Erin Davies 3/10/03
BIOCHEM218 Final Project

A Critical Review of Computational Methods Used to Manage Microarray Data Sets

Introduction

Transcriptional profiling techniques, such as oligonucleotide and cDNA microarrays, are powerful technologies that enable biologists to study regulated gene expression at a global level. These methods are versatile, and can be implemented to dissect complex signaling networks, to rigorously define a compendium of cell or tissue-specific gene expression profiles, and to plot how the transcriptome of an organism changes during development. Biomedical applications for microarray technology are also being researched, and are focused on developing exquisitely sensitive molecular diagnostic reagents that can distinguish between related tumors and/or disease states.

Although such holistic methods have the potential to produce a wealth of information, the utility of large data sets is ultimately limited by the extent to which individuals leverage existing computational tools to relate and organize the data. At present, a systematic approach to microarray data analysis has not been prescribed by the field, leaving the choice of computational methods to the individual. The outcomes of commonly employed clustering algorithms are not identical, and often lend themselves to different interpretations of the data (5). This inherent subjectivity highlights the need to critically assess which computational method(s) will maximize the information content of the data set with respect to the biological question(s) being addressed. Several parametric and/or statistical approaches have been introduced to grapple with the computational challenges inherent in microarray experimentation, including hierarchical clustering, self-organizing maps (SOMs), k-means clustering, support vector machines (SVMs), and probabilistic relational models (PRMs). Each method aims to relate and reorganize the raw data set so that coexpressed genes are grouped together; sequence identifiers and functional annotation are often provided as well. In addition. reductions in dimensionality of the dataset, e.g. by primary component analysis (PCA), should allow researchers to hone in on gene(s) or cellular processes of interest with greater facility and precision (6). A critical evaluation of existing computational methods adopted for microarray data analysis is presented herein.

Experimental Protocol

Transcriptional profiling experiments aim to compare the relative abundance of mRNAs present in a cell population of interest relative to a reference sample. For example, one may wish to know what global changes in gene expression are elicited in a particular cell type in response to an environmental stress or pharmacological compound. Alternatively, one could learn how the gene expression signature of a particular mutant of interest differs from wildtype. Similar protocols are used for both oligonucleotide and cDNA microarrays; the latter is described in detail (for a comprehensive review, see references 4 and 5). First, mRNA is isolated from experimental and control cell populations, and differentially labeled cDNA populations are generated by reverse transcription in the presence of a fluorophore. Traditionally, experimental cDNAs are labeled with Cy5, whereas control cDNAs are labeled with Cy3. Purified cDNAs are then mixed in a 1:1 ratio, and are competitively hybridized to the microarray. Slides are subsequently washed and read using a fluorimeter, and the fluorescent intensiy of each spot (i.e., cDNA sample corresponding to a single gene) is read separately for the control (green) and experimental (red) channels. Spots that appear yellow have approximately equal amounts of control and experimental samples bound, whereas spots that are increasingly red or green are enriched for the experimental or reference samples, respectively. Black spots correspond to genes that are not expressed appreciably in either of the mRNA populations assayed. Software is available to normalize the fluorescent intensities of the two assay channels and to minimize noise, but will not be discussed in this review.

After normalization, the expression ratio of the experimental to control values is calculated for each spot, and is typically recorded as a $\log_2$[Cy5/Cy3] ratio in an n-dimensional expression matrix (n = number of genes in the experiment). Herein, each row represents a separate gene, whereas each column represents a single experiment. Conventional wisdom states that differentially expressed genes must reproducibly post a change in gene expression of ~2-fold in several independent trials (5). It should be noted that raw intensities can be recorded in lieu of expression ratios if comparisons of absolute gene expression levels are desired (5). This approach may rectify problems stemming from sample complexity in developmental profiling experiments, wherein low-level, developmentally-relevant changes in gene expression are overshadowed by abundant, tissue specific transcripts that may not be present in the reference sample. Alternatively, expression ratios can be transformed using methods that accentuate changes in gene expression profiles across several experiments, such as mean-centering(5). Such methods improve the signal to noise ratio prior to running the data through a clustering algorithm.

Experimentalists employ clustering algorithms to simply n-dimensional expression matrices in ways that reveal trends in the dataset, namely groups of co-expressed genes or tissue-specific expression signatures. Indeed, Eisen et. al. were among the first to demonstrate that hierarchical clustering isolates groups of co-expressed genes that are known to share a common function and/or operate in a common process (4). Expression matrices can also be clustered by column, e.g., by experiment, to identify subsets of differentially expressed genes in different cell or tumor types (5,6). Unsupervised clustering methods operate on raw data sets without ancillary information, and often depict expression patterns for each cluster graphically (2,4,5). Supervised clustering methods, on the other hand, incorporate unique identifiers and/or functional annotation into the analysis (2,5,6). Co-expression, a manifestation of similarity between two gene expression vectors, may be assessed using either geometric (Euclidean) distance methods or statistical approaches. The choice of distance metric is a nontrivial one, since different classifications are produced using different algorithms (5). Details on commonly used similarity measurements are included in subsequent discussion sections.

Unsupervised Clustering Methods: Hierarchical Clustering

Hierarchical clustering is the most frequently employed method for microarray data management. It is an agglomerative approach that computes similarity between gene expression vectors using distance methods that are reminiscent of those used for phylogenetic analysis (4,5). A dendrogram is produced using pairwise-similarity scores for genes in the expression matrix; each leaf represents an expression profile for a single gene, and co-expressed genes branch off of common nodes. Similarity scores are reflected by the branch lengths for any pair of genes in the tree. Data in the gene expression matrix can then be reordered according to the branching pattern delimited by the dendrogram, along with visual displays of individual gene expression profiles. This method was pioneered in the Brown and Botstein labs at Stanford, as reported in (4).

Similarity scores are first calculated in a pair-wise fashion for gene expression vectors in an n-dimensional expression matrix using a distance metric of choice. A node is generated for the highest scoring pair, its average gene expression vector is computed, and the distance between the node and the remainder of the matrix is recalculated. This process is iterated n-1 times, until all gene expression profiles have been incorporated into a single tree.

Geometric distance methods for computing similarity scores assume that distances are metric; i.e., the triangle inequality holds true. Therefore, a generalization of the Pythagorean theorem can be used to calculate the distance between gene expression vectors in n-dimensional expression space (5). Alternatively, correlation coefficients can be used as similarity measures (4,5). The dot product is calculated for each pair of gene expression vectors in the data set; thus correlation coefficients range in value from –1 to 1. For example, Eisen et. al. used the Pearson correlation coefficient in their pioneering study on hierarchical clustering of microarray data (4). The latter method may be preferable, since similarity scores are a reflection of the ‘shape’ of the expression profiles, rather than the magnitudes of the signals in question (4). The chosen distance method is then incorporated into a clustering algorithm; average-linkage clustering is routinely used, as its groupings have been shown to be biologically significant in a number of studies (4,5). This method is similar to the UPGMA tree-building method, which calculates distances based on average expression profiles for each cluster, and joins the two clusters separated by the smallest average distance (5). Additionally, average cluster profiles can be weighted according to size; i.e., the number of genes in a cluster (4,5). Randomization of the expression matrix by row, column, or both disrupts the clustering pattern generated using weighted averages and a Pearson coefficient, suggesting that the derived relationship is specific to the gene expression profiles in question (4). A significant task left to the researcher, however, is to determine whether the clusters produced are biologically significant. An example of an annotated cluster from Eisen et. al. is shown below; each row corresponds to a gene expression profile for an object that is involved in oxidative phosphorylation (4). Data from several independent studies, each corresponding to a column, was pooled for analysis.

Figure 1: An example of a regulon produced by the hierarchical clustering method of Eisen et. al. (4), which shows a group of co-expressed genes that are involved in aerobic respiration. Notice that the gene expression profiles are similar, but not identical over the range of conditions tested.

The hierarchical clustering software package introduced by the Brown and Botstein labs (TREEVIEW (4)) has proven to be an invaluable tool for the burgeoning functional genomics field because of its simplicity and the intuitive nature of its tree-diagrams and graphical output. From such clustering studies researchers can rediscover fundamental relationships that are known to exist in the cell, e.g., genes that encode functionally related proteins are often co-regulated, as well as gaining insight as to potential functions for novel and/or uncharacterized genes that cluster with genes of known function (4). Coincident expression patterns over range of experimental conditions increases the likelihood of uncovering a common transcriptional regulatory program for a given set of genes, and computational suites that assay noncoding regulatory regions for conserved transcripton factor binding sites may be informative in such cases. Cross-talk between signaling pathways may also be reflected in the complement of differentially expressed genes observed under different experimental conditions.

However, there are several pitfalls to be wary of when using hierarchical clustering algorithms. Like all clustering methods, the number of clusters produced and cluster composition vary with the choice of distance metric, and thus the subjective nature of the analysis should be kept in mind. Since the tree-building process is iterative, poorly aligned profiles have the potential to be propagated without recourse (5,6). Such a scenario may result in poorly-delimited, noisy clusters that may obscure relevant relationships in the dataset. Although the phylogenetic paradigm is useful, the similarity scores and branching patterns are not tantamount to evolutionary relationships amongst sequences. The use of a weighted averaging method in the agglomerative approach described here becomes increasingly problematic as the size of the dataset increases, since the weighted average may not accurately reflect the expression profiles of genes (or subgroups of genes) within the cluster (5). It should also be noted that both the geometric and statistical distance metrics described above assume linear relationships between objects. Perhaps methods that assume nonlinear relationships between objects, such as use of the Spearman coefficient or others that allow many-at-once comparisons, may better suit the needs of systems biologists who wish to construct a comprehensive network of specific transcriptomes.

k-Means Clustering

The use of multivariate, non-hierarchical clustering algorithms, such as the k-means approach, is also common practice amongst biologists mining microarray data. This unsupervised method partitions expression profiles into a predetermined number of clusters, such that similarity scores are maximized within clusters and minimized between clusters (5,6,10). Since the number of groupings, k, is specified by the user, hierarchical clustering or primary component analysis (PCA) are often performed first to estimate of the number of regulons in the data set (5). PCA is a technique that reduces the dimensionality of the dataset through a mathematical manipulation that projects data in n-dimensional space onto a Cartesian coordinate system (5). Viewing clusters in three dimensions provides a visually intuitive interface for the user to make general assessments regarding the diversity and content of the dataset. Alternatively, clusters may be optimized by running the algorithm with different k values (10). Since the success of the analysis rests on the quality of the clusters produced, many experimentalists tend to over-estimate the expected cluster number (6,10). Secondary tests, such as motif-finding algorithms and functional annotation, can later be applied to validate cluster membership for given objects. Although Euclidean distance metrics are usually employed in such analyses, Tavazoie et.al. note that this convention was arrived at arbitrarily (10).

The clustering procedure begins by random assignment of expression profiles to k clusters, followed by calculation of intra- and inter-cluster distances (5). An iterative process of shuffling objects between clusters and recalculating distances within and between clusters then ensues until the algorithm converges. This method has successfully been used by the Altman lab to distinguish between two types of lymphomas (6), and Tavazoie and colleagues. have shown that clusters of differentially expressed genes often participate in a common biological process (10).

Figure 2: Transcriptional profiles of two clinically distinct lymphomas by k-means clustering: (a) germinal cell subtype; (b) activated subtype (6). Note that a confidence score is listed to the right of each gene expression profile (see below).

A particularly nice feature of the SYSTAT 7.0 platform used by Tavazoie et. al. is that correlation coefficients are calculated for each expression profile within a cluster, allowing one to quantitatively evaluate how well the groups reflect the contributions of individual objects (10). Such statistical criteria can subsequently be used to order expression profiles within a cluster. This approach may produce clusters that are more stable than those produced solely from binary comparisons, e.g., the hierarchical clustering algorithm. Researchers may also address hypothesis-driven questions by ‘seeding’ clusters with expression profiles of interest, such as a molecular signature that is diagnostic of a particular cancer type, prior to running the k-means algorithm for a given expression matrix (5,6). In summary, k-means clustering is still a subjective process that is sensitive to the experimenter’s assumptions about expression profile diversity within the matrix, but represents an improvement over hierarchical clustering algorithms based on its computational rigor and the potential for the clustering process to be informed by biological knowledge.

Self-Organizing Maps

Another unsupervised clustering method that has been validated for large-scale data analysis is the generation of self-organizing maps (SOMS). This is a divisive clustering approach that redistributes a user-defined, two-dimensional set of nodes into n-dimensional gene expression space (5,9). Each node is assigned a reference vector, and a training set of randomly generated vectors is employed to optimize the algorithm for the chosen starting geometry, e.g., a rectangular or hexagonal array (5). Then, a single gene expression profile is selected at random, its closest node is identified, and the reference vectors in the nodal network are readjusted; this process is repeated until the algorithm converges. Node positions are readjusted as a function of proximity to the data point in question and the number of iterations that have been completed to that point; a weighting factor learned from the test set ensures that the closest nodes are moved more than distant nodes (5,9). Thus, each node comes to define a cluster of similar gene expression profiles, and adjacent clusters are likely to contain genes that have related expression patterns and/or kinetics.

Figure 4a: A graphical depiction of how a SOM is generated (9). Nodes (light gray circles) are initially positioned in a rectangular grid, but are subsequently redistributed in n-dimensional expression space to reflect the organization of related gene expression vectors (dark circles).

Figure 4b: Average expression profiles from neighboring nodes in an SOM; adjacent clusters (e.g., cluster 0 and cluster 1) are more similar than distant clusters (e.g., cluster 0 and cluster 4). Each error bars on each trace correspond to the distribution of individual gene expression profiles within the cluster. N corresponds to the number of genes in the cluster (9).

Tamayo et. al. conducted a proof-of-principle study demonstrating that SOMs can cluster microarray data in biologically meaningful ways (9). In addition, they provide a software package, GENECLUSTER, that displays the output in simple graphical form (9). The primary advantages of using SOMs are that the nodes are fit to the data according to a learned weight function, and therefore the positions of the nodes reflect the distribution of objects in expression space (5,9). Nodal organization in an SOM is therefore less arbitrary than the relationships that are derived from pairwise comparisions of similarity scores, such as hierarchical clustering methods. Therefore, adjacent nodes in an SOM are more closely related in profile composition than are two distantly situated nodes. Although use of this algorithm is not guaranteed to yield discrete clusters, changing the starting nodal geometry should, in theory, provide a starting point for addressing such problems (9). The use of SOMs in conjunction with hierarchical clustering or PCA may also give a reasonable estimate of the number of nodes required to effectively partition the dataset.

Supervised Clustering Methods

Unlike the unbiased clustering methods described above, supervised machine learning uses preexisting biological knowledge to classify microarray expression data. Often, neural networks or probabilistic models are implemented to ascertain the likelihood of membership in a group, e.g., a particular functional unit or cellular process, such as aerobic respiration or a tumor subtype. Membership criteria are learned from a user-defined training set that is assembled from citations in the literature, database searches, and experimental observations; the set must include both true positive and negative examples so that the algorithm's performance can be quantitatively evaluated and optimized (2,3,5). Although such classification schemes have the potential to act as exquisitely sensitive medical diagnostic tools, their utility for the *de novo* discovery of regulons is currently limited. Like HMM-based motif finding algorithms, supervised classification systems for microarray data may suffer from overfitting the training set data. If true, the positive predictive value of the algorithm would be high for previously identified gene expression signatures, but would potentially misclassify novel expression patterns that are nonetheless germane to the process or tissue type described in the true positive training set. Another implicit assumption in such binary classification schemes is that subtle differences in gene expression profiles must be discernible by the algorithm. (Presumably, this holds true for all of the methods described herein; see Figure 1).

It must be stressed that the efficacy of such classification schemes rests on the quality of the training set as well as the chosen algorithm, since false annotations can easily lead to misclassifications (3,5). The user must also verify whether there is explicit evidence for transcriptional regulation of the process in question, since other types of biological relationships cannot be captured by the learned methods described here (8). These issues may be a greater concern to systems biologists interested in constructing network of transcriptomes, as is further discussed below.

Support Vector Machines

Support vector machines (SVMs) are binary classifiers that categorize samples as either members or non-members with respect to a computationally-defined hyperplane in n-dimensional expression space (3,5). It is necessary to invoke comparisons in n-dimensions because it is often impossible to obtain effective separation of the data set in input space (3). Although several solutions that are consistent with the classification scheme may be possible in ‘feature space,’ the most stringent separation criteria are generally used to distinguish between members and outliers in experimentally-derived data sets (3). Effective separation is also contingent on the appropriate choice of two parameters: the kernel function and the magnitude of the penalty for violating the ‘soft margin,’ as explained below (3,5).

The kernel function is a flexible similarity metric that expresses relationships amongst gene expression vectors in feature space as dot products in input space; this obviates the need to specify the coordinates of gene expression vectors in higher dimensional space explicitly (3). Additionally, the complexity of the kernel function can be increased by raising the expression to a higher power: $K(X, Y) = (X \bullet Y + 1)^d$, where X and Y are gene expression vectors, and d is a positive integer (3). In essence, features for all d-fold interactions are taken into account amongst entries in input space, allowing the researcher to empirically derive which kernel function best separates the dataset. Another concern in classification schemes arises from the relative imbalance engendered by the number of true positives versus negatives in the dataset; this is particularly vexing if the magnitude of the noise in the negative dataset exceeds that of the positive signal (3). In such cases, incorrect false negative classifications are likely to confound data analysis. To counteract such implicit trends in the data, a soft margin and/or modified kernel function may be introduced. Soft margins allow true positive training examples to fall on the wrong side of the hyperplane (or surface), whereas the modified kernel function includes a diagonal element that corrects for the numerical imbalance between the number of positive and negative objects in the dataset (3,5).

Brown and colleagues evaluated the performance of four SVMs with kernel functions of increasing complexity and several standard machine learning approaches, including decision trees and Fisher’s linear discriminant, on previously published yeast microarray data (3). (The details of the latter two algorithms will not be discussed in detail). Briefly, each method was tested for its ability to correctly identify known members of a functional class, e.g. ribosomal proteins, from whole genome microarray datasets. Although SVMs with higher kernel functions (d=2 or 3) were superior to the other methods assayed, none of the algorithms correctly identified all known members for any of the functional groups tested. It should also be noted that the test dataset was composed of objects that were known to cluster into discrete expression classes *a priori*, and the authors do not speculate as to how well the best SVMs would perform on ‘unknown’ datasets. Further studies should be conducted to test the positive predictive value of SVMs for novel functional classification.

Many false positive and false negative designations reported in (3) resulted from discrepancies in database annotations versus mechanistic information gleaned from biochemical data. If classification strategies are to be an effective means of microarray data mining, a concerted effort should be made improve functional annotation strategies. Gene ontology databases should develop a controlled, yet expandable vocabulary that is uniformly applied within the research community, and classification schemes should include information about molecular events, such as reaction mechanisms or signal transduction cascades, that may be used to validate group membership. Additionally, the software packages introduced for classification analysis should have features that enable the user to manually curate the output so that the groupings accurately reflect our biological understanding of the process in question.

A logical extension of the binary classification methods mentioned here are relational classification strategies that group objects together on the basis of different criteria, such as functional annotation, expression level, and shared patterns of transcription factor binding sites in upstream regulatory regions. Such dynamic, query-specific methods promise to be powerful, and current efforts aimed at 'higher order' clustering methods of this nature are being applied to *S. cerevisiae* microarray datasets (e.g., Segal et. al. (8)). Methods that prove to be sufficient for yeast datasets may not be immediately applicable to microarray data derived from higher eukaryotes, which typically have larger intergenic regions and gene families, as well as the potential for more complex transcriptional regulatory mechanisms. Regardless of the model organism in question, however, it should be emphasized that even an exhaustive network of transcriptional regulators and their targets is not sufficient to explain any biological process in its entirety. Studies of translational and post-translational regulation are also needed to fully explain aspects of cellular physiology.

Segal et. al. built a versatile clustering program using probabilistic relational modeling, and have demonstrated its efficacy on microarray expression data from yeast (8). Their approach draws on the principles of Bayesian logic, and treats normalized expression ratios, transcription factor binding sites, and functional classifiers as equally weighted random variables in the algorithm. Unlike "two-way clustering" methods, which ignore differences amongst individual gene expression profiles in a cluster, PRM-based classifications are subjective groupings that reflect similarities that only exist in subsets of the array data; i.e., genes 1 and 2 may contain binding sites for a common transcription factor, but may exhibit different transcriptional responses under conditions X and Y (8). This dynamically shifting perspective should reveal novel relationships that are not evident from analysis of co-expressed genes alone. The structure of the PRMs described by Segal et. al. is reminiscent of the organizational strategy used to construct relational databases, e.g. the Altman lab's pharmacogenomics database (7), and thus this strategy may be most useful to those experimentalists who routinely analyze microarray data from a variety of sources.

In a PRM, a Bayesian network unites a set of defined variables, e.g., gene cluster, cellular function, expression level, each of which is associated with a conditional probability distribution (CPD (8)). The latter is a type of classification test that is used to determine the likelihood that the data point in question is a member of the group. Typically a rooted, bifurcating tree diagram is the preferred visual representation of the model; each parent (node) has two children (leaves) that correspond to the test outcomes (member/non-member (8)). Such trees are, of necessity, computationally complex. Interestingly, PRMs are reminiscent of the maximum likelihood methods used to infer evolutionary distance and to build phylogenies. Although Brown and colleagues report that the decision tree algorithm tested in their simulations performed poorly relative to SVMs, they note that parameters were not adjusted to optimize performance of the algorithm (3). Given the work of Segal et. al. and others who have provided proof-of-principle experiments to validate PRMs, it is likely that their versatility and capacity for discovering new relationships amongst differentially expressed genes will enable such platforms to effectively compete with SVMs in the future.

Concluding Remarks

Whole-genome transcriptional profiling technologies afford molecular biologists with a wealth of data, yet provide little information to researchers unless specific, testable hypotheses are coupled with computational savvy. Microarray data analysis aims to identify trends that are implicit in the data set, and to reorganize the data in ways that are simple, intuitive, and meaningful. It is important for molecular biologists to realize that microarray data management is a subjective process, despite the computational rigor inherent in many of the existing algorithms. As the output of each organizational method is different, it is preferable to run a dataset through several algorithms before conclusions about specific relationships are made. In addition, results obtained from microarray screens should always be validated and extended by performing traditional ‘wet bench’ experiments to demonstrate physiological relevance.

This paper is not intended to be an exhaustive overview of microarray clustering programs, but rather provides a critical introduction to commonly used and emerging methods that have produced biologically significant results. Both unsupervised and supervised classification methods are described; examples of the former include hierarchical clustering, k-means clustering, and SOMs, whereas SVMs and PRMs fall under the latter category. Hybrid techniques, e.g. k-means clustering using ‘seed’ vectors, may ultimately prove to be the most useful way of conducting data analysis in certain hypothesis-driven experiments. The utility of unsupervised clustering methods stems from their use for discovery of novel regulon and other relationships *de novo*, whereas classification schemes are perhaps more useful as clinical diagnostic agents. Although the signal to noise ratio must be greater for unsupervised versus supervised methods, the positive predictive value of the unstructured methods is often greater (1).

Of the unsupervised methods described here, hierarchical clustering represents the most ‘primitive’ algorithm, albeit the most commonly used method of microarray data management (4,5). The integrity of the clusters produced using the hierarchical algorithm is dependent on the distance metric used; statistical methods are preferable to Euclidean measures, and weighted, UPGMA-derived clusters are generally preferred over minimal or maximal neighbor joining methods. Hierarchical clustering can provide an informed estimate of the number of discrete clusters in a dataset, and can be used to visualize how an object’s expression profile relates to that of other co-expressed genes within a cluster. Such analyses can then be extended using an approach that is computationally rigorous, such as k-means clustering or generation of SOMs. SOMs are inherently attractive because nodes map themselves to the data, whereas the data is partitioned amongst clusters in the k-means approach. SOMs, however, are not guaranteed to yield discrete, robust clusters. k-means clustering can also be used in conjunction with supervised methods, or as a diagnostic tool.

Classification tools for mining microarray data, such as SVMs, seem to afford little capability for exploratory studies, and rather are useful for diagnostic purposes. PRMs may greatly improve the predictive value of supervised methods, since clusters can be organized according to different attributes specified in the skeleton, or structure, of the model. The greatest challenge facing developers of PRMs will be generating software packages that perform well on datasets derived from higher eukaryotes, as well as yeast. Experimentalists also have an obligation to push the development of chip technologies further, which will do much to address common concerns about conclusions derived from microarray experiments. For example, the issue of whether upregulated genes are direct or secondary targets of a given transcription factor can be addressed by a search for the relevant transcription factor binding sites in upstream regulatory sequence, or by performing a chIP chip experiment using an epitope-tagged version of the transcription factor as bait in the pull-down reactions. Cross-disciplinary training and collaboration between molecular biologists, computer scientists, and statisticians will ensure that progress continues to be made in understanding cellular physiology and disease etiology at a global level.

References

1. Altman, Russ. Microarray Data Analysis: Clustering and Classification Methods. BMI 214/CS274, Lecture notes 4/18/02.

2. Brazma, Alvis and Jaak Vilo. Gene Expression Data Analysis (2000). *FEBS Letters* 480;
17-24.

3. Brown, Michael P.S. et. al. Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines (2000). *PNAS USA* 97(1); 262-267.

4. Eisen, Michael B. et.al. Cluster Analysis and Display of Genome-Wide Expression Patterns (1998). *PNAS USA* 95; 14863-14868.

5.Quackenbush, John. Computational Analysis of Microarray Data (2001). *Nature Reviews*
Genetics 2; 418-427.

6. Raychaudhuri, Soumya et.al. Basic Microarray Analysis: Grouping and Feature Reduction (2001). *TRENDS in Biotechnology* 19(5); 189-193.

7. Rubin, DL et. al. Representing Sequence Data for Pharmacogenomics: An Evolutionary
Approach Using Ontological and Relational Models (2002). *Bioinformatics* Jul;18 Suppl 1:S207-15.

8. Segal, Eran et.al. Rich Probabilistic Models for Gene Expression (2001). *Bioinformatics*
17, 243S-252S.

9. Tamayo, Pablo et.al. Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation (1998).

PNAS USA 96; 2907-2912.

10. Tavazoie, Saeed et. al. Systematic Determination of Genetic Network Architecture (1999). *Nature Genetics* 22; 281-285.