EXPLORING GENE EXPRESSION DATA WITH CLASS SCORES

PAUL PAVLIDIS
Columbia Genome Center, Columbia University, pp175@columbia.edu

DARRIN P. LEWIS and WILLIAM STAFFORD NOBLE[a]
Department of Computer Science, Columbia University,
{dplewis,noble}@cs.columbia.edu

To appear in *Proceedings of the Pacific Symposium on Biocomputing*, 2001.

Abstract

We address a commonly asked question about gene expression data sets: "What functional classes of genes are most interesting in the data?" In the methods we present, expression data is partitioned into classes based on existing annotation schemes. Each class is then given three separately derived "interest" scores. The first score is based on an assessment of the statistical significance of gene expression changes experienced by members of the class, in the context of the experimental design. The second is based on the co-expression of genes in the class. The third score is based on the learnability of the classification. We show that all three methods reveal significant classes in each of three different gene expression data sets. Many classes are identified by one method but not the others, indicating that the methods are complementary. The classes identified are in many cases of clear relevance to the experiment. Our results suggest that these class scoring methods are useful tools for exploring gene expression data.

1 Introduction

Researchers interested in discovering meaningful patterns in gene expression data often ask, "What functional categories of genes are most interesting in the data?" This question is usually answered by indirect means, by making use of two fundamentally different general methodologies: "supervised" and "unsupervised."[1] [b] In this paper, we describe an analysis method which we term "semi-supervised," which combines elements of supervised and unsupervised approaches using existing classifications. This approach attempts to circum-

[a] Formerly William Noble Grundy: see www.cs.columbia.edu/~noble/name-change.html

[b] The supervised and unsupervised methodologies can be used to seek information about the genes on the arrays, or the samples from which the RNA was extracted; here we focus on the analysis of genes.

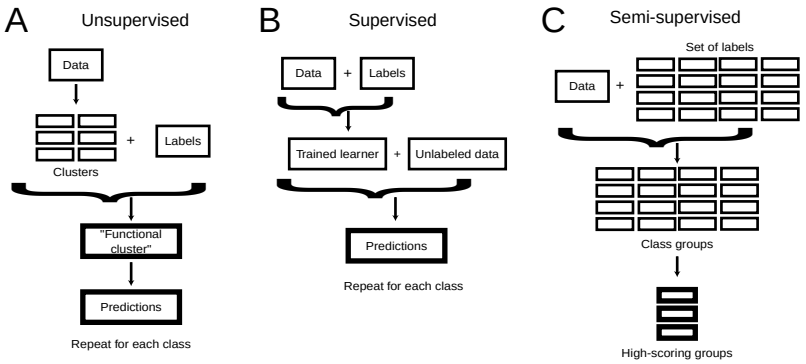

Figure 1: **Schematics of three general methodologies for analyzing gene expression data.** Boxes represent sets of genes or descriptions of gene classifications ('labels'), and arrows represent transitions between analysis stages. Thick outlines indicate the main outputs of each method. **A. Unsupervised.** The data is divided into clusters based on profile similarity. A *post hoc* analysis using classification labels can be used to identify "functional clusters" which contain many genes from the same class. Unannotated genes in these classes are predicted to have a related function. This *post hoc* analysis can then be repeated for multiple classes. **B. Supervised.** Data together with classification labels are used to train a learning algorithm, which can then be used to make predictions about unannotated genes. The learner must be retrained to recognized each class. **C. Semi-supervised.** Data together with a constellation of classification labels are used simultaneously to partition the data into class groups, based entirely on the labels. Scoring methods (the topic of this paper) are then used to identify groups that have particular characteristics.

vent some of the limitations of the supervised and unsupervised methods, and is designed to directly identify "interesting" gene classes.

An important concept for our discussion is that of a "gene class." We define a gene class as a group of genes with related functions, or which are otherwise grouped together based on biologically relevant information. For example, a class could represent a signal transduction or metabolic pathway, the members of a protein complex, or an enzymatic activity. A gene can be a member of any number of classes, and hundreds if not thousands of gene classes can be defined [2]. The goal of computational analysis is often to identify new class members, but here we are primarily concerned with making direct use of the existing gene classifications.

Before describing semi-supervised gene expression analysis, it is useful to describe the approaches from which it is derived. The three general methods (supervised, unsupervised, and semi-supervised) are depicted schematically in Figure 1. The unsupervised approach is perhaps the most familiar to gene expression researchers, who often use clustering algorithms to identify genes with

similar expression patterns.[3,4] Clustering is unsupervised because the only input is the expression data, without any additional use of prior knowledge about the genes (Figure 1A). Genes with similar expression patterns are grouped together by clustering, without any knowledge of the genes' functions. Using clustering as a functional genomics technique thus requires *post hoc* interpretation of clusters in terms of the functions of the genes in the clusters. For example, if in a given cluster many genes are found to be in the same class, the experimenter might hypothesize that other genes in the cluster have related functions, and that the function is relevant in some way to the biological process under investigation. Another way to use unsupervised methods is for class discovery: genes which cluster together are hypothesized to have some functional or regulatory relationship. A strength of the cluster-driven approach is that it encourages an exploratory approach to the data, but it does not automatically generate hypotheses about the functions of the genes in the clusters.

In constrast, prior knowledge about genes is directly exploited by supervised methods, such as support vector machine classification [5,6] (Figure 1B). This type of algorithm is often referred to as a "learner." The learner is trained by a "teacher" to identify a particular gene class. The learner can predict the classifications of previously unanotated genes.[5] Thus, the supervised approach can be used to do the same job as the unsupervised method by complementary means. Supervised methods can yield superior performance in grouping genes of particular functions together,[5] but require identification of the class of interest ahead of time.

The semi-supervised approach is intermediate between the supervised and unsupervised approaches (Figure 1C). A score is assigned to each of a large number of predefined gene classes, and classes with high scores are considered potentially more interesting than classes with low scores. Thus in the semi-supervised approach, a large collection of teachers is available, but only some of the teachers provide "true" classifications. The goal of the learner in this case is to select the true classes from among this large collection of candidate classes. In comparison, in unsupervised learning, there is no teacher to provide the true classifications, while in supervised learning there is only a single teacher from which to learn the classifications.

To implement semi-supervised learning, we consider three methods for scoring classes: the tendency of genes in the class to be co-expressed, the significance of the expression profiles in the context of the experimental design, and the learnability of the gene class. The first method measures how well the genes in the class cluster together, that is, how similar their expression profiles are. The method we apply uses the average pairwise correlation between the expression profiles in a class [7]. Although this co-expression measure is a powerful means for scoring classes, profile similarity alone is too limiting as a

metric for class importance. This is because while it may sometimes be true that genes which cluster together have related functions, it is certainly not always the case that genes with related functions cluster together.

The second scoring method measures the statistical significance of the expression pattern of each gene with respect to the experimental design. Using statistical methods such as analysis of variance (ANOVA), each gene can be assigned a p-value corresponding to the probability that the variations in gene expression across the conditions could have been observed by chance. Such analysis of each gene is commonly conducted in expression studies to assess which genes changed expression level during the experiment. Although such scores cannot be used as a means of identifying new members of a class, or in class discovery, we show here that the scores for the genes in a class can be meaningfully combined to provide a score for a class as a whole.

The third scoring method measures the learnability of a candidate gene class. The particular score we use here is a p-value derived from the total hold-one-out cross-validated error rate of a k-nearest neighbor classifier. This metric measures the distinctness of genes within the class relative to genes outside of the class.

Some previous work suggests that the semi-supervised approach is likely to be fruitful. Gerstein and Jansen (2000) have shown how classes can be ranked by coexpression,[7] while Hakak *et al.* [8] used the statistical signficance of individual genes to assess the significance of one class of interest. Mirnics *et al.*[9] compared the distributions of expression ratios in gene classes to that of the bulk data. Zien et al.[10] report the use of "conspicuousness" (related to the statistical significance approach we describe) and "synchrony" (essentially the same as expression pattern similarity) alone and in combination, as a means of identifying biologically relevant biochemical pathways among sets of hypothetical pathways. Ben-Dor *et al.*[11] discuss tissue classification and class discovery based on "surprise scores" that are similar to the statistical measures we describe here. However, the use of these methods as a general means for identifying gene classes of interest in a data set does not appear to have been fully explored.

In this paper, we apply the similarity-based, statistical-significance-based, and learnability-based methods to three previously published gene expression data sets, using publically-available gene classifications. In all three cases we show how to calculate p-values that can be used to accurately assess the significance of a particular class score. All three types of scores identify interesting classes of genes in all three data sets. Importantly, we show that the methods are to a large extent complementary, with each giving high scores to classes that the other does not.

Data set	Type	Arrays	Cond.	Genes	Classes	Reference
Yeast	Spot	79	79	2465	145 (MIPS)	Eisen *et al.*, 1998
Brain	Affy	24	6×2	5552	581 (GO)	Sandberg *et al.*, 2000
Cancer	Affy	38	3	5092	397 (GO)	Golub *et al.*, 1999

Table 1: **Summary of the three gene expression data sets.** The type of array, either spotted cDNA[14] (spot) or Affymetrix olignucleotide[15] (Affy), is listed, together with the number of arrays, conditions (Cond), genes and classes present. Genes were counted only if they were a member of at least one class. In the brain data, six brain regions are examined in two mouse strains. In the "classes" columns, MIPS and GO refer to the classification scheme (MIPS functional catalog,[16] or Gene Ontology,[2] respectively).

2 Methods

We used three publically available gene expression data sets to evaluate our methods. The data sets were chosen to represent a range of situations where the methods we describe might be useful. The first ("yeast") is from Eisen *et al.*;[3] and consists of 79 experiments in a variety of conditions. The conditions include different time points during the cell cycle and during the responses to various stresses (heat, cold, etc.). There is only one array per condition. The second ("brain") is from the work of Sandberg *et al.*,[12] and consists of replicate analysis of six brain regions in two mouse strains, for a total of 24 arrays. The last ("cancer") is from the work of Golub et al.,[13] who performed microarray analysis of acute leukemias. Each sample is from an individual patient, and was identified by Golub *et al.* as belonging to one of three groups of tumor type. We used the "training" data set from their work. The data sets are summarized in Table 1.

Our classification schemes were drawn from publically available databases. For the yeast data, we used the MIPS functional catalog[16] (www.mips.biochem.mpg.de). For the brain data and the cancer data, we used the publically available Gene Ontology[2] classifications (www.geneontology.org). Both schemes are hierarchical; that is, they consist of nested descriptions of genes that increase in detail as one descends down the hierarchy. Thus we expect a certain amount of redundancy in our results, as similar classes will receive similar scores. While we have not attempted to address this replication issue directly, we did find that restricting the classes to a particular size range to be useful for reducing the complexity of the results. Thus we limited our analysis to classes that had between 5 and 200 members. The number of classes meeting our criteria for each data set are listed in Table 1.

In our first scoring approach, we wished to calculate a measure that represents how coregulated the genes in a class are. The measure we used is the average of the Pearson correlation coefficient for the pairwise comparisons of

genes in the class, omitting comparisons of genes to themselves.[7] If the expression vectors for the genes in a class are correlated, then the average correlation between the genes will be high.

Some genes (or more precisely, UniGene clusters) were represented more than once on the Affymetrix arrays, and this replication can skew the average score for a class. To deal with this issue, we gave each member of a set of n replicates a weight of $1/n$ in calculation of the average, and comparisons between replicates were not included in the average correlation. We note that this correction is crude; not all replicates are equivalent because the various "replicates" can come from different sequences representing different splice variants, or probe sets which are of varying sensitivity, and thus should not truly be considered replicates. We apply this simple correction to ameliorate the problems caused by giving the replicates the same weight as unreplicated genes, but leave a more thorough treatment as a topic for further study.

To convert the raw average corrrelations into p-values, the background distribution of scores expected under the null hypothesis was determined empirically by generating scores for $500,000$ randomly selected sets of genes. Separate distributions were calculated for each class size for each expression data set. For small classes this distribution is quite broad, while for large classes it is narrower (not shown). The p-value for a class was then calculated as the fraction of simulated classes of the same size which had higher scores than the real class. The smallest p-value we could thus directly measure is thus $1/500000$ (2×10^{-6}). Classes with p-values less than this value were provisionally set to 1×10^{-6}. This p-value is the "correlation score" for a class, and is calculated for all classes.

Our second measure applies statistical measures of significance of the expression pattern with respect to the experimental design. For the brain and cancer data sets, we used ANOVA[17] to obtain a separate significance score for each gene, in the form of a p-value. ANOVA is a standard statistical method for testing hypotheses about multiple means. In this context, genes with low p-values show more significant changes in expression between groups. For the brain data, we focused on genes showing differences among the six brain regions in two mouse strains in a two-way ANOVA, while for the cancer data we generated p-values for differences among the three tumor types (ALL-Bcell, ALL-Tcell and AML) in a one-way ANOVA. We used the $-\log_{10}(p\text{-value})$ as the score for each gene in our subsequent calculations. For the yeast data, which had no replication across the 79 conditions, we used the standard deviation of the expression Zien *et al.*[10]

The average of the log transformed p-values for the genes in a class forms the basis of the class score. This summation is equivalent to calculating the joint probability of the genes in the class under the null hypothesis, assum-

ing independence of the genes (an assumption which is certainly not correct, but our results indicate that this simplification is acceptable). These average values can be converted to p-values in a manner identical to that we used for the correlation scores, by calculating the average $-\log_{10}(p\text{-value})$ for $500,000$ randomly chosen sets of genes to generate a background distribution, with a separate distribution calculated for each class size. To deal with replicated genes, we used a $1/n$ weighting scheme analogous to that described for the classification scores. Because these scores take the experimental design into consideration, we refer to it as the "Experiment score."

The third method we tested measures the learnability of the class by a simple supervised learning algorithm, yielding a "learnability score." In order for a class to be learnable, the genes must not only cluster together in space (i.e., be co-expressed), but also be sufficiently distinct from other genes in the data set to be distinguishable as a class. The degree to which this is possible using the k-nearest-neighbor (KNN) algorithm forms the basis of our third method. The KNN classifier predicts the label of an unclassified example as the label belonging to the majority of the k closest examples in Euclidean space. Because KNN is unique among supervised learning algorithms in that there is no training step, we can efficiently compute hold-one-out cross-validation error rates. These rates form the basis for the scoring scheme. In this work we set k to one. The use of a different learning algorithm might yield different results than those we report here.

To convert these raw scores into p-values, the null distribution can be calculated analytically, instead of empirically as for the correlation and experiment scores. The calculation is based on the observation that, for randomly labeled data, the probability of KNN misclassifying a randomly selected data point X depends only on the size $\mathcal{P}$ of the gene class and the size D of the entire data set. Say that example X belongs to the positive class $\mathcal{P}$. To encounter an error on this example, KNN must place X into the negative class $\mathcal{N}$, which can only occur if fewer than $\lfloor \frac{k}{2} \rfloor$ of the next k points, chosen at random, have labels $\mathcal{P}$. This outcome is expressed in the following conditional probability: $\mathbf{Pr}(X_{\mathcal{N}}|X \in \mathcal{P}) = \left(\sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} \binom{P-1}{i}\binom{D-P}{k-i}\right) / \binom{D-1}{k}$, where $X_{\mathcal{N}}$ denotes example X being classified in class $\mathcal{N}$ by KNN. This probability, along with prior probabilities derived from the class sizes, yields the overall probability of a false positive or false negative misclassification $\mathbf{Pr}(X_{\mathcal{N}}|X \in \mathcal{P})\mathbf{Pr}(X \in \mathcal{P}) + \mathbf{Pr}(X_{\mathcal{P}}|X \in \mathcal{N})\mathbf{Pr}(X \in \mathcal{N})$, which can be used to compute a binomial cumulative distribution. In this way, a p-value can be obtained for any KNN cross-validated total error.

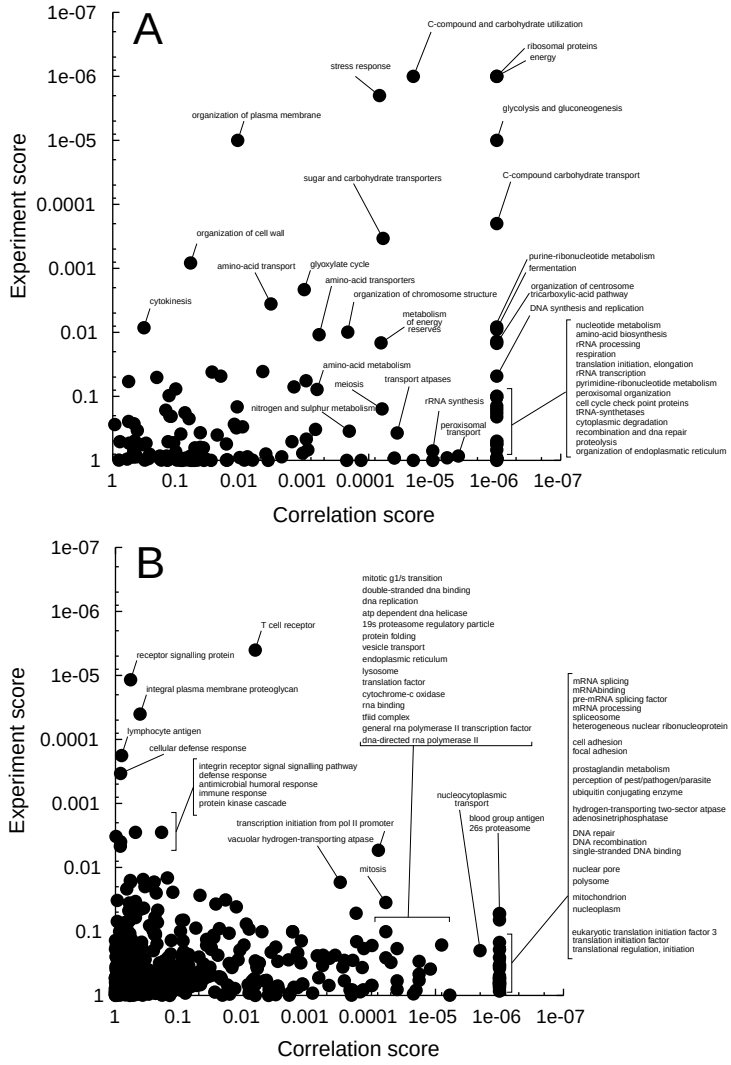

Figure 2: **Summary of "experiment" and "correlation" results.** See next page for legend.

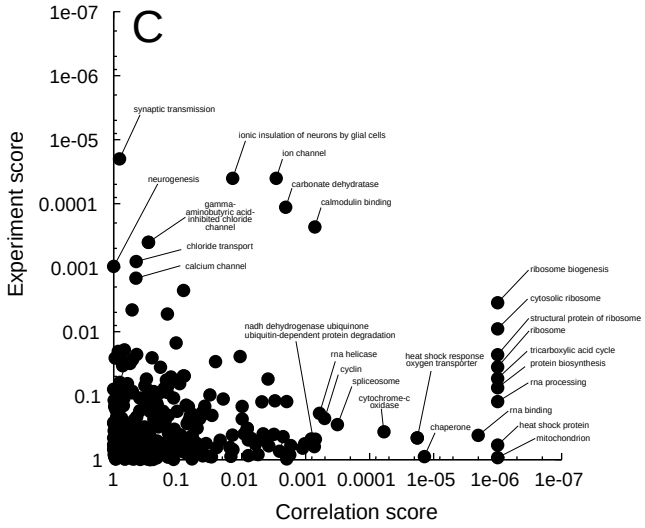

Figure 2: **(Continued) Summary of "experiment" and "correlation" results. A.** Yeast data. **B.** Cancer data. **C.** Brain data. In all three panels, each point represents a single gene class. The correlation score is plotted on the horizontal axis while the experiment score is plotted on the vertical axis. Text labels indicate the identities of some individual high-scoring classes and groups of classes.

Class	p-value
Brain	
histogenesis and organogenesis	3.861^{-4}
Cancer	
structural protein of ribosome	1.969^{-23}
protein biosynthesis	5.160^{-7}
RNA binding	1.158^{-6}
cell motility	1.876^{-4}
Yeast	
Transport facilitation	1.550^{-10}
lipid fatty-acid and isoprenoid biosynthesis	2.682^{-4}
lipid fatty-acid and isoprenoid metabolism	3.869^{-4}
glycolysis and gluconeogenesis	3.058^{-4}
sugar and carbohydrate transporters	3.816^{-4}
tricarboxylic-acid pathway	5.00^{-4}
rRNA processing	5.807^{-4}

Table 2: **Summary of the KNN results.** Only classes which were not given p-values less than 10^{-3} by another method are listed. Closely related classes are indented. The total number of significant classes identified by this method were: Brain: 1; Yeast: 22; Cancer: 4. Most of these were identified by the correlcation score method.

3 Results

We measured correlation, experiment, and learnability scores for the yeast, cancer, and brain data sets. The results for the correlation and experiment scores are summarized in Figure 2. The learnability method yielded fewer significant classes than the other methods, so its results are summarized in Table 2.

As predicted (see Methods), due to the hierarchical nature of the classifications some of the high-scoring classes shown in Figure 2 are closely related to each other. For example, in Figure 2B, multiple classes closely corresponding to mRNA splicing factors (mRNA splicing, spliceosome, etc.) are given high correlation scores. This redundancy makes it somewhat difficult to make an accurate count of how many classes are given high scores. However, some important trends are discerned by inspecting the data.

Of the three methods, the learnability measure yielded the fewest "interesting" classes. However, some of the classes it identifies are different than the ones identified by the other methods (Table 2). Thus it forms a useful complement to the other two methods, and has in addition the advantage of computational speed.

We observed that several classes consisting of "housekeeping" genes, such as the ribosomal proteins and "RNA processing," are given high correlation scores in all three data sets, but not necessarily high experiment scores. The appearance of these classes in three disparate data sets suggests that such housekeeping genes show a very high coordination of expression levels that is not dependent on the experimental context.

In contrast to the correlation scores, high experiment scores tended to be given to classes that are highly specific to the experimental design. For example, the highest experiment-scoring class in the cancer data (Figure 2B) is "T-cell receptor," which is appropriate considering that the tumors studied fell into groups depending on whether they were derived from T-cells or B-cells.[13] Similarly, the highest scoring classes in the brain data set were "synaptic transmission," "ion channels" and "ionic insulation of neurons by glial cells," all of which might be relevant to functional differences among the brain regions studied.[12] In the yeast data set, fewer classes received high experiment scores without also receiving high correlation or learnability scores. The major exceptions are "organization of plasma membrane" and possibly "stress response." The former class consists primarily of permeases for sugars and other small molecules. The latter class consists of genes that change expression level in response to stress. These classes are relevant because the yeast data was gathered during various stressful conditions and metabolic states.[3]

4 Discussion

Our contributions in this work are three-fold. First, we provided an explicit description of the class scoring problem, formulating it as intermediate between supervised and unsupervised approaches. Second, we described three methods for semi-supervised analysis, which capture different features of the data. Finally, we demonstrated the use of these methods on real data, showing they reveal interesting biologically relevant features of the data. In our view, one of the chief appeals of the semi-supervised method is that it uses prior knowledge in ways that unsupervised methods cannot, while maintaining a flexibility that supervised methods lack.

Interestingly, the three methods we used often give different classes high scores; that is to say, they are complementary in the kinds of information they provide. This result is particularly apparent in the comparison of experiment scores to correlation scores for the cancer and brain data sets. The learnability score yields only a small number of additional high-scoring classes. Of the three methods, the experiment scores appear to be the most specific for each data set, while the correlation scores, and to some extent learnability scores, tended to focus on "housekeeping" classes. It will be interesting to see if this trend is evident as we examine additional data sets.

There are several issues we encountered during our experiments that suggest avenues for future research and improvements to the methods. Most obviously, we are at the mercy of the existing annotations. A major reason for this limitation is the current incompleteness of annotations based on the Gene Ontology. Thus our methods should prove to be even more useful as database annotations improve. Because some classes are redundant, for our purposes some simplification of the classification schemes would also be desirable.

Another issue is our assumption, for the experiment-score analysis, that the ANOVA p-values for each gene are independent. This is clearly not the case. At one extreme, some genes are represented more than once in a data set. In general, the correlation structure of the data will affect the statistical significance of a given gene pattern. There are many methods for correcting p-values in such a situation,[18] but we have not attempted to apply them here and leave this as an issue for future study.

A final issue is the requirement for a computationally intensive determination of the background distribution of experiment and correlation scores. It is possible that this computation can be avoided by estimating the distributions.[7] Even if such estimates are not exact, they are likely to provide a reasonable calibration of the scores for the effect of class size. We note that we can also probably afford to sacrifice some precision in p-value computation, because as long as the method provides guidance through the hundreds of gene classes,

we consider it a success.

Acknowledgments

This work was supported by an Award in Bioinformatics from the PhRMA Foundation, and by National Science Foundation grants DBI-0078523 and ISI-0093302.

5 References

1. A. Brazma and J. Vilo. *FEBS Letters*, 23893:1–8, 2000.
2. M. Ashburner, C. A. Ball, et al. *Nat Genet*, 25(1):25–9., 2000.
3. M. B. Eisen, P. T. Spellman, et al. *Proc Natl Acad Sci U S A*, 95(25):14863–8., 1998.
4. P. Tamayo, D. Slonim, et al. *Proc Natl Acad Sci U S A*, 96:2907–2912, 1999.
5. M. P. Brown, W. N. Grundy, et al. *Proc Natl Acad Sci U S A*, 97(1):262–7., 2000.
6. T. R. Hvidsten, J. Komorowski, et al. In *Proceedings of the Pacific Symposium on Biocomputing*, pages
7. M. Gerstein and R. Jansen. *Curr Opin Struct Biol*, 10(5):574–84., 2000.
8. Y. Hakak, J. R. Walker, et al. *Proc Natl Acad Sci U S A*, 98(8):4746–51., 2001.
9. K. Mirnics, F.A. Middleton, et al. *Neuron*, 28:53–67, 2000.
10. A. Zien, R. Kuffner, et al. In *Proc Int Conf Intell Syst Mol Biol*, pages 407–417, 2000.
11. A. Ben-Dor, N. Friedman, et al. In *Proc Int Conf Intell Syst Mol Biol*, pages 31–38, 2000.
12. R. Sandberg, R. Yasuda, et al. *Proc Natl Acad Sci U S A*, 97(20):11038–43., 2000.
13. T. R. Golub, D. K. Slonim, et al. *Science*, 286(5439):531–7., 1999.
14. J. DeRisi, L. Penland, et al. *Nat Genet*, 14(4):457–60., 1996.
15. D. J. Lockhart, H. Dong, et al. *Nat Biotechnol*, 14(13):1675–80., 1996.
16. H. W. Mewes, D. Frishman, et al. *Nucleic Acids Res*, 28(1):37–40., 2000.
17. J. H. Zar. Prentice Hall, 1998.
18. P.H. Westfall and S.S. Young. John Wiley & Sons, Inc., New York, 1993.