Extracting Conserved Gene Expression Motifs from Gene Expression Data

T.M. Murali, S. Kasif

Pacific Symposium on Biocomputing 8:77-88(2003)

EXTRACTING CONSERVED GENE EXPRESSION MOTIFS FROM GENE EXPRESSION DATA

http://genomics10.bu.edu/murali/xmotif

T. M. MURALI
Bioinformatics Program, 48 Cummington St., Boston University, Boston MA 02152

SIMON KASIF
Bioinformatics Program and Department of Biomedical Engineering, 48 Cummington St., Boston University, Boston MA 02152

Abstract

We propose a representation for gene expression data called conserved gene expression motifs or xMOTIFs. A gene's expression level is conserved across a set of samples if the gene is expressed with the same abundance in all the samples. A *conserved gene expression motif* is a subset of genes that is simultaneously conserved across a subset of samples. We present a computational technique to discover large conserved gene motifs that cover all the samples and classes in the data. When applied to published data sets representing different cancers or disease outcomes, our algorithm constructs xMOTIFs that distinguish between the various classes.

1 Introduction

Gene expression plays an important role in cell differentiation, development, and pathological behaviour. DNA microarrays[1,2] offer biologists the remarkable ability to monitor the expression levels of thousands of genes in a cell simultaneously. High-throughput gene expression analysis promises to produce new insights into cell function as well as stimulate the development of new therapies and diagnostics.

When a gene's expression level is measured across a variety of samples, the expression values usually span a wide range. Biologically, these values correspond to a small number of distinct states that the gene is in, e.g., up-regulated or down-regulated. Since the up-regulation of a gene is a temporal process, it is often difficult to determine a gene's state based on a set of noisy expression values. However, on average, it might be possible to differentiate the expression level of an up-regulated gene in one tissue type (e.g., a cancer tissue) from its level in another type of tissue (e.g., a healthy tissue) where the gene is not expressed with the same abundance.

Motivated by this observation, we say that a gene's expression level is *conserved* across a subset of samples if the gene is in the same state in each of the samples in this subset. A *conserved gene expression motif* or xMOTIF is a subset of genes whose expression is simultaneously conserved for a subset of samples; we say that each of these samples *matches* the motif. In this paper, we use a range of expression values to represent a gene's state. If we map each gene to a dimension, each sample to a point, and each expression value to a coordinate value, an xMOTIF is identical to a multi-dimensional hyper-rectangle that is bounded in the dimensions corresponding to the conserved genes in the motif and unbounded in the other dimensions. See Figure 1 for examples. Mathematically, the expression values of a sample that matches a motif satisfy a conjunction of inequalities, two for each gene in the motif.

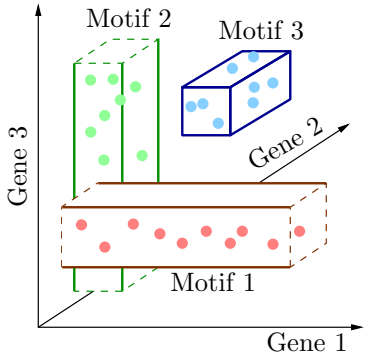

Figure 1: Example of xMOTIFs. Genes 2 and 3 are conserved in Motif 1, Genes 1 and 2 are conserved in Motif 2, and all three genes are conserved in Motif 3. A dashed face of a box indicates the dimension along which the box is unbounded.

In this work, we address the task of identifying xMOTIFs in gene expression data. We concentrate on data sets where each sample belongs to a particular class, e.g., different types of cancer,[3] cancerous and healthy tissues,[4] and patients with different survival rates.[5]

We believe that this work is the one of the first to suggest representing gene expression data concisely in the form of xMOTIFs. Such a representation has several potential biological advantages and applications. First, by comparing and contrasting the gene motifs for different classes, we can identify genes that are conserved in multiple classes but are in different states in different classes. If the classes correspond to different diseases or to diseased and normal tissues, such genes are possible drug targets. Second, if the genes in a motif are believed to interact in a pathway, the information present in the motif about which genes

are highly expressed and which are suppressed could be used to deduce and refine the structure of the pathway. Third, by requiring that multiple genes be simultaneously conserved across the samples matching a motif, we might be able to characterise sub-classes in the data that no gene on its own provides high-quality evidence for.

Before attempting to develop an algorithm for computing xMOTIFs, it is useful to consider the properties that an xMOTIF should have. Computing one motif for each sample makes the representation over-specific. Therefore, we desire that each motif for a class should be matched by a large fraction of the samples in that class, if not all the samples in that class. Each motif should contain as many conserved genes as possible. While a motif that contains one or two genes is biologically feasible, it may not be statistically significant since such a motif could appear with high probability in randomly-generated data. On the other hand, a motif that contains too many genes may be too restrictive since no sample may match the motif. Motivated by these observations, we propose the following formal definition of an xMOTIF:

Definition 1.1 *Given a set of genes whose expression levels are measured across a set of samples and user-defined parameters* $0 < \alpha, \beta < 1$*, a* conserved gene expression motif *or* xMOTIF *is a pair* (C, G)*, where* C *is a subset of the samples and* G *is a subset of the genes, that satisfies the following conditions:*

Size: *the number of samples in* C *is at least an* α*-fraction of all the samples,*

Conservation: *every gene in* G *is conserved across all the samples in* C*, i.e., the gene is in the same state in all the samples in* C*, and*

Maximality: *for every gene not in* G*, the gene is conserved in at most a* β*-fraction of the samples in* C*.*

The maximality condition enforces a balance between the number of genes in the motif and the number of samples matching the motif. If we add a gene to the motif, then the number of samples matching the new motif will decrease by a fraction of at least β, a cost we may not be willing to pay.

Given this definition, the gene expression data may contain many xMOTIFs. Among all xMOTIFs, we are interested in the largest xMOTIF, the one that contains the maximum number of conserved genes. In order to cover all the classes completely using xMOTIFs, we adopt the following iterative algorithm: find the largest xMOTIF in the data, remove the samples that satisfy this motif from the data, find the largest motif in the remaining data, and continue in this manner until all samples satisfy some motif.

Our approach has several desirable features. (i) We allow a gene to appear in more than one motif and in motifs representing different classes, modelling

the possibility that the gene's expression level may be regulated in multiple conditions. (ii) By not deleting the samples that match a computed xMOTIF, we can allow samples to appear in different motifs. This property may be useful when a sample belongs to multiple classes or when we are interested in discovering new classifications of the samples. (iii) The system need not be told beforehand how many motifs to compute.[a] (iv) Using this approach, we can find xMOTIFs with vastly different numbers of genes; we are likely to discover motifs with many genes in earlier iterations and motifs with fewer genes in later iterations.

Our definition of an xMOTIF is based on the notion of "projective cluster" developed by Procopiuc et al. [6] in the context of problems in computer databases and computer vision. It can be shown that the problem of computing the largest xMOTIF is NP-complete by transforming the problem of computing the maximum-edge bipartite clique in a bipartite graph[7] to the motif computation problem. In Section 4, we present a probabilistic algorithm that exploits the mathematical structure of xMOTIFs to compute the largest xMOTIF efficiently. This algorithm extends the technique developed by Procopiuc et al. to compute projective clusters.

2 Previous Research

Previous work on the computational analysis of gene expression data falls into two broad categories. When the samples belong to distinct classes, researchers have used well-known techniques such as nearest-neighbour rules, support vector machines and feature selection to build a diagnostic tool that can distinguish between the various classes based on the expression profiles of predictive genes.[3,8] When the goal is to analyse the data in an unsupervised manner, techniques such as k-means clustering and hierarchical clustering[9] are common. See the paper by Tibshirani et al. [10] for a survey of such clustering methods. Unfortunately, it appears difficult to modify techniques like k-means clustering or hierarchical clustering to compute xMOTIFs since these techniques attempt to cluster in the space spanned by all the genes. Bayes networks[11] and graph-based clustering algorithms[12] have also been used to model and analyse gene expression data.

Recently, some researchers have developed techniques that simultaneously cluster genes and samples, an idea that seems closest to xMOTIFs. Since this approach to analysing gene expression data is relatively new, different papers use different criteria for simultaneously clustering genes and samples. As a

[a] While the user provides the parameter α, we may compute far less than $1/\alpha$ xMOTIFs, since α is a lower bound on the number of samples that match the largest xMOTIF.

result, it is difficult to compare these techniques. Here, we attempt to characterise in what terms our approach is different from other similar ideas.

Hartigan's block clustering algorithm[13,10] repeatedly rearranges the rows and columns of the matrix of gene expression values so that the rearranged matrix contains several disjoint blocks (contiguous sub-matrices) of highly correlated values. This technique requires that a gene be present only in one cluster, a condition that may not always be useful biologically (consider a situation when a gene's expression is regulated in different diseases). Cheng and Church adopt an approach called "biclustering"[14]. They compute sub-matrices or biclusters that have small "mean squared residue," a measure of the variance in the sub-matrix. Tanay, Sharon, and Shamir[15] adopt a graph-theoretic approach to biclustering. They represent gene expression data as a bipartite graph, whose nodes and edges are assigned weights under a graph model that they define. In this framework, a bicluster is a dense bipartite subgraph of the original graph. They develop algorithms to compute biclusters with large weights. Getz, Levine, and Domany apply a technique called "coupled two-way clustering."[16] They repeatedly apply a hierarchical clustering algorithm to different subsets of genes and samples. The sub-matrices they output correspond to stable clusters generated by this process.

All these techniques partition or cover the expression matrix by sub-matrices such that the expression values in each sub-matrix are highly coherent according to a suitable measure. The key property of an *x*MOTIF is that each gene in the *x*MOTIF is conserved across all the samples in that sub-matrix. It does not appear that the other techniques can capture this property. Finally, Hastie et al. [17] present a technique called "gene shaving" that tries to extract coherent and small clusters of genes that vary as much as possible across the samples; in essence, their goals are complementary to ours.

3 Determining Gene States

In this section, we describe our technique for computing the states corresponding to a gene. In our approach, a state is simply a range of expression values that is statistically significant. Similar ideas have been adopted by other researchers.[18] Thus, if we have n samples, there are $\binom{n}{2}$ possible states for each gene, each corresponding to one of the sub-intervals spanned by the expression values. Clearly, not all these states are biologically interesting. Intuitively, a state is interesting if it contains far more expression values than we would expect the state to contain if the expression values were generated at random. Formally, as a null hypothesis, we assume that gene expression values are generated by a uniform distribution. We define a state $[a, b]$ to be "interesting"

if the expression values in it are unlikely to have been generated by a uniform distribution. We compute the p-value of the decision that $[a, b]$ is interesting,[b] order states by p-value, and consider only those states whose p-value is less than a user-defined parameter. When the samples belong to different classes, we also adopt a "supervised" version of this idea. We define a state $[a, b]$ to be interesting if there is a class such that the set of expression values of the samples from that class that lie in $[a, b]$ are unlikely to have been generated by a uniform distribution. For each class, we calculate a p-value and assign the smallest p-value to $[a, b]$.

In practice, we also discard those intervals that contain more than a user-specified number of expression values. The rationale for this step is that even if an interval containing a large number of expression values is statistically significant, it may not be biologically interesting since it is unlikely to help us in distinguishing between the various classes.

4 Algorithm

We are now ready to describe our algorithm for computing the largest xMOTIF. The input to the algorithm is a set of genes, a set of samples, an expression value for each gene-sample pair, and for each gene, a list of intervals representing the states in which the gene is expressed in the samples.

To determine an xMOTIF, we have to compute the set G of conserved genes, the states that these genes are in, and the set C of samples that match the motif. We observe that if we are given (i) the set G of conserved genes, (ii) the states of the conserved genes, and (iii) one sample c that matches this motif, we can compute the remaining samples in C simply by checking for each sample c' whether the genes in G are in the same state in c and c'. Informally, c is a "seed" from which we can compute the entire motif.

This observation is the starting point of our algorithm. Suppose we know a sample c that matches the largest xMOTIF. Given such a sample c, how can we compute the genes in the largest motif and the states they are in? Suppose we are given a set D of samples with the following properties: (i) for every sample c' in D and for every gene in the largest motif, there is exactly one state such that the gene is in that state in samples c and c' and (ii) for every gene g that is not in the largest motif, there exists a sample c' in D such that gene g is not in the same state in samples c and c'. We call D a *discriminating*

[b] If k values lie in the interval $[a, b]$ and if the gene's expression values fall in the range $[0, 1]$, then the probability that an expression value falls inside $[a, b]$ is $b - a$. Therefore, the p-value of the decision that $[a, b]$ is interesting (rejecting the null hypothesis) is the sum $\sum_{k \leq i \leq n} \binom{n}{i} (b-a)^i (1-(b-a))^{n-i}$.

set. The key property of a discriminating set is that given the seed sample c and such a set D, we can compute the largest xMOTIF by including exactly those gene-states that satisfy these conditions and exactly those samples that agree with c on all these gene-states.

Algorithm 1 describes the steps of our probabilistic algorithm. We assume that for each gene, the intervals corresponding to that gene's states are disjoint.[c] It proceeds by selecting n_s samples uniformly at random from the set of all samples. These samples act as seeds. For each random seed, we select n_d sets of samples uniformly at random from the set of all samples; each set has s_d elements. These sets serve as candidates for the discriminating set. For each seed-discriminating set pair, we compute the corresponding xMOTIF as explained above. We discard the motif if less than an α-fraction of the samples match it. Finally, we return the motif that contains the largest number of genes. Suppose our data consists of n samples and m genes. We can extend the arguments made by Procopiuc et al. to prove that by choosing $n_s = O(1/\alpha)$, $s_d = O(\log m/\log(1/\beta))$, and $n_d = O(1/\alpha^{s_d})$, we can compute the largest xMOTIF with probability greater than $1/2$ in time $O(nn_sn_d) = nm^{O(\log(1/\alpha)/\log(1/\beta))}$. We can increase the probability of success by repeatedly executing this algorithm.

Algorithm 1 FINDMOTIF(): algorithm for computing the largest xMOTIF.

1: **for** $i = 1$ to n_s **do**
2: Choose a sample c uniformly at random.
3: **for** $j = 1$ to n_d **do**
4: Choose a subset D of the samples of size s_d uniformly at random.
5: For each gene g, if g is in the state s in c and all the samples in D, include the pair (g, s) in the set G_{ij}.
6: C_{ij} = set of samples that agree with c in all the gene-states in G_{ij}.
7: Discard (C_{ij}, G_{ij}) if C_{ij} contains less than αn samples.
8: **return** the motif (C^*, G^*) that maximises $|G_{ij}|, 1 \leq i \leq n_s, 1 \leq j \leq n_d$.

5 Results

We have implemented this algorithm in C++ under the Linux operating system. We present our analysis of three data sets: an ALL-AML data set,[3] a

[c] In practice, since the intervals for a gene often overlap, we modify Step 5 of the algorithm as follows: For each gene g, let I_g be the set of all states s such that g is in state s in sample c and all the samples in D. We pick a state s in I_g uniformly at random and include the pair (g, s) in the motif G_{ij}.

colon cancer data set,[4] and a B-cell lymphoma data set.[5] A detailed description of our results is available at http://genomics10.bu.edu/murali/xmotif.

For each data set, we computed 50 genes that were most informative about the class distinctions in the data.[d] For each gene in the data set, we computed its score using the "twoing" rule[19] and selected the genes with the 50 best scores. For each of these genes, we computed all the states as described earlier. Finally, we executed the *x*MOTIF-finding algorithm on each data set.

In all these experiments, we set the number of seeds $n_s = 10$ and the number of determinants $n_d = 1000$. The size s_d of the discriminating set varied from 7 to 10. The quality of our results did not change much if we varied these parameters slightly. For each data set, our algorithm took about 2–5 minutes to compute all the *x*MOTIFs when running on a computer equipped with an 800 MHz Intel Pentium III processor.

Our overall algorithm uses information about which class each sample belongs to only to compute the set of informative genes. The *x*MOTIF algorithm does not explicitly take class labels into account. As a result, samples from different classes may match a computed *x*MOTIF.

5.1 ALL-AML data

The ALL/AML data set[3] consists of the expression levels of roughly 6,800 human genes measured using an Affymetrix oligonucleotide array from bone marrow samples collected from 47 patients suffering from acute lymphoblastic leukaemia (ALL) and 25 patients suffering from acute myeloid leukaemia (AML).

For each gene, we considered only those states that had p-values at most 10^{-10} and contained at most 50 expression values. Our algorithm computes five *x*MOTIFs that cover the data. The samples matching four of these *x*MOTIFs are almost exclusively ALL patients. The fifth motif is matched almost exclusively by AML patients. The number of conserved genes in ALL-related motifs ranges from 11 to 21 while the AML-related motif contains 7 conserved genes. Table 1 displays information on each motif in the order they were computed.

The algorithm is able to compute motifs that distinguish between the two types of leukaemia almost perfectly. A total of 30 distinct genes appear in ALL-related motifs. Only one gene, TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47), is conserved both for ALL patients and for AML patients. In all the motifs, almost every gene is in an under-expressed state (relative to the total expression range of the gene).

[d] The number 50 is somewhat arbitrary. We chose it because Golub et al. build a classifier using 50 genes for the ALL-AML data set.

To obtain another view of the degree of similarity between the samples that match a motif, we computed average intra- and inter-motif distances. To compute the distance between two samples that match a motif, we used the standard Manhattan (also called rectilinear or L_1) distance, except that we ignored genes that were not conserved in that motif and we divided the distance by the number of conserved genes in that motif. We defined the distance between two samples in different motifs similarly. We observed that all the average intra-motif distances ranged between 250 and 350. All the average inter-motif distances between ALL-related motifs ranged between 310 and 515 but the average distance between ALL motifs and the AML motif ranged from 1675 to 3115. These results indicate that the ALL-related and AML-related motifs captured distinct regions of the gene expression space.

ALL-AML motifs			
#genes	#samples	#ALL	#AML
18	19	18	1
11	17	15	2
7	20	1	19
21	9	8	1
19	5	5	0

Colon cancer motifs			
#genes	#samples	#tumour	#normal
11	15	14	1
13	18	2	16
12	11	9	2
6	10	9	1
1	8	6	2

B-cell lymphoma motifs			
#genes	#samples	#alive	#dead
15	14	11	3
17	14	0	14
3	10	10	0

Table 1: Motifs computed for the data sets.

5.2 Colon cancer data

Alon et al. [4] present a data set containing 62 samples of colon epithelial cells collected from patients suffering from colon cancer; 40 of these samples are

collected from tumours and 22 from normal colon tissues of the same patients. They measured the absolute expression levels of about 6500 human genes using an Affymetrix oligonucleotide array.

We use the supervised algorithm for assigning p-values to gene states and discarded all states that contained more than 40 values or had p-value greater than 10^{-10}. Our algorithm computed five *x*MOTIFs. At least 75% of the samples matching four of these *x*MOTIFs were tumorous tissues. Almost 90% of the samples matching the fifth *x*MOTIF were normal tissues. Only one gene was conserved in the fourth tumour-related *x*MOTIF but the other *x*MOTIFs contained between 6 and 13 genes. Only one gene, P03001 Transcription Factor IIIA, was conserved in both types of motifs. If we ignore the fourth tumour-related *x*MOTIF (since it contains only one conserved gene), distances between tumour-related *x*MOTIFs ranged from 90 to 265, while all the distances between the normal *x*MOTIF and the tumour *x*MOTIFs were greater than 610.

5.3 B-cell lymphoma data

Alizadeh et al. [5] describe a data set of 96 normal and malignant lymphocyte samples whose expression levels they measured using a specialised "Lymphochip." Alizadeh et al. had information of the survival rates of the patients suffering from diffuse large B-cell lymphoma (DLBCL). Of the 40 such patients, 22 survived (data available at http://llmpp.nih.gov/lymphoma/data.shtml). We applied the supervised version of our algorithm to this classification (using only gene states with p-value at most 10^{-3}) and obtained three motifs, with two motifs being conserved across patients who survived and one motif conserved across patients who died. Only three genes were conserved across both classes. One of these is a NF-κB family member that is frequently amplified in diffuse large cell lymphoma. The other two are ESTs of unknown function. In each motif, the conserved genes appear in states ranging from relatively under-expressed to relatively over-expressed. The separation between the *x*MOTIFs in terms of distance was not as clear as for the other data sets.

6 Conclusions

We have introduced a useful and concise representation of gene expression data in the form of conserved gene expression motifs or *x*MOTIFs. These motifs capture the degree of conservation in the gene expression profiles of the samples belonging to a class at two levels: (i) each gene in the motif is similarly-expressed in each of the samples and (ii) all the genes in the motif are simultaneously conserved in all these samples. We believe that this representation has the

potential to capture many key biological properties implicitly present in gene expression data. We have implemented a system to compute large xMOTIFs that cover all the classes in the data. Our analysis of three publicly-available data sets shows that our algorithm can compute xMOTIFs that appear to distinguish between the classes in these data sets quite well. Our technique has the potential to find clinically and biologically relevant subdivisions in gene expression data.

In our current formulation, we require that each gene in a motif be expressed in the same state across all samples matching the motif. A useful generalisation of this concept is the requirement that the gene's expression levels in a motif obey a specified distribution. Extending our algorithm to this case appears to be a challenging problem.

Acknowledgements We would like to than Geoff Cooper and Michael Schaffer for useful discussions. We thank Yang Su for providing us with the gene-selection software. An NSF-KDI grant partly funded this research.

References

1. S Fodor, R Rava, X Huang, A Pease, C Holmes, and C Adams. Multiplexed biochemical assays with biological chips. *Nature*, 364(6437):555–6, 1993.
2. M Schena, D Shalon, R Davis, and P Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.
3. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
4. U Alon, N Barkai, D A Notterman, K Gish, S Ybarra, D Mack, and A J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–50, 1999.
5. A A Alizadeh, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, H Sabet, T Tran, X Yu, J I Powell, L Yang, G E Marti, T Moore, J Hudson, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, and L M Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

6. Cecilia Magdalena Procopiuc, Michael T. Jones, Pankaj K. Agarwal, and T. M. Murali. A monte-carlo algorithm for fast projective clustering. In *Proceedings of the 2002 International Conference on Management of Data*, 2002.
7. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, NY, 1979.
8. Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, 2002.
9. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
10. R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown. Clustering methods for the analysis of dna microarray data. Technical report, Department of Statistics, Stanford University, 1999. Available at http://www-stat.stanford.edu/ tibs/ftp/jcgs.ps.Z.
11. N Friedman, M Linial, and I Nachman. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–20, 2000.
12. R Sharan and R Shamir. Click: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:307–16, 2000.
13. J. A. Hartigan. Direct clustering of a data matrix. *J. Amer. Statis.*, 67:123–129, 1972.
14. Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. ISMB*, 2000.
15. Amos Tanay, Roded Sharon, and Ron Shamir. Biclustering gene expression data. In *Proceedings of ISMB 2002*, pages S136–S144, 2002. Available at http://www.math.tau.ac.il/ roded/publications.html.
16. G. Getz, E. Levine, and E. Domany. Coupled two-way clustering of dna microarray data. *Proc. Natl Acad. Sci. USA*, 97:12079–12084, 2000.
17. T Hastie, R Tibshirani, M B Eisen, A Alizadeh, R Levy, L Staudt, W C Chan, D Botstein, and P Brown. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, 1(2), 2000.
18. Jinyan Li and Limsoon Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, 18(5):725–34, 2002.
19. S. K. Murthy, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–33, 1994.