in the process of image analysis that should be used to evaluate these programs.

### 3.1. Image capture

A good program should be able to cope with as many standard types of DNA array image as possible. For fluorescent double hybridisations, the format is usually TIFF, with 16 bits of colour deep per channel (a channel corresponds to a fluorochrome). Some scanners are able to use three or more channels. A flexible image analysis program should also be able to capture radioactive images (GEL images, from Phosphorimager-like devices). This kind of image is used in macroarray experiments.

### 3.2. Adjustment of the grid

This is a key step in the process of quantification. The adjustment of the grid should be fast and able to be adapted to several possible patterns of

distribution of spots. In the case of microarrays, the spots are often distributed as discrete sectors of rows and columns that are easy to define, but in some macroarray membranes the distribution is semi-regular, with spots being set in the space between columns or rows.

### 3.3. Precise detection of the spots

After the grid has been defined and positioned, a more precise localization of each spot should be carried out. This is due to the fact that some spotter robots can displace the spots slightly. In the case of macroarrays, this effect is magnified by the hybridisation process and can produce a distortion in the flexible membrane.

### 3.4. Detection and annotation of low quality spots

For all arrays, some spots present an irregular shape or lack detectable intensity because they are not well set or the hybridisation has not been correctly achieved. A good image analysis program should be able to detect these cases with some precision, as they should then be discarded from the analysis.

### 3.5. Background estimation

Estimation of the background ideally should be local, since each spot has a surrounding background that can be of different magnitude to other areas of the image.

### 3.6. Precise quantification of the spots

There are several quantification algorithms. The simplest are based on the mean intensity of the pixels of the spot when considered as a circle, after removing the calculated background, whilst more complex algorithms will take into account the shape of the spots and the distribution of the intensity values of each pixel in order to estimate the quality of the hybridisation.

### 3.7. Output results

The output of these programs is usually a table relating each spot to the corresponding values of intensity, background and quality measures and coordinates of the image and the grid, etc. However, there is as yet no standard output format.

## 4. Storage and organisation of experimental results

This is a critical and often underestimated step. Reproducibility and comparison of different experiments (usually from different laboratories) are both highly dependent on the existence of a common well-defined storage structure.

There is currently considerable international effort underway to standardise the way the information is stored, so that it can then be shared with other laboratories and combined with other experimental results. These international projects generally encompass: definition of the minimum set of data to store; development of relational database schema for storing the data; tools for submitting and retrieving the data; and definition of standard data structures (usually in XML format).

The most relevant public DNA array repository projects are currently (in alphabetic order):

- ArrayExpress
- GeneX
- GEO
- SMD

### 4.1. ArrayExpress European Bioinformatics Institute (EBI) <http://www.ebi.ac.uk/arrayexpress/>

This is the initiative where most effort has been made to date to define the set of information that should be stored for gene expression experiments. On the establishment of the closely-related Microarray Gene Expression Database Group (MGED, <http://www.mged.org>), four working groups were created to discuss the methods and components for a DNA array repository system. These groups deal with:

- Experiment description and data representation standards
- Microarray data XML exchange format
- Ontologies for sample description
- Normalisation, quality control and cross-platform comparison

A document describing the ‘minimum information about a microarray experiment’ (MIAME) was published recently (Brazma et al., 2001). The goal of the MIAME is to specify the minimum information that must be reported on a gene expression experiment. The ArrayExpress scheme then follows these specifications. Currently, most of the other database proposals tend to follow the MIAME document as well. Also, a data exchange format in XML (called MAGE-ML) has been proposed by the XML working group.

#### 4.2. GeneX. National Center for Genome Research (NCGR), University of California <http://www.ncgr.org/genex>

This system is designed to be installed locally and provides a free set of utilities, with some coming from external sources. The tools included are:

- Curation tool: A Java tool to assist the user in the correct submission of data to the database
- Database: An implementation of the proposed structure for storing the data from gene expression experiments. Currently this database is ready to run on a PostgreSQL manager
- XML data exchange protocol: A specific XML format definition (GeneXML) for exchanging data between users and platforms
- Query and analytical routines: A set of tools and programs that deal with the task of preparing and processing the data in order to get biologically relevant results and conclusions. These include clustering algorithms, viewing tools, principal component analysis and statistical packages (the *R* package).

Among the best features of this project is the fact that it is an open project where anyone can participate and provide information on bugs.

Also, all software used is free for academic users, including the database manager. The possibility of storing BLAST results for sequences is also an advantage here.

#### 4.3. GEO (Gene Expression Omnibus), National Center for Biotechnology Information (NCBI), USA <http://www.ncbi.nlm.nih.gov/geo/>

This repository is the only one that already accepts gene expression data from external sources. The data to store is divided into four types:

- Submitter: contact and login information on the submitter
- Platform: information on the reagents and materials used to make the gene expression measurements
- Sample: information on the mRNA samples and the experimental conditions, including the signal measurements generated
- Series: information on the relationship between different samples, analyses and experiments

GEO curators have developed several web tools and detailed file formats (simple omnibus format in text ‘SOFT’) to facilitate the submission of data from any user.

Also, the retrieval of data from the database is very easy and intuitive thanks to the well-known ENTREZ system (in this case called ENTREZ Probeset).

#### 4.4. SMD (Stanford Microarray Database) Stanford University <http://genome-www5.stanford.edu/microarray/smdl>

The SMD is an on-line repository of gene expression data developed at Stanford University, where microarrays were ‘born’. Currently it is possible (and very easy) to retrieve public gene expression data in raw or prepared format, ready to be used with external analysis software. Images of the hybridisations are also available. To submit your own data, you must be a registered user.

As has been implemented using the ORACLE database system, this database is very robust

technically and will support huge amounts of data. This feature is critical for a generic repository of gene expression data, where it is expected that the volume of information to be stored will grow rapidly.

## 5. Comparison of expression profiles

The possibility of determining in a single experiment the expression level of thousands of genes opens up the possibility of obtaining answers to biological questions from a genomic perspective. In general, there are two types of experiments: those involving the comparison of two conditions (typically the condition of interest versus a reference) and those involving the study of many conditions (e.g. time courses, dosage series, series of patients, tissues, etc.). Comparison of two conditions involves determination of those genes whose expression levels change significantly with respect to the reference condition and this just requires the application of a simple test. On the other hand, multi-condition experiments provide answers to more complex questions and involve more sophisticated methods for their analysis.

Typically, multi-condition experiments are represented by a matrix of gene expression values, with genes in rows and conditions in columns. Depending on the experiment, the values of gene expression can be used to classify conditions (columns) or gene expression profiles (rows). Both cases involve an initial grouping step, either to obtain sets of conditions with similar gene expression values or to obtain sets of genes with similar expression profiles for the conditions studied. Classification of different types of cancers is a typical example of the first type of experiment. The molecular signature of the different tumoral tissues has been demonstrated to be a valuable diagnostic tool (see, for example, [Alizadeh et al., 2000](#); [Scherf et al., 2000](#)). The second type of experiment usually involves the study of time series or dosage series to detect which genes display highly-correlated expression patterns. These genes are likely to be playing similar roles in the cell (see, for example, [Eisen et al., 1998](#); [Wen et al., 1998](#); [Brown et al., 2000](#)). The grouping is usually

performed using clustering methods. The various properties of the distinct clustering methods make them suitable for one, both or none of the types of comparison.

There are two different ways of approaching the clustering problem. If no external information is used in the clustering process then this is ‘unsupervised’ clustering. There is then another family of techniques, known as supervised clustering methods, which makes use of the information available on the groups that are sought.

Unsupervised clustering comprises of a number of techniques that produce arrangements of the data based on a distance function. Despite the arsenal of methods used, the optimal way of classifying gene expression data by unsupervised methods is still open to debate. We will discuss some of the virtues and pitfalls of the most frequently used methods. A full discussion is outside the scope of this article.

[Table 1](#) shows a list of the methods that are currently most used for clustering, arranged on the basis of their properties and underlying algorithm.

Clustering techniques can be used in combination with other exploratory techniques, such as principal component analysis (PCA), that help the user to view the complexity of the data in a two- or three-dimensional space, allowing groups of genes to be identified. Other related techniques, such as singular value decomposition (SVD) or correspondence analysis have also been applied to the clustering of gene expression patterns. Nevertheless, some authors have pointed out that these exploratory techniques can produce misleading results when applied to gene expression data.

Data can be clustered in two different ways: in a hierarchical or non-hierarchical manner. Hierarchical clustering allows detection of higher-order relationships between clusters of profiles, whereas the majority of non-hierarchical classification techniques work by allocating expression profiles to a pre-defined number of clusters, with no assumptions as to the inter-cluster relationships. Aggregative hierarchical clustering in its different variants (average-linkage, single-linkage, complete-linkage, etc.) ([Sneath and Sokal, 1973](#)) is still one of the most popular choices for the analysis of patterns of gene expression. This is in part due to

Table 1  
The most commonly-used methods for clustering and their properties

	Non-hierarchical	Hierarchical	Properties
Standard	SVD; <i>k</i> -means quality cluster	Aggregative hierarchical clustering	Sensitive to noise. Slow
ANN-based	SOM	SOTA	Robust. Fast
Properties		Provides information on the relationships between clusters	

the availability of software for running these methods, either within standard statistical packages or those designed specifically for gene expression data (Eisen et al., 1998). Standard aggregative hierarchical clustering produces a representation of the data in the shape of a binary tree, where the most similar patterns are clustered in a hierarchy of nested subsets (Sneath and Sokal, 1973). This method has been used to analyse various different datasets, from yeast (Eisen et al., 1998) to human cells (Wen et al., 1998; Scherf et al., 2000).

As an alternative to hierarchical clustering, non-hierarchical methods, such as quality clusters or *k*-means (Tavazoie et al., 1999) have been used. These algorithms start with a pre-defined number of clusters and, by iterative reallocation of cluster members, minimise the overall intra-cluster dispersion. A criticism of this approach concerns the fact that the number of clusters must be fixed from the outset of the procedure. Other authors have proposed different versions of a progressive *k*-means procedure that finds the number of different clusters from the data itself and is independent of an a priori specified number of clusters.

Standard hierarchical clustering works very well for clustering conditions (represented by columns, i.e. a small number of items), but several authors (Tamayo et al., 1999) have noted that standard clustering methods are not very robust when applied to clustering thousands of gene expression profiles. In addition, typical runtimes of standard methods based on distance matrices can range from  $N^2$  to  $N^4$ , which makes them very slow when thousands of items are to be analysed. Neural networks have been proposed as an alternative for overcoming some of the above-mentioned problems (Tamayo et al., 1999; Törönen et al., 1999; Herrero et al., 2000). Unsupervised neural net-

works, such as self-organising maps (SOM) (Kohonen, 1990) or the self-organising tree algorithm (SOTA) (Dopazo and Carazo, 1997; <http://www.almabioinfo.com/sota>), provide a more robust framework, appropriate for clustering large amounts of noisy data. Because of their properties, neural networks are suitable for the analysis of gene expression patterns. They can deal with real-world data sets containing noisy, ill-defined items with irrelevant variables and outliers and whose statistical distributions do not need to be parametric.

Nevertheless, SOMs have some inherent problems. Firstly, they use a topology-preserving neural network. In other words, the number of clusters is arbitrarily fixed from the beginning, such as for *k*-means. In addition, the training of the network (and consequently, the definition of the clusters) depends on the number of items in each cluster. Thus, the clustering obtained is not proportional. If irrelevant data (e.g. invariant, 'flat' profiles) or some particular type of profile is over-represented, a SOM will produce an output in which this type of data will populate the vast majority of clusters. As a consequence of this, the most interesting profiles tend to map into only a few clusters and therefore the resolution is lower for these. Contrary to this, the clustering obtained using SOTA is proportional to the heterogeneity of the data, instead of to the number of items in each cluster. Thus, regardless of whether a given type of profile is abundant, all similar items will remain grouped together in a single cluster and will not affect the rest of the clustering process. This is because SOTA is distribution-preserving, while SOM is topology-preserving (Dopazo and Carazo, 1997).

The SOTA structure grows from the root of the tree, where all expression profiles are mapped to

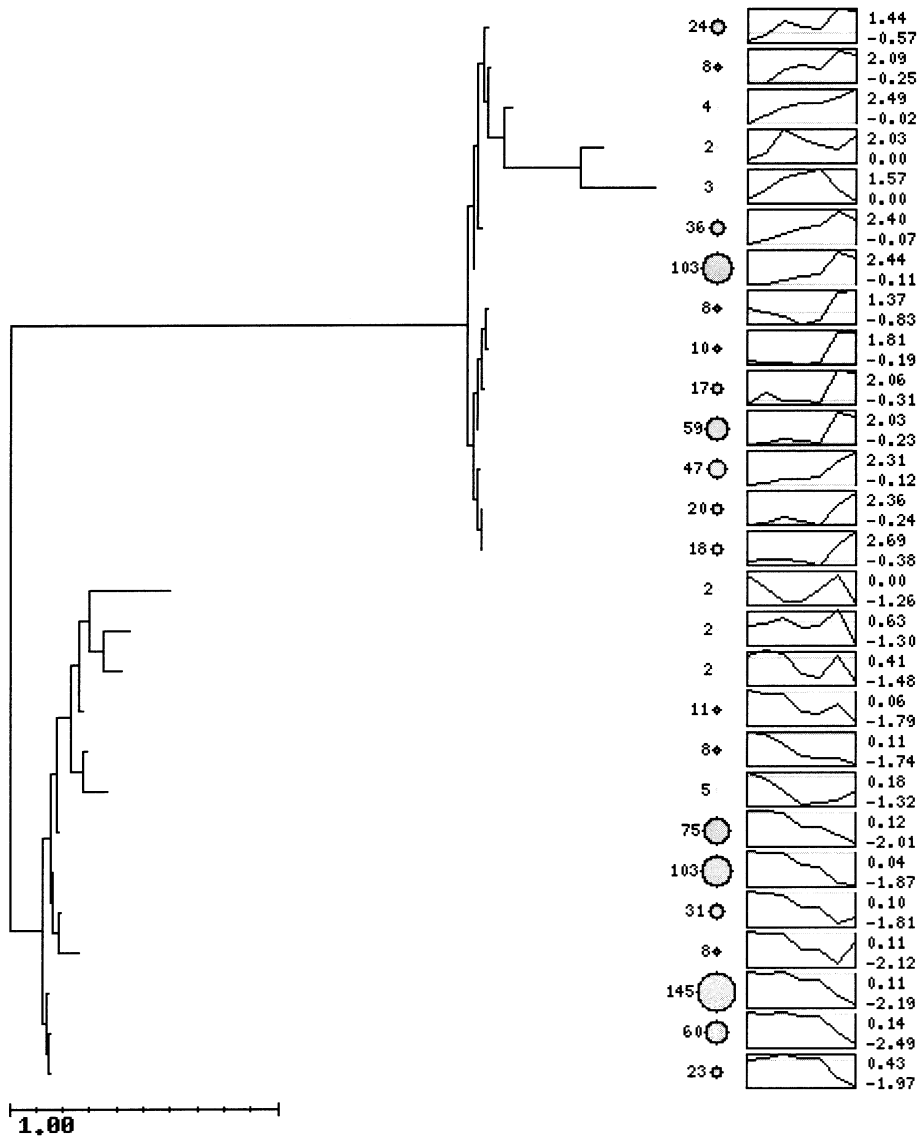


Fig. 2. Dendrogram obtained using the SOTA system for the data from the diauxic shift experiments (DeRisi et al., 1997). Only the 834 most significant patterns have been considered. Using a confidence level of 90%, the patterns have been clustered into 28 groups. The righthand part of the figure shows the size of each cluster and the average expression profile associated with it.

just one node, out towards the leaves, which contain only one profile if the tree is developed completely (Fig. 2). The algorithm proceeds by creating new terminal nodes, expanding the network from the node with the most heterogeneous population of associated gene expression profiles. The heterogeneity is measured using a dispersion

value (defined as the mean value of the distances between the node's own profile and the expression profiles associated with the node). The growth ends when the maximum dispersion value among all the terminal nodes reaches a certain threshold set by the user. The final structure can therefore be asymmetrical, including branches with different

numbers of nodes and can be stopped at the desired level. SOTA is one of the fastest available algorithms for performing hierarchical clustering.

Since the comparison operations are performed between the data and the average profiles in the nodes, the absence of some points (missing values) in a vector corresponding to a particular gene expression profile will have a negligible effect on the whole process of the network training. This renders unnecessary the use of methods for estimating missing values, required if average linkage or similar methods are used.

For most biological problems, there is some prior information available that can be exploited to produce clusters and further classify new data. There are a number of methods for supervised clustering, which are able to ‘learn’ from this information the features defining each cluster. The learning process makes use of a training set and later employs this knowledge to classify additional data. Support vector machines (SVM) were the first example of the application of machine learning methods to the classification of gene expression profiles. SVM are able to use prior information on the classes studied, for instance MIPS functional classes (Brown et al., 2000). SVM have also been used for classifying conditions. Supervised neural networks have also been applied recently to classifying conditions. Neural networks, in contrast to SVM, are able to discriminate amongst many different classes, and this is preferable for multi-class problems.

## 6. Functional interpretation

The main result of expression array experiments is the discovery of sets of genes with similar gene expression patterns (expression-based gene clusters). The underlying assumption is that these gene clusters are related by their participation in common biological processes (Lockhart et al., 1996). The operations carried out to define the ‘biological meaning’ of these clusters typically involve consulting functional annotations in different sequence databases such as SWISS-PROT (Bairoch and Apweiler, 2000) or other specialised databases. This information is often insufficient

and bibliographic information must be consulted, usually by following the links to selected Medline abstracts provided in some sequence databases. Since only a small fraction of these pointers provide direct information on gene function, further references are usually collected by querying Pubmed (<http://www.nlm.nih.gov/entrez/medline.html>) directly with gene names. In practice, analysis of the results of a full experiment can imply thousands of references, making systematic analysis of the differences between gene groups impractical. This situation will become increasingly complex for experiments involving larger systems, such as the human genome.

Development of methods to extract information on the common biological characteristics of gene clusters has received little attention, even if there is an obvious need for protocols for summarising the vast amounts of data in a comprehensive way, algorithms for selecting information that could be of use to human experts and tools for guiding them through the analysis.

The most promising developments in this area currently involve direct extraction of information from the related literature and databases, such that an initial idea of the functional characteristics associated with the various gene clusters can be suggested automatically, thus bypassing the time-consuming, tedious analysis that would otherwise be carried out by hand.

Some interesting approaches have been published that make use of the literature in relation to the analysis of DNA arrays. Of these, we can cite the following.

### 6.1. Medline

MedMiner (Tanabe et al., 1999) uses a pre-defined list of keywords which have been compiled for different domains of molecular biology and medicine in order to filter the abstracts returned from a Medline search and to select the sentences that best describe the document. In addition, information from GeneCards (<http://bioinfo.weizmann.ac.il/cards>) is used to obtain synonyms for the genes specified by the user and to extend the query. This information is presented via web pages that allow quick browsing of the results. It has

proved to be useful to some extent for the analysis of DNA array results as the overwhelming amount of text related to the genes involved in an experiment becomes easier to handle.

Jenssen et al. (2001) constructed a network of gene relationships for human simply by counting co-occurrences of gene symbols obtained from a public repository in Medline abstracts. These relationships were then compared to the results obtained by clustering the data from DNA expression arrays. This simple approach gives very interesting results because genes that are functionally related can show totally different expression patterns (and hence belong to different clusters). But for the cases where they appear in the same abstracts, their relationship is not evident in the experiments. This information can then be used to propose new experiments.

Shatkay et al. (2000) developed a method that detects similar documents to a given seed document. It is not based on a static similarity measure of the word frequencies in the different abstracts, but tries to detect the similarity of the ‘theme’ between texts, and associate abstracts and the query document. As a ‘side product’, keywords for each theme are extracted that aid interpretation by users. The objective is somewhat similar to that of Jenssen et al. (2001), as genes with the same themes in different clusters point to a relationship between groups that was not detected in the experiments. The main problem with the method is that a ‘kernel document’ for each gene has to be selected and the automatic procedure for doing so was not put forward by the authors, which will limit the application of this method to large-scale experiments, as the selection of this initial document will influence the results considerably.

## 6.2. Gene expression information system for human analysis

Gene expression information system for human analysis (GEISHA; Blaschke et al., 2001; Oliveros et al., 2000) is based on statistical evaluation of texts from scientific literature associated with gene expression clusters. As with other statistical methods (Andrade and Valencia, 1998; Blaschke et al., 1999), GEISHA is based on the frequency of

occurrence of words in a text corpus associated with gene clusters. The significant patterns detected are then used to characterise the corresponding groups of genes or proteins.

The GEISHA system uses clusters of genes as a framework for clustering the text corpus. Each cluster of genes is therefore associated with a subset of the literature. The algorithm calculates the frequency of relevant words in these literature clusters and compares these frequencies with those of the background, in order to assess their statistical relevance. A similar procedure is applied to the selection of combinations of pairs of words (as biological information is often expressed in composite terms, such as ‘DNA polymerase’ or ‘RNA polymerase’) and finally, to the extraction of complete sentences specific to the various gene clusters.

By comparing the frequency of a term in a given cluster with its frequencies in other clusters, the statistical significance of a term for a given cluster can be computed (simply said, a significant term is one that appears in one or few clusters with a frequency significantly higher than in all the other clusters). This significance is currently calculated using a Z-score. Composite terms (e.g. ‘DNA analysis’ or ‘cell cycle’) are analysed by comparing the frequency of a word pair with the expected value based on the frequencies of the individual words, with selection of those exhibiting significantly higher co-occurrence.

The results of the GEISHA system have been extensively compared to annotations provided by databases and human experts, showing that, in many cases, GEISHA was able to extract relevant or alternative information to that provided by other sources.

An example of the performance of the GEISHA system can be found in Fig. 3. Here the SOTA algorithm has been used to cluster the genes into hierarchical groups and GEISHA has then been used to extract relevant information for the groups. In the first example (Fig. 3a), a clear correspondence between the quality of the expression clusters and the associated biological information is shown. The parent cluster here holds genes that correspond to histones and diverse functions, such as cell wall maintenance and

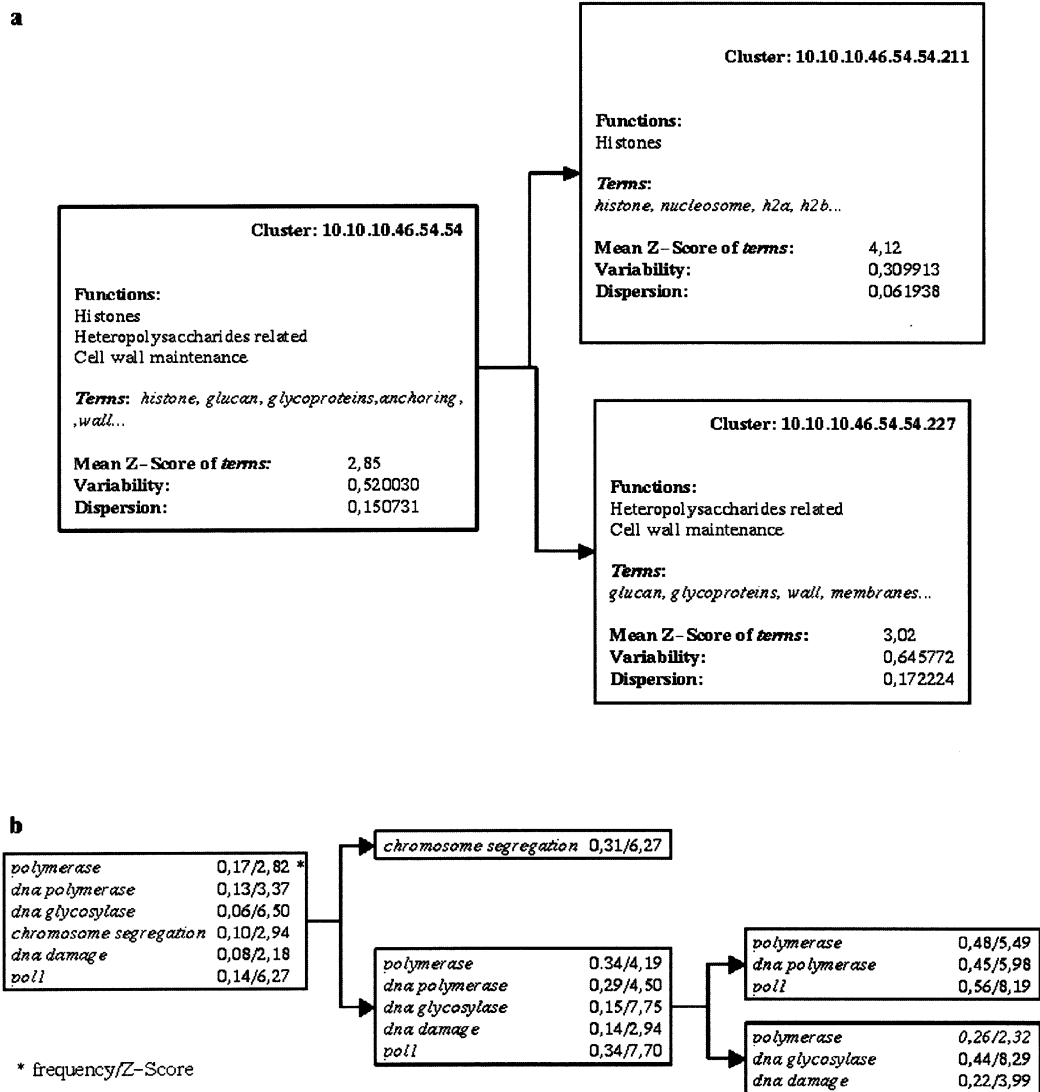


Fig. 3. Example of the clustering of expression profiles and biological terms. (a) Cluster of expression profiles for histone genes and (b) cluster of DNA-associated proteins. The figure shows two snapshots of the clustering process to illustrate the general tendency toward an increase in the frequency and specificity of the associated biological terms. Also included in (b) is a case where the reliability of the terms (as measured by the Z-score) decreases, indicating the simultaneous occurrence of different functions within a group of genes with similar expression patterns.

transport of polysaccharides. In most cases, the associated keywords clearly represent these functions (histone, glucan, wall, etc.). This cluster then gives rise to two daughter clusters: the first contains all the histone genes, with clear enhance-

ment in the similarity of the corresponding expression patterns and the quality of the corresponding terms (the average Z-score increases, with new keywords appearing such as 'histone', 'chromatin' and 'nucleosome'). The

other daughter cluster contains the genes related to cell wall maintenance, synthesis and transport of polysaccharides that do not share a unique biological function and consequently, the keyword *Z*-score only increases slightly.

The second example (Fig. 3b) illustrates a more complex reality. The parent cluster of proteins with different functions can be well described by the term ‘chromosome segregation’, which is significant for the description of some of the functions associated with the parent node. However, when the parent node is further divided, this term happens to be very specific for one of the daughter nodes. A similar observation is made in respect of other terms such as ‘DNA glycosylase’, ‘DNA polymerase’ or ‘DNA damage’.

An interesting case is the term ‘polymerase’, for which the *Z*-score rises until the last two nodes are divided. At this point, in one of the nodes its significance increases, while in the sister node it decreases to a value even lower than that of the parent group. This fact indicates that the information related to ‘polymerase’ is still present in both groups, showing that they are linked, but in only one of them does it become dominant, as the second group tends toward functions related to DNA glycosylation.

## 7. Conclusions

The analysis of DNA arrays has become one of the main research areas in computational biology and new methods and applications are continually being developed.

Here we have reviewed the current state of the principal areas involved: image analysis, data management, array design, data clustering and functional interpretation. We expect that new methodologies and tools in these areas will help improve the results obtained from experiments. We have focused this review on the possibilities offered by new methodologies in data clustering and functional analysis and have described some of the solutions that our groups propose for these.

## References

- Akman, L., Aksoy, S., 2001. A novel application of gene arrays: *Escherichia coli* array provides insight into the biology of the obligate endosymbiont of tsetse flies. *Proc. Natl. Acad. Sci. USA* 98, 7546–7551.
- Alizadeh, A.A., Eisen, M.B., et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511.
- Andrade, M.A., Valencia, A., 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14, 600–607.
- Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* 28, 46–48.
- Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., Small, P.M., 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* 284, 1520–1523.
- Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A., 1999. Automatic extraction of biological information from scientific text: protein–protein interactions. *ISM B99*, 60–67.
- Blaschke, C., Oliveros, J.C., Valencia, A., 2001. Mining functional information associated with expression arrays. *Func. Integr. Genom.* 256–268 (see also <http://www.mont-blanc.cnb.uam.es/geisha/index.html>).
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29 (4), 365–371.
- Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.J., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97, 262–267.
- Cho, R.J., Campbell, M.J., Winzler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* 2, 65–73.
- Cohen, B.A., Mitra, R.D., Hughes, J.D., Church, G.M., 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186.
- Cummings, C.A., Relman, D.A., 2000. Using DNA microarrays to study host–microbe interactions. *Emerg. Infect. Dis.* 6, 513–525.
- Debouck, C., Goodfellow, P.N., 1999. DNA microarrays in drug discovery and development. *Nat. Genet.* 21, 48–50.
- deRisi, J.L., Iyer, V.R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

- Dopazo, J., Carazo, J.M., 1997. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* 44, 226–233.
- Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.
- Evans, W.E., Relling, M.V., 1999. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286, 487–491.
- Ferea, T.L., Brown, P.O., 1999. Observing the living genome. *Curr. Opin. Genet. Dev.* 9, 715–722.
- Geiss, G.K., Bumgarner, R.E., An, M.C., Agy, M.B., van't Wout, A.B., Hammersmark, E., Carter, V.S., Upchurch, D., Mullins, J.I., Katze, M.G., 2000. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology* 266, 8–16.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.
- Gray, N.S., Wodicka, L., Thunnissen, A.M., Norman, T.C., Kwon, S., Espinoza, F.H., Morgan, D.O., Barnes, G., LeClerc, S., Meijer, L., et al., 1998. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533–538.
- Hayward, R.E., DeRisi, J.L., Alfadhli, S., Kaslow, D.C., Brown, P.O., Rathod, P.K., 2000. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* 35, 6–14.
- Herrero, J., Valencia, A., Dopazo, J., 2000. A hierarchical unsupervised growing neural network for clustering gene expression patterns (submitted).
- Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., Hood, L., 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934.
- Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J.J., Boguski, M.S., et al., 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.
- Jenssen, T.K., Lægreid, A., Komorowski, J., Hovig, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28.
- Kohonen, T., 1990. The self-organizing map. *Proc. IEEE* 78, 1464–1480.
- Kozian, D.H., Kirschbaum, B.J., 1999. Comparative gene-expression analysis. *Trends Biotechnol.* 17, 73–78.
- Leemans, R., Loop, T., Egger, B., He, H., Kammermeier, L., Hartmann, B., Certa, U., Reichert, H., Hirth, F., 2001. Identification of candidate downstream genes for the homeodomain transcription factor labial in drosophila through oligonucleotide-array transcript imaging. *Genome Biol* 2: Research 0015.0011–0015.0019.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, P.O., 1996. Expression monitoring by hybridization to high density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680.
- Loftus, S.K., Chen, Y., Gooden, G., Ryan, J.F., Birzniers, G., Hilliard, M., Baxevasis, A.D., Bittner, M., Meltzer, P., Trent, J., Pavan, W., 1999. Informatic selection of a neural crest-melanocyte cDNA set for microarray analysis. *Proc. Natl. Acad. Sci. USA* 96, 9277–9280.
- Marton, M.J., deRisi, J.L., Bennett, H.A., Iyer, V.R., Meyer, M.R., Roberts, C.J., Stoughton, R., Burchard, J., Slade, D., Dai, H., et al., 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* 4, 1293–1301.
- Marton, M.J., DeRisi, J.L., Bennet, H.A., Iyer, V.R., Meyer, M.R., Robert, C.J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Basset, D.E., Hartwell, L.H., Brown, P.O., Friend, S.H., 1998. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* 4, 1293–1301.
- Oliveros, J.C., Blaschke, C., Herrero, J., Dopazo, J., Valencia, A., 2000. Expression profiles and biological function. *Genome Inform. Ser. Workshop Genome Inform.* 11, 106–117.
- Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., Blattner, F.R., 1999. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.* 27, 3821–3835.
- Rockett, J.C., Luft, J.C., Garges, J.B., Krawetz, S.A., Hughes, M.R., Kirn, K.H., Oudes, A.J., Dix, D.J., 2001. Development of a 950-gene DNA array for examining gene expression patterns in mouse testis. *Genome Biol* 2: Research 0014.1–0014.9.
- Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., et al., 2000. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244.
- Shatky, H., Edwards, S., Wilbur, W.J., Boguski, M., 2000. Genes, themes and microarrays. Using information retrieval for large-scale gene analysis. *ISM B2000*, 317–328.
- Sneath, P.H.A., Sokal, R.R., 1973. *Numerical Taxonomy*. Freeman, San Francisco.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96, 2907–2912.
- Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L., Weinstein, J.N., 1999. MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* 27, 1210–1217.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285.
- Törönen, P., Kolehmainen, M., Wong, G., Castrén, E., 1999. Analysis of gene expression data using self-organizing maps. *FEBS Letts.* 451, 142–146.

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R., 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* 95, 334–339.

Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., Schoolnik, G.K., 1999. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc. Natl. Acad. Sci. USA* 96, 12833–12838.