Statistical Approach to DNA Chip Analysis

N.M. SVRAKIC,* O. NESIC,† M.R.K. DASU,‡ D. HERNDON,‡ AND J.R. PEREZ-POLO†

**Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri 63110; †Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston, Texas 77550; ‡Shriners Hospital for Children, Department of Surgery, University of Texas Medical Branch, Galveston, Texas 77550*

ABSTRACT

Statistical methods for analyzing data from DNA microarray experiments are reviewed. Specifically, we discuss common experimental setups, methods for data reduction and clustering, and analysis of time-course experiments. While early microarray studies focused mainly on the basic methodological and technical aspects of DNA arrays, emphasis has shifted to biological, medical, and clinical applications. We mention several of these and present results from our recent research as illustrative examples. New developments in this ever-growing field are outlined.

I. Introduction and Outline

The recent development of DNA array technology, or DNA arrays — also called microarrays, gene chips, DNA chips, and biochips — has enabled researchers to monitor simultaneously levels of thousands of genes as they are expressed in tissues, cell lines, patient samples, etc., at particular times and under a variety of different conditions. Developed in 1996 (Lockhart *et al.*, 1996), this new technology has experienced a remarkable growth in popularity and utility, as reflected in the number of papers published on the subject. A casual search of the literature shows that the number of publications dealing with DNA array technology has increased nearly exponentially over the past several years. In 1997, about eight papers were published on the subject, followed by 23 in 1998, 94 in 1999, 296 in 2000, and 1156 in 2001. From incomplete data for the first half of 2002, one finds about 800 papers citing some aspect of microarray technology.

While early microarray studies mostly focused on the basic methodological and technical aspects of DNA arrays (e.g., data normalization, error correction, replication), emphasis has shifted to biological, medical, and clinical applications. DNA chips are being used in pharmacogenomics and pharmacogenetics, toxicogenomics, human disease studies, disease screening, profiling and classification, diagnosis and clinical applications, and basic biological science studies. In each case, the experimental design has to be planned to fit the questions being

addressed. Following is a brief list describing the most-frequently used experimental designs.

In a typical experimental situation with microarrays, one may want to:

1. Compare gene expressions obtained from *two or more different tissues* — for example, healthy versus diseased tissue — in order to compare or classify them. This type of experimental design has been used, for instance, to get clues regarding mechanisms and causes of disease processes or to classify specific clinical varieties of cancers/tumors according to their expression profiles, in order to better predict prognosis (Golub *et al.*, 1999; Dirix and van Oosterom, 2002).

2. Compare gene expression data obtained from the *same tissue* or cell line *at different time points*, in order to follow the time course of expression. This design can be used to monitor temporal gene expression patterns during the cell cycle (Spellman *et al.*, 1998; Tamayo *et al.*, 1999) or during development (Wen *et al.*, 1998), temporal progression of a disease (Agrawal *et al.*, 2002; Pomeroy *et al*., 2002; Spies *et al*., 2002), or response to a treatment (Nesic *et al.*, 2002; Sotiriou *et al.*, 2002).

3. Compare gene expression data obtained from *different parts* of the *same tissue*, in order to reconstruct spatial distribution patterns of gene expressions. An expanded version of this design, called voxelation, was used recently by Brown *et al*. (2002), who correlated microarray data with the site of gene expression in tissues by creating signatures of expression patterns in coronal hemisections at the level of the hippocampus of the human brain. By combining the data for the entire surface of a volume of brain section, a three-dimensional spatial pattern of gene expression was generated. This important study (Peterson, 2002) combines DNA array technology and brain-imaging technique, like functional magnetic resonance imaging (fMRI), to represent the expression patterns of the whole organ.

Irrespective of the type of microarray employed (e.g., cDNA, oligonucleotide, spotted), such experiments generate tens of thousands of data points per each measurement. In addition, depending on the experimental design, or the number of samples, or the number of time points, the complete data set to be analyzed often contains hundreds of thousands of gene expression levels. These data are most commonly presented in tabular form (Quakenbush, 2001), called an expression matrix (see Table I).

In Table I, "Gene 1" is the name of the first gene, "Gene 2" the second gene, and so on. The column labeled "Experiment 1" lists data obtained from the first microarray (or under one condition, or measured at one time point), the column labeled "Experiment 2" lists the data obtained from the second microarray (or under a second condition, or measured at a second time point), etc. Finally, the entry (number) "$E_{i,k}$" is the measured expression level of gene *i* in the experiment *k*, so that the entry "$E_{1,1}$" is the expression level of "Gene 1" in "Experiment 1," the entry "$E_{1,2}$" is the expression level of "Gene 1" in "Experiment 2," etc.

TABLE I
Schematic Tabulation of a Typical Gene Expression Data Set ("Expression Matrix") in a Complete Microarray Experiment

	Experiment 1	Experiment 2	Experiment 3
Gene 1	$E_{1,1}$	$E_{1,2}$	$E_{1,3}$
Gene 2	$E_{2,1}$	$E_{2,2}$	$E_{2,3}$
Gene 3	$E_{3,1}$	$E_{3,2}$	$E_{3,3}$

Typically, a single microarray contains many thousands or tens of thousands of different probes ("Genes") and the complete experimental design may require measurements from tens or even hundreds of such microarrays ("Experiments"). Complete data are collected in a matrix similar to Table I of a size that may be $10{,}000 \times 50$, which translates to half a million entries to be analyzed.

The main difficulty in statistical analysis of such data sets stems primarily from the fact that one must deal with a small number of samples or "Experiments" (i.e., cell lines, patients, time points), relative to the large number of probes ("Genes"). Moreover, the unnormalized, raw expression levels of different genes in the same experiment (or under one condition, or at one time) (i.e., the numbers $E_{1,1}$, $E_{2,1}$, $E_{3,1}$…) may have values that range over several orders of magnitude — from values close to unity to values on the order of 10^5. The ultimate goals are to establish how the expression level of some gene changes from experiment to experiment and to identify groups of genes that exhibit similar coexpression patterns. Statistical methods designed to deal with these issues continue to be adapted and developed, since they are crucial for providing useful data and for extracting reliable biological information from DNA array experiments. This chapter reviews some of these methods, starting from the most basic and working towards more complex ones. Some of our results, obtained by using Affymetrix GeneChips, are described briefly in the form of illustrative examples.

II. Fold Changes in Expression Levels

The critical issue is statistical handling of expression data, as one typically wants to identify genes of potential interest and search for those that are systematically up- or downregulated across experiments. For this limited purpose, it suffices to perform simple statistical analysis of gene expression levels. Early papers reported such analyses by presenting a list of genes that show $\geq$ 2-fold change in expression level. But even with this simple analysis, care must be taken because of the aforementioned "large number of probes vs. small

number of experiments" problem. This is especially important if one wants to attach statistical significance to the observed changes. For instance, to determine the expression level of a single gene in one experiment (or under one condition), one needs to make several replicate measurements — the more the better. Performing many replicate experiments, however, often is not feasible, due to the high cost of DNA chips or the limited amount of RNA or DNA material available. Nevertheless, it has been our experience, in agreement with more-formal reliability studies (Lee *et al.*, 2000), that at least three replicates per experiment must be made to have reasonable statistical confidence in the expression values obtained. Once the expression value of a gene has been established through replicate experiments under one condition, one wants to compare that with the expression value of that same gene under some other condition. A usual way to make the comparison is through a 2-sided t-test, assuming normal distribution of replicated expressions, or by some other non-parametric method. With three or so replicates per experiment, the statistical significance of the difference between the two experiments typically is not very impressive and only those genes that exhibit large up- or downregulation between the two experiments can be identified with some confidence. Thus, due to the various sources of errors or chance variations between two measurements, DNA arrays cannot be used with great confidence to detect small (i.e., less than 1.5-fold changes) in expression levels across experiments. Even with this constraint, one is left with, for example, 1000 genes that are identified as significantly changed. To assign some confidence level to this finding, one can perform t-tests, one for each gene, requiring 1000 total tests. Then, to correct for the repeated testing, one can impose the usual Bonferroni correction to the individual significance levels (i.e., require that the p-value for each gene be 1000 times smaller than, say, 0.05). Unfortunately, under these settings, the Bonferroni condition turns out to be extremely restrictive and almost no genes with significantly changed expression levels are detected with required statistical confidence.

One way out of this impasse was suggested by Tusher and coworkers (2001), who proposed a new method, significance analysis of microarrays (SAM). The SAM procedure assigns "observed score" to each gene, depending on that gene's expression level scaled by the standard deviation of replicated measurements. Next, a number of "balanced" permutations of expression values are performed and a similar score in each case is assigned, which is then finally averaged over all permutations to compute the "expected score." The scatter plot of observed vs. expected scores is then used to identify significant changes in gene expression. With the additional adjustable threshold parameter, parameter Δ in the original article (Tusher *et al.*, 2001), one can control the overall false discovery rate (FDR), the percentage of genes discovered to be potentially significant by chance alone. With an FDR of $\approx$ 5–10%, which is deemed acceptable, one is still

left with dozens of genes that show statistically significant changes in expression levels. SAM analysis has become a standard statistical technique for detecting groups of genes with potentially significant change in their expression levels. (SAM software is available at http://www-stat.stanford.edu/~tibs/SAM/index.html.) Following such analysis, one can compile a table of significantly over- or underexpressed genes, under different circumstances, with the expectation that these genes most actively participate in the phenomenon under study.

III. Data Classification and Clustering

To go a step beyond simple recording of fold changes of gene expression levels, various methods of data reduction and classification have been devised to identify groups of genes that show similar expression patterns. To present the results of such classification, it is useful to have an intuitive visual representation (Eisen *et al.*, 1998). This often is achieved by drawing dendrograms and/or color-coded representations of similarly expressed genes. The most-common approach to perform classification or grouping of data is by one of the many clustering methods. Even though clustering methods are deterministic and reproducible, they still are subjective, since they may yield different results depending on the selected algorithm, normalization, distance metrics, etc. The challenge is to select the most-suitable one for the purpose of the experiment, so that the clustering produces results appropriate to the question being asked or the hypothesis being tested.

To illustrate the issues involved, consider Table I. With each "Gene" (e.g., "Gene 2") from this table, one can associate an "expression vector" with entries "$E_{2,1}$, $E_{2,2}$, $E_{2,3}$…" that are simply read off from the *row* corresponding to that gene. In other words, the expression vector of a gene contains expression levels of that gene in different experiments. The number of components (dimension) of the expression vector equals the number of experiments (N_E) and the number of expression vectors equals the number of genes (N_G). Geometrically, one can think of the expression vector as a point (tip of the expression vector) in the N_E-dimensional "expression space," so that each gene is uniquely assigned a single point. The dimensionality of the expression space is equal to the number of experiments (typically, between 10 and 100), while the number of points in this space is equal to the number of genes (typically, several thousand). In order to group the genes (or points in expression space) into clusters, one needs to define some measure of distance between them. The most-straightforward and most commonly used one is the geometric, Euclidean distance between the two points (expression vectors) *i* and *j*, the square of which is defined as

$$D_{i,j}^{\,2} = \sum_{k} (E_{i,k} - E_{j,k})^2$$

where the sum runs over all experiments *k* and Es are the appropriate expression levels from Table I. The most-similar points are the ones with the shortest Euclidean distance between them. Another possibility is to use some nongeometric measure of similarity, such as the Pearson correlation, which basically measures how similar are the directions in which the two expression vectors point. Thus, one attributes greatest similarity to the points with the highest correlation score. This method is widely used but has the drawback that it may sometimes falsely attribute high correlation score to expression vectors that are dissimilar. This may happen when there is an outlier in the data, such that overall expression levels of two genes are unrelated but for a single experiment, in which there is a common large peak, thus producing artificially high correlation. This can be remedied by employing a different correlation measure, a jackknife correlation, which is robust to single outliers, as proposed by Heyer *et al.* (1999). Many other distance measures can be used but discussing them is beyond the scope of this chapter.

As a somewhat "orthogonal" procedure to gene clustering, it often is useful to perform experiment clustering. To achieve this, one can represent each experiment by an "experiment vector," with its entries read off from the corresponding *column* of the expression matrix (see Table I). Thus, "Experiment 2" would be represented by an experiment vector with entries "$E_{1,2}$, $E_{2,2}$, $E_{3,2}$, etc." The number of experiment vectors equals the number of different experiments, while the length (dimension) of this vector equals the number of genes. Each point in this "experiment space" corresponds to one experiment. By introducing appropriate distance measure between two experiment vectors, one then can cluster experiments according to their similarity.

Clustering experiments is particularly useful as a preliminary step to discover, for instance, eventual gross discrepancies between microarrays that may occur with faulty arrays or because of other systematic errors. As an illustration, we recently reported on a microarray experiment involving burn injury in rats (Spies *et al*., 2002), where gene expressions in the skin tissues from burned rats and normal rats were compared at four time points (2 hours, 6 hours, 24 hours, and 240 hours after the injury). We used three replicate experiments for each group: thus, 3 replicates $\times$ 2 groups $\times$ 4 time points $=$ 24 experiments (arrays). After clustering of experiments (arrays), it was discovered that one of the 24 arrays differed markedly from the rest. In this particular array, only about 800 genes were expressed, while in all others, the number of expressed genes averaged around 4000. This difference was immediately visible in the clustering of experiments. The faulty array was discarded from further analysis and a proper one was substituted.

Experiment clustering also can be used to determine the overall effect of treatment, or healing, on global expression profiles. For instance, in a recent study of spinal cord injury (SCI) in rats (Nesic *et al.*, 2002), we compared

expression levels from spinal cord tissues of 1) rats with injured cord, 2) rats with injured cord that were treated with N-methyl-D-aspartate (NMDA) receptor antagonist MK-801, and 3) control (sham) animals. We used three replicates per group and performed hierarchical clustering of nine experiments. The resulting dendrogram, shown in Figure 1, correctly demonstrated that, overall, the injured and MK-801-treated groups are more similar to each other than to the sham group.

In a similar manner, one can use experiment clustering to follow the overall healing process. To again cite an example from our study of burn injury in rats (Spies *et al*., 2002), after all 24 experiments (arrays) were clustered, it was evident that samples from burned skin 10 days after the injury were more similar to control (unburned) samples than to samples from other burned groups. This gave us the molecular imprint of the onset of healing process that already had started 10 days post-injury.

With the defined distance measure, whether in the expression space or in the experiment space, the next step is to select the appropriate clustering algorithm.

A. HIERARCHICAL CLUSTERING ALGORITHMS

Most gene-clustering algorithms are hierarchical. These methods are derived (Eisen *et al.*, 1998) from algorithms used to construct phylogenetic trees; the most-similar genes are clustered first, while those with more-diverse profiles are subsequently included in a stepwise hierarchy of increasing diversity. This means that, in the first clustering step, the single most-similar expression profiles are linked to form nodes, the most similar of which are linked further in the second clustering step, and so forth, until all nodes finally are linked and the complete hierarchical tree of proximities (dendrogram) is obtained. Starting from the second clustering step and higher, each node may consist of two or more objects. The distances between nodes must be recomputed at each step. This can be done,

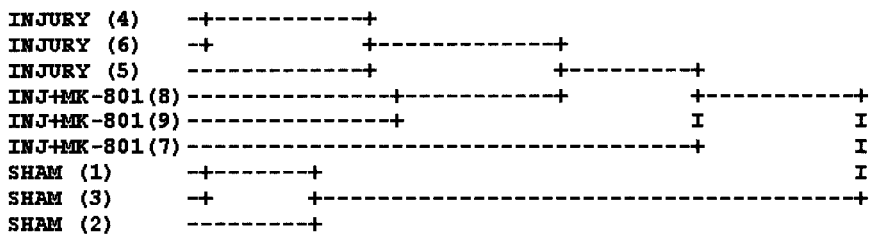

FIG. 1. Dendrogram obtained by hierarchical clustering using average linkage between normalized values of expression levels for 1322 probe pairs in nine DNA chips: three sham samples (1–3), three injured spinal cord samples (4–6), and three injured spinal cord samples treated with MK-801 (7–9).

for example, by computing the distance between the nodes as the average distance between its objects, as in the *average linkage* procedure, or as the distance between two of its closest objects, as in the *nearest-neighbor linkage* procedure. Other options include distances computed between the centers of mass of clusters or their modifications. In most cases, however, average linkage procedure is considered acceptable.

These different linkage choices are made to compensate for potential problems with hierarchical clustering. Namely, as clusters grow in size, at higher levels of hierarchy, the expression vector that represents the cluster may no longer be representative of any of the genes in the cluster. Thus, actual expression patterns of the genes themselves become less relevant on higher levels of hierarchy. If a gene is assigned to the "wrong" cluster, this error cannot be corrected later under hierarchical clustering.

B. NONHIERARCHICAL CLUSTERING ALGORITHMS

1. K-means Clustering

Sometimes, when *a priori* knowledge exists about the number of clusters that should be obtained, one can use nonhierarchical K-means clustering to partition the data. In this procedure, one first specifies the number of clusters (K), then randomly assigns expression vectors to them. Distances between clusters are recomputed, expression vectors are reassigned to the nearest cluster, and the procedure is iterated until the point is reached when no new assignments are made. The K-means clustering procedure simply partitions expression data into K groups and does not produce a dendrogram, although one can be constructed later by a hierarchical procedure.

2. Self-organizing Maps

Another frequently used nonhierarchical procedure is self-organizing maps (SOMs), a neural network-based procedure for clustering. In this algorithm, one also specifies in advance the number of clusters, chosen usually as the nodes of a grid. The nodes are mapped into K-dimensional space, initially at random, and then iteratively adjusted. During each iteration, a data point is randomly selected and the node is moved towards that point by the amount proportional to its proximity, so that more distant nodes are moved the least amount. In this way, neighboring points in the initial geometry are mapped to nearby points in the data space. This process usually is iterated tens of thousands of times. SOMs are particularly useful for exploratory data analysis, in order to expose the global patterns in the data.

C. PRINCIPAL COMPONENT ANALYSIS

A somewhat more-familiar method for data reduction is the singular value decomposition (SVD), or principal component analysis (PCA) as it is known in statistics (Alter *et al.*, 2000; Yeung and Ruzzo, 2001). In this procedure, expression data from the "genes $\times$ experiments" expression space are transformed to diagonalized "eigengenes $\times$ eigenexperiments" space, where eigengenes (or eigenexperiments) are unique orthonormal superpostion of genes (or experiments). PCA is essentially a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, with each succeeding component accounting for as much of the remaining variability as possible. This transformation represents the data in the new reduced coordinate space, in which individual genes or experiments appear to be classified into groups of similar functions or similar cellular state or phenotype. A simple illustration of PCA is given in Figure 2, in which the first principal component of a two-dimensional data set is shown by a straight line.

So far, the discussion has focused on the most common and most frequently used clustering algorithms. One should note, however, that other approaches exist, such as Bayesian and neural network algorithms, together with their numerous variants and modifications, that have been implemented in many DNA array studies.

FIG. 2. Principal component analysis (PCA) of a two-dimensional data cloud. The straight line shown is the direction of the first principal component, which gives an optimal (in the mean-square sense) linear reduction of dimension from two to one.

D. BIOMEDICAL APPLICATIONS OF CLUSTERING

Use of clustering/classification procedures in microarray experiments has been particularly fruitful in cancer research because cancers are complex, multigenic diseases with a natural control group for the analysis — noncancerous tissue (Alon *et al.*, 1999; Lin *et al.*, 2002). This was studied in prostate cancer behavior (Singh *et al.*, 2002), where a set of gene expression differences between healthy and diseased tissues was detectable at the time of diagnosis. Alternatively, clustering procedure can be used to compare cancerous tissues of the same type and to distinguish between clinical subtypes, as was done in two types of breast cancer (Hedenfalk *et al.*, 2001). The procedure also was very efficient in finding genes that distinguish between small blue cell tumors and leukemias (Tibshirani *et al.*, 2002) as well as in the discovery of a new subset of melanomas (Bittner *et al.*, 2000). The general conclusion drawn from these and other studies is that different cancers can be classified by the characteristic expression patterns of not more than dozens of genes. With more than 200 types of cancer, DNA microarray experiments are becoming an important tool to distinguish between their types and subtypes on the molecular level.

E. CLUSTERING OF TIME-COURSE EXPERIMENTS

An important class of DNA array experiments, in which data classification and clustering have been used successfully, are time-course experiments. In this setup, genome-wide expressions are measured at different time points in order to discover the temporal pattern in the course of development, or during a response to a treatment, or during a healing process. In this context, we mention the important pioneering work by Tamayo *et al.* (1999), where the temporal patterns of gene expression during the yeast cell cycle were classified by the SOM algorithm. The expression measurements were taken at 16 equally spaced, 10-minute intervals over two cell cycles (160 minutes), yielding a total of 30 different patterns. The classification was able to successfully extract yeast cell-cycle periodicity as the most prominent feature in the data and to select the appropriate group of genes that participate in the cycling process.

Following this work, numerous articles have reported results of temporal gene expression patterns under a variety of conditions. These include the temporal gene expression mapping of central nervous system development in rat's cervical spinal cord (Wen *et al.*, 1998); response of human bronchial cells to smoke and hydrogen peroxide (Yoneda *et al.*, 2001); differentially expressed genes in human myometrium during pregnancy and labor (Aguan *et al.*, 2000); and a range of experiments in toxicogenomics that measure response following exposure to toxicants, to identify drugs that provoke adverse reaction (Castle *et al.*, 2002). A large-scale study of development and metabolic pathways in mice, with approximately 1.8 million measurements of gene expressions based on 294

microarray analyses of 49 adult and embryonic tissues (Miki *et al.*, 2001), is perhaps the best illustration of the versatility of time-course DNA array experiments.

The time points where expression levels are measured in time-course experiments need not be equally spaced, since biologically important events often occur over different time scales. To give an example (Spies *et al.*, 2002), we analyzed the time course of healing and recovery of burn wounds in rats, with measurements made at the following four time points: 2 hours, 6 hours, 24 hours, and 240 hours after the burn injury.

The goal of our study was to identify local responses and initial cellular responses to skin thermal injury by comparing expression profiles in burned and unburned rat skin tissue. The associated genomic events include differential expression of genes involved in cell survival and death, growth regulation, metabolism, inflammation, and immune response. The dynamics of these events is most clearly seen when genes with similar temporal expression patterns are clustered together.

With only four data collection time points, the temporal change of the ratio of burned vs. unburned expressions can be analyzed in considerable detail. Note that, in this case, there are 27 possible dynamical patterns of temporal development. At each time point, the value of the expression ratio (burned vs. unburned) can be 1) increased with respect to the previous time, 2) decreased, or 3) remain the same as the value at previous time. Thus, starting at some base value at time 1, the expression ratio can change/not change at later time points 2, 3, and 4, giving $3^3 = 27$ possibilities of development (Figure 2).

More generally, with t time points, there are $3^{(t-1)}$ dynamical patterns. This number can become quite large quickly with increasing numbers of time points. For instance, with 17 time points equally spread over a 160-minute interval during two yeast cell cycles, as used in the already-mentioned work of Tamayo *et al.* (1999), there are over 43 million (3^{16}) mathematically possible different patterns. Of course, in this and similar cases, it is quite unrealistic and, indeed, completely unnecessary to consider all possible patterns in detail. What one needs is to classify existing data into a small number of characteristic patterns (clusters) with some global features like "clusters with peak expressions at 25–45 minutes and 85–105 minutes" (Tamayo *et al.*, 1999) that correspond to some meaningful biological events. This is precisely what a clustering algorithm such as SOM performs: it searches through the large "pattern space" for the small set of characteristic patterns that reflects global features of the entire set. Alternatively, in dynamical system parlance, one can think of the final characteristic clusters as attractors and of sets of patterns assigned to them as points that fall into their domains of attraction.

Returning to the example at hand, the complete set of 27 patterns that can be obtained with four time points is shown in Figure 3. The total number of genes in our data is 781.

Note that Figure 3 shows only the *generic shapes* of possible patterns, not the actual data, and that the four time points (1–4) are drawn as equidistant to simplify the graphics. By simple visual inspection, one can see that the "population" of patterns, specified by the number "N" in the figures, differs widely between the patterns, from N = 1 in pattern 4 to N = 133 in pattern 17. This indicates that some types of expression patterns are much more frequent than others. In particular, of 781 genes in this example, 88% (685) of them show significant over- or underexpression between time points 1 and 2 (i.e., in the period between 2 and 6 hours following the injury). The remaining 12% (96) do not change their expression levels appreciably during the same period. Moreover, just eight genes (patterns 2 and 3) exhibit increased (decreased) late-stage activity only, during the 24- to 240-hour period, without significant change in their activity prior to this time. The dynamics that emerges suggests that in the early phase (i.e., an hour or so after the injury), most genes (88%) involved in the entire 10-day process change their activity. This is consistent with the dynamics of the wound-healing process, which can be divided into an early phase of abrupt energy depletion and necrosis, followed by a two-stage inflammatory phase, delayed cell death, formation of granulation tissue, and matrix formation and remodeling (Spies *et al.*, 2002).

The particular set of genes participating in these processes can be analyzed in detail by examining each cluster separately. Consider, for illustration, cluster 6 (i.e., pattern 6) that includes 24 genes, as shown in Figure 4. The numbers 1–24 in this figure label the particular genes that belong to the cluster (their names are listed in the separate table, not included). In this and other figures, the time points are drawn as equidistant to simplify the drawing.

With images like this, one can see that all genes in this cluster show a peak of activity at time 3 (24 hours post-injury), as specified by generic pattern 6 from Figure 3. A better view of the patterns of change is obtained in the three-dimensional rendering of this image, where the third dimension is the value of the expression ratio (see Figure 5).

Yet another view of cluster 6, this time from the direction of time axis, shows more precisely the amount of over- or underexpression of the genes involved (Figure 6). From this view, one can simply read off the amount by which the genes in this cluster from the burned tissue are over- or underexpressed with respect to the unburned one. With this information from all clusters, and knowledge of the particular genes involved (especially their metabolic functions, protein products, etc.), one can reconstruct patterns of biological activity during the wound-healing process. A detailed presentation of this and other analyses for all clusters will be published separately (Nesic *et al.*, 2002).

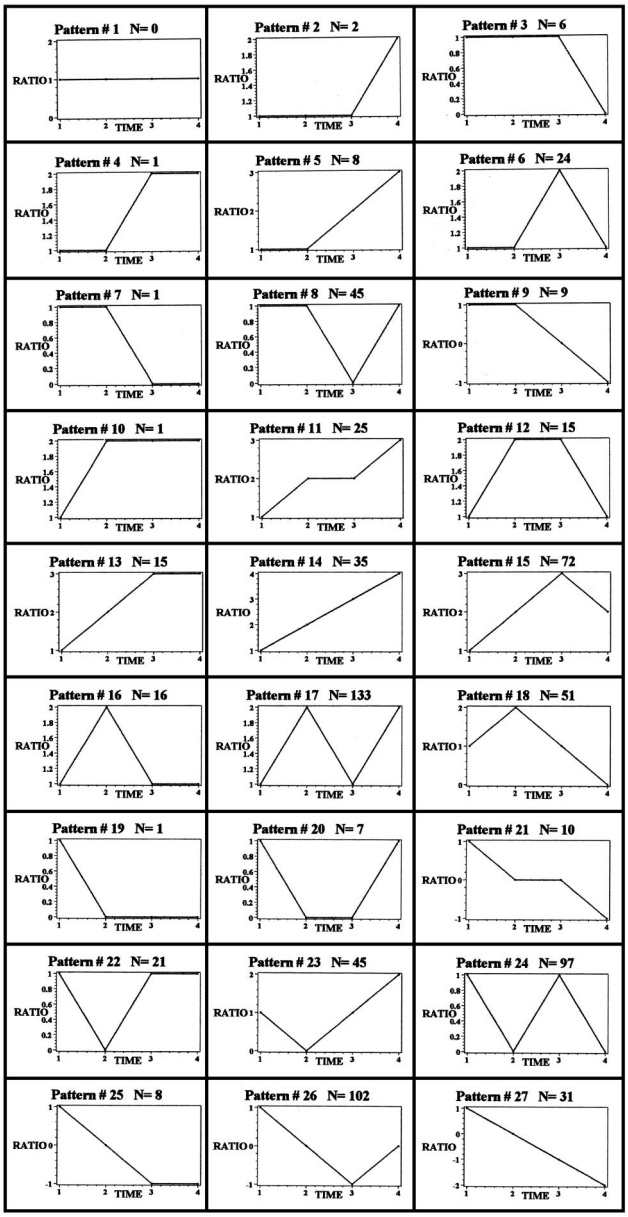

FIG. 3. Generic shapes of 27 possible dynamical patterns of gene expressions from four time-point measurements. Number "N" inside each box denotes the number of genes in our data that exhibit that pattern. The labeling (1–27) of patterns is arbitrary.

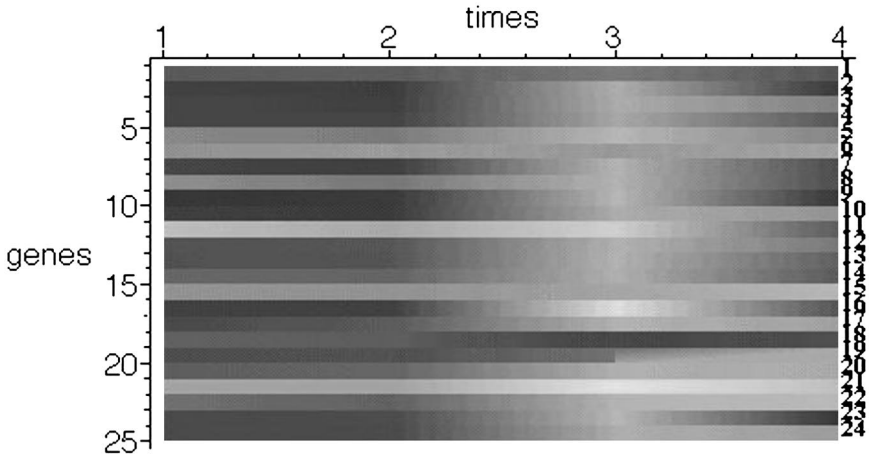

FIG. 4. Representation of the ratio of expression levels for 24 genes in cluster 6 over time points 1–4 (2 hours, 6 hours, 24 hours, and 240 hours, respectively). Note the common peak of expression ratio at time point 3, visible as lighter shade.

Clustering of gene expression patterns can be improved by including additional, more-complex relationships beyond simple coexpression that are implicit in time-course patterns. For example, a gene may activate or control another gene downstream in the pathway, thus introducing a time-delayed response. Another possibility is that two genes have opposing influences on each other, so that when the activity of one increases, the activity of the other decreases, producing inverted correlation. A study in this direction has been reported by Qian *et al.* (2001) where, instead of simple direct correlation, four different correlation measures between gene expression patterns have been taken into account: 1) simultaneous correlation, 2) time-delayed correlation, 3) inverted correlation, and 4) inverted and time-delayed correlation. The method was applied to the yeast cell-cycle data set of Tamayo *et al.* (1999) and new interactions were identified, implying new biological relationships between genes. Still, with this and other improvements such as time warping (Aach and Church, 2001) and dynamical modeling (Holter *et al.*, 2001; Ramoni *et al.*, 2002), much research remains to be done on the systematic classification and clustering of gene expression patterns from time-course DNA array experiments.

IV. Beyond Simple Clustering: Genetic Regulatory Networks

Clustering gene expressions into similar patterns usually is performed with the expectation that "coexpressed genes are coregulated," a plausible assumption

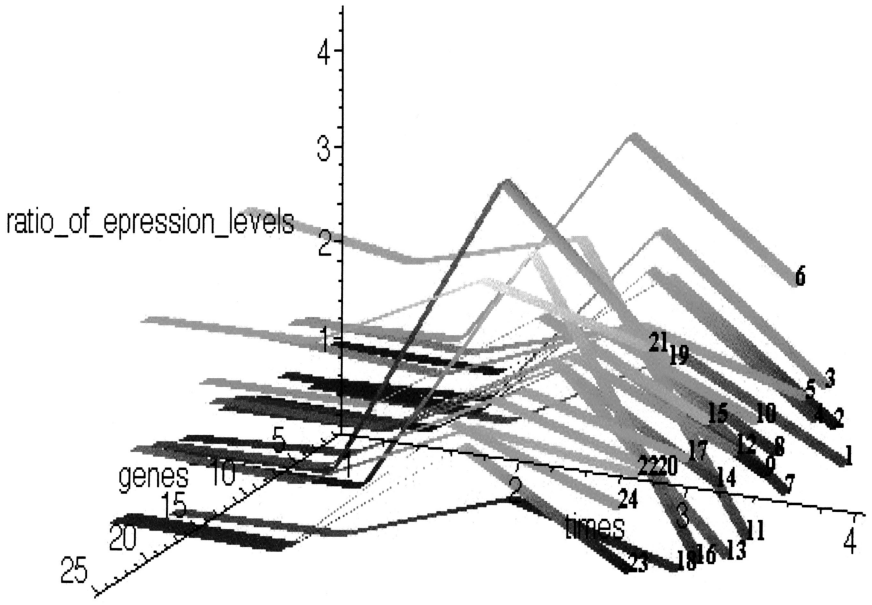

FIG. 5. Three-dimensional view of the ratio of expression levels for 24 genes in cluster 6. All genes exhibit common dynamical pattern: initial stagnation, then peak of activity at time 3 (24 hours after the injury), followed by a drop in activity at time 4 (10 days after the injury). Numbers 1–24 label the particular genes involved, specified in a separate table (not provided).

(Spellman *et al.*, 1998) that is, however, not universally true. Simultaneous detection of overexpression of two different genes does not necessarily imply that they are regulated by the same pathway, even if they appear together in the same cluster after a stimulus is applied to the cells. Many stimuli are known to initiate several different genomic-scale processes simultaneously, so that observed synchronicity in gene expression at certain times may be purely coincidental.

In order to move beyond simple coexpression, one has to establish which genes in some cluster also share common regulatory elements that control their expression levels (a group of genes regulated by a common element has been dubbed "regulons" in recent literature). The final goal of such analysis is to construct genetic regulatory networks and to identify the function of many thousands of novel genes (Tavazoie *et al.*, 1999). This approach has been successful in yeast (Lyons *et al.*, 2000), for which the complete sequence of promoter regions is known. Unfortunately, this is not the case with mammalian or other systems, where untranslated first exons, followed by introns greater than 10 kb in size, can make promoter identification

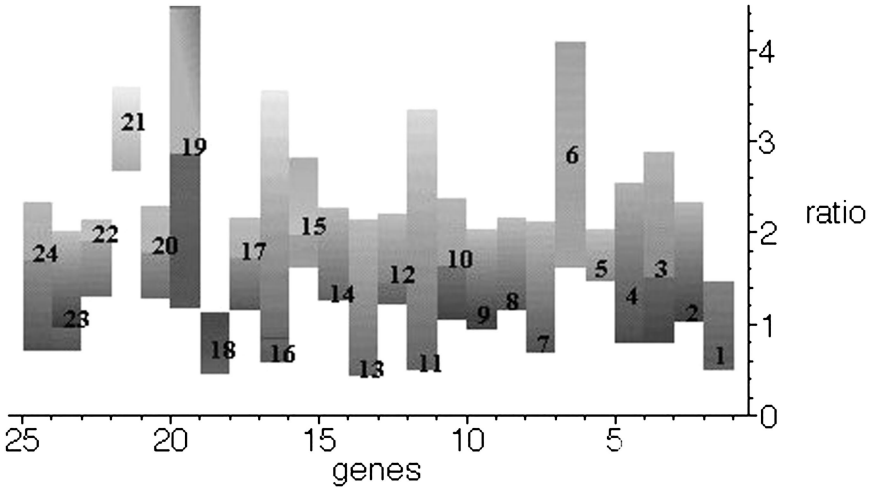

FIG. 6. Side view of the gene expression patterns shown in Figure 5, exhibiting the range of expression change over all time points.

extremely difficult. In many organisms, the promoter regions have not been fully sequenced. To construct the network for the phenomenon in question, one must use statistical algorithms for clustering and motif discovery in combination with genomic data, cis-regulatory analysis, and known molecular biology of the process studied. In spite of all the difficulties, several genetic regulatory networks, or parts of them, have been constructed. Davidson *et al.* (2002) have mapped a gene regulatory network in sea urchin embryo that controls the specification of endoderm and mesoderm. Such studies reveal that, in addition to comprehensive gene expression maps (Kim *et al.*, 2001) obtained by DNA array measurements, one needs as much other genome-wide information as can be mustered to unravel the intricate patterns of genetic interactions in biological processes.

Simultaneously with this development, additional knowledge is accumulating regarding the statistical nature of naturally occurring networks. Many biological networks (e.g., genetic, metabolic) exhibit "small world"-scale free behavior (Watts and Strogatz, 1998). This means that although the network may possess thousands of nodes, the path leading from one node to another is remarkably short. Such architecture may serve to minimize transition times between metabolic states or provide robustness against mutations (Fell and Wagner, 2000; Jeong *et al.*, 2000; Wagner, 2000). These new insights, combined with the knowledge of biological processes, may lead us for the first time towards understanding biology at the systems level (Kitano, 2002).

REFERENCES

Aach J, Church GM 2001 Aligning gene expression time series with time warping algorithm. Bioinformatics 17:495–508

Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D, Yeatman TJ 2002 Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling**.** J Natl Cancer Inst 94:513–521

Aguan K, Carvajal JA, Thompson LP, Weiner CP 2000 Application of a functional genomics approach to identify differentially expressed genes in human myometrium during pregnancy and labour. Mol Hum Reprod 6:1141–1145

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

Alter O, Brown PO, Botstein D 2000 Singular value decomposition for genome-wide expression data processing and modeling Proc Natl Acad Sci USA 97:10101–10106

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

Brown VM, Ossadtchi A, Khan AH, Cherry SR, Leahy RM, Smith DJ 2002 High-throughput imaging of brain gene expression. Genome Res 12:244–254

Castle AL, Carver MP Mendrick DL 2002 Toxicogenomics: a new revolution in drug safety. Drug Discov Today 7:728–736

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown TC, Livi CB, Lee PY, Revilla R, Alistair G, Rust AG, Pan Z-j, Schilstra MJ, Clarke PJC, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H 2002 A genomic regulatory network for development. Science 295:1669–1678

Dirix LY, van Oosterom AT 2002 Gene-expression profiling to classify soft-tissue sarcomas. Lancet 359:1263–1264

Eisen MB, Spellman PT, Brown PO, Botstein D 1998 Cluster analysis and display of genome-wide expression patterns Proc Natl Acad Sci USA 95:14863–14868

Fell D, Wagner A 2000 Small world of metabolism. Nat Biotech 189:1121–1122

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri M, et al. 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Hedenfalk I, Duggan DMS, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O-P, Wilfond B, Borg A, Trent J 2001 Gene-expression profiles in hereditary breast cancer. N Engl J Med 344:539–548

Heyer LJ, Kruglyak S, Yooseph S 1999 Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9:1106–1115

Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR 2001 Dynamic modeling of gene expression data. Proc Natl Acad Sci USA 98:1693–1698

Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A-L 2000 The large-scale origination of metabolic networks. Nature 407:651–654

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GM 2001 A gene expression map for *Caernohabditis elegans.* Science 293:2087–2092

Kitano H 2002 Systems biology: a brief overview. Science 295:1662–1664

Lee MM-L, Kuo FC, Whitmore GA, Sklar J 2000 Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridization. Proc Natl Acad Sci USA 97:9834–9839

Lin YM, Furukawa Y, Tsunoda T, Yue CT, Yang KC, Nakamura Y 2002 Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas**.** Oncogene 21:4120–4128

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo V, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL 1996 Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680

Lyons TJ, Gasch AP, Gaither LA, Botstein D, Brown PO, Eide DJ 2000 Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. Proc Natl Acad Sci USA 97:7957–7962

Miki R, Kadota K, Bono H, Mizuno Y, TomaruY, Carninci P, Itoh M, Shibata K, Kawai J, Konno H, Watanabe S, Sato K, Tokusumi Y, Kikuchi N, Ishii Y, Hamaguchi Y, Nishizuka I, Goto H, Nitanda H, Satomi S, Yoshiki A, Kusakabe M, DeRisi JL, Eisen M.B, Iyer VR, Brown PO, Muramatsu M, Shimada H, Okazaki Y, Hayashizaki Y 2001 Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci USA 98:2199–2204

Nesic O, Svrakic NM, Xu GY, McAdoo D, Westlund KN, Hulsebosch CE, Ye Z, Galante A, Soteropoulos P, Tolias P, Young W, Hart RP, Perez-Polo JR 2002 DNA microarray analysis of the contused spinal cord: Effect of NMDA receptor inhibition. J Neurosci Res 68:406–423

Peterson AS 2002 Pixelating the brain. Genome Res 12:217–218

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR 2002 Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415:436–442

Qian J, Dollet-Filhart M, Lin J, Yu H, Gerstein M 2001 Beyond synexpression relationship: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol 314:1053–1066

Quakenbush J 2001 Computational analysis of microarray data. Nat Rev Genet 2:418–427

Ramoni MF, Sebastiani P, Kohane S 2002 Cluster analysis of gene expression dynamics. Proc Natl Acad Sci USA 99:9121–9126

Singh D, Febbo P, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR 2002 Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1:203–209

Sotiriou C, Powles TJ, Dowsett M, Jazaeri AA, Feldman AL, Assersohn L, Gadisetti C, Libutti SK, Liu ET 2002 Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. Breast Cancer Res 4:R3

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B 1998 Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 9:3273–3297

Spies M, Dasu MRK, Svrakic NM, Nesic O, Barrow RE, Perez-Polo JR, Herndon DN 2002 Gene expression analysis in burn wounds of rats. Am J Physiol Regul Integr Comp Physiol 283(4):R918–R930

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T 1999 Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907–2912

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM 1999 Systematic determination of genetic network architecture. Nat Genet 22:281–285

Tibshirani R, Hastie T, Narasimhan B, Chu G 2002 Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99:6567–6572

Tusher VG, Tibshirani R, Chu G 2001 Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–5121

Wagner A 2000 Robustness against mutations in genetic network of yeast. Nat Genet 24:355–361

Watts D, Strogatz S 1998 Collective dynamics of 'small-world' networks. Nature 393:440–442

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R 1998 Large scale temporal gene expression mapping of the central nervous system development. Proc Natl Acad Sci USA 95:334–339

Yeung KY, Ruzzo WL 2001 Principal component analysis for clustering gene expression data. Bioinformatics 17:763–774

Yoneda K, Peck K, Chang MM-J, Chmiel K, Sher Y-P, Chen J, Yang P-C, Chen Y, Wu R 2001 Development of high-density DNA microarray membrane for profiling smoke- and hydrogen peroxide-induced genes in a human bronchial epithelial cell line. Am J Respir Crit Care Med 164:S85–S89