Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-003
AIM-2004-002
CBCL-234

January 27, 2004

Selecting Relevant Genes with a Spectral Approach

Lior Wolf, Amnon Shashua, and Sayan Mukherjee

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Selecting Relevant Genes with a Spectral Approach

Lior Wolf[a], Amnon Shashua[†], and Sayan Mukherjee[‡a]

[a] Center for Biological and Computational Learning
McGovern Institute for Brain Science
Cambridge, MA 02139

[†]School of Engineering & Computer Science, Hebrew University, Jerusalem, 91904 Israel

[‡] Whitehead Institute/MIT Center for Genome Research

{liorwolf,sayan}@mit.edu, shashua@cs.huji.ac.il

ABSTRACT

Array technologies have made it possible to record simultaneously the expression pattern of thousands of genes. A fundamental problem in the analysis of gene expression data is the identification of highly relevant genes that either discriminate between phenotypic labels or are important with respect to the cellular process studied in the experiment: for example cell cycle or heat shock in yeast experiments, chemical or genetic perturbations of mammalian cell lines, and genes involved in class discovery for human tumors. In this paper we focus on the task of unsupervised gene selection. The problem of selecting a small subset of genes is particularly challenging as the datasets involved are typically characterized by a very small sample size — in the order of few tens of tissue samples — and by a very large feature space as the number of genes tend to be in the high thousands. We propose a model independent approach which scores candidate gene selections using spectral properties of the candidate affinity matrix. The algorithm is very straightforward to implement yet contains a number of remarkable properties which guarantee consistent sparse selections. To illustrate the value of our approach we applied our algorithm on five different datasets. The first consists of time course data from four well studied Hematopoietic cell lines (HL-60, Jurkat, NB4, and U937). The other four datasets include three well studied treatment outcomes (large cell lymphoma, childhood medulloblastomas, breast tumors) and one unpublished dataset (lymph status). We compared our approach both with other unsupervised methods (SOM,PCA,GS) and with supervised methods (SNR,RMB,RFE). The results clearly show that our approach considerably outperforms all the other unsupervised approaches in our study, is competitive with supervised methods and in some case even outperforms supervised approaches.

1. INTRODUCTION

In DNA microarray expression studies, estimated abundances of thousands of mRNA species in different tissue samples are obtained through hybridization to oligonucleotide or cDNA arrays. Biological class differences manifest themselves as significant differences in the expression levels of a relatively *small* set of genes, resulting in the observed overabundance of mRNA. The set of relevant genes is typically small since the majority of the active cellular mRNA is not affected by the biological differences. In other words, a significant difference in biological characteristics (such as a normal cell versus a tumor cell from the same tissue) does have a gene expression level manifestation, but the set of genes that is involved can be rather small. For example, previous work on classification of tumor tissue samples based on gene expression profiles has shown that in many cases, cancer types can be discriminated using only a small subset of genes whose expression levels strongly correlated with the class distinction [1, 2]. Identifying highly relevant genes from the data is therefore a fundamental problem in the analysis of expression data.

Relevant genes can be selected either in a supervised or unsupervised fashion. A tissue sample consists of a vector in R^n describing the expression values of n genes/clones. In a *supervised* setting each tissue sample is associated with a label — typically binary or trinary (negative, positive, control) — denoting its class membership. In an *unsupervised* setting, the class labels are omitted or unknown. A variety of algorithms exist for supervised gene selection: signal-to-noise [1], recursive feature elimination [3], t-test metrics [4], Wilcoxon rank sum test [4], and gene shaving [5]. These studies all make the implicit assumption that that relevant genes are discriminative genes.

However, discrimination is not the only measure of relevance and there are many studies where the objective does not necessarily consists of some measure of discrimination. Some examples include finding genes relevant for cell cycle [6], analyzing a compendium of expression profiles with different mutations [7], and the problem of class discovery [8]. Unsupervised methods for selecting relevant features have been applied in these types of problems using singular value decomposition (SVD) [6], principle components analysis (PCA) and iterative principle components analysis, a.k.a gene shaving, [5, 9], max-surprise [10], self organizing maps [8], and hierarchical clustering [11].

In this paper we focus on the task of unsupervised gene selection. Gene selection, unlike other applications of feature selection in the machine learning literature, is characterized by first and foremost by a very small sample size q — typically in the order of few tens of tissue samples — and by a relatively very large feature space R^n as the number of genes tend to be in the thousands ($n \approx 10^4$). Coupled with the notion that applications in the domain of unsupervised gene selection (such as "class discovery") require one to discover things which are unknown or unexpected, it follows that the unsupervised gene selection process should be model-independent as much as possible. Motivated by this fact, we approach the gene selection task as a process of dividing the tissue

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profi t or commercial advantage and that copies bear this notice and the full citation on the fi rst page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specifi c permission and/or a fee.
RECOMB'04, Submitted.
Copyright 2003 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

samples into k (typically $k = 2, 3$) clusters where the goal is to find the gene subset which maximizes the clusters coherency. In other words, we assume that if one knew which were the relevant genes to begin with then the tissue sample values corresponding to the selected genes would be naturally clustered in into k sets. Our task therefore, is to find (among the exponentially many gene subsets) that subset which maximizes the cluster coherency.

The task of gene selection is somewhat different from the that of interpreting patterns of gene expressions [8, 12]. In the latter case, by regarding the quantitative expression levels of n genes over q samples as defining n points in R^q, one employs a clustering technique for grouping together genes which have similar expression profiles and use the cluster averages as expression profiles. Various clustering techniques have been proposed from direct visual inspection [12] to employing self organization maps (SOM) [8]. The visual inspection approach does not scale well to larger data sets, is best suited for data with an expected pattern (like cyclic cell lines) and thus is less appropriate for discovering unexpected patterns. The SOM clustering approach, as with all exploratory data analysis tools, involves manual inspection of the data to extract insights. Gene selection in comparison is an open problem. Rather than grouping the gene expression levels into clusters, one seeks to distinguish a small set of genes which are relevant to the biological classification of the tissue samples. As a result of this distinction, rather than grouping the gene expression levels we look for a subset of genes for which the corresponding tissue sample values are coherently divided into k (2 or 3) clusters. The notion of clustering is still there but in an indirect manner — *the goodness of clustering is used a score for the gene selection process.* More on the distinction between selection and pattern seeking, and comparison with SOM results are discussed in Section 4.1.

The clustering score in our approach is measured indirectly. Rather than explicitly performing a clustering phase per gene selection candidates, we employ spectral information in order to measure the cluster arrangement coherency. Spectral algorithms have been proven to be successful in clustering [13], manifold learning or dimensionality reduction [14, 15], approximation methods for NP-hard graph theoretical questions [16]. In a nutshell, given a selection of genes, the strength (magnitude) of the leading k eigenvalues of the affinity matrix constructed from the corresponding expression levels of the selected genes are directly related to the coherence of the cluster arrangement induced by the subset of selected genes. More details are described in Section 2.

It is worthwhile noting that unsupervised gene selection differs from dimensionality reduction in that it only selects a handful of genes (features) which are "relevant" with respect to some inference task. Dimensionality reduction algorithms, for example PCA, generate a small number of features each of which is a combination of all of the original features. A main purpose of expression analysis is to extract a set of genes that are of interest from the perspective of the biological process being studied. In general, it is assumed that each such process involves a limited number of genes. For this reason feature combination methods are not as desirable as methods that extract a small subset of genes. The challenge in the selection process is to overcome the computational burden of pruning an exponential amount of gene subsets. The $Q - \alpha$ algorithm [17] which we propose using as a basis for our approach handles the exponential search space by harnessing the spectral information (the sum of eigenvalues of the candidate affinity matrix) in such a manner where a computationally straightforward optimization guarantees a sparse solution, i.e., a selection of genes rather than a combination of the original genes.

In the next section we will describe the workings of the spectral approach to gene selection and the mechanism by which a sparse solution (i.e., a selection) is obtained. In the sections that follow we apply our algorithm on a variety of gene expression datasets. We show that for datasets with label information our unsupervised algorithm performs as well as the leading supervised algorithms and is significantly superior to previously proposed unsupervised algorithms. We also show that genes selected using the unsupervised $Q - \alpha$ overlap with those selected using supervised procedures and have some biological relevance.

2. METHODS: SELECTING GENES WITH THE $Q - \alpha$ ALGORITHM

The array based technologies, cDNA and oligonucleotide, for studying gene expression levels provide static information about gene expression (i.e. in which tissue(s) the gene is expressed) and dynamic information (i.e. how the expression pattern of one gene relates to those of others). In general the raw data have to be corrected for different experimental conditions by a normalization procedure sometimes followed by a logarithmic transformation to the absolute intensities or ratios. This gives the data matrix whose rows correspond to genes and whose columns correspond to tissue samples. We assume that exactly one value for each gene/sample is given, which may be achieved over repeated measurements for samples or genes.

Let the microarray data matrix be denoted by M. The gene expressions levels form the rows of M denoted by $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$ and normalized to unit norm $\|\mathbf{m}_i\| = 1$. Each row vector represents a gene sampled over the q trials. The column vectors of M represent the q samples (each sample is a vector in R^n). As mentioned in the previous section, our goal is to select rows (genes) from M such that the corresponding candidate data matrix (containing only the selected rows) consists of columns that are coherently clustered in k groups. The value of k is user dependent and is typically 2 or 3 denoting the expected number of different biological classes contained in the tissue samples. The challenge in this approach is to avoid the exponential number of row selections and preferably avoid explicitly clustering the columns of the candidate data matrix per each selection.

Mathematically, to obtain a clustering coherency score we compute the "affinity" matrix of the candidate data matrix defined as follows. Let $\alpha_i \in \{0, 1\}$ be the indicator value associated with the i'th gene, i.e., $\alpha_i = 1$ if the i'th gene is selected and zero otherwise. Let A_α be the corresponding affinity matrix whose (i, j) entries are the inner-product (correlation) between the i'th and j'th columns of the resulting candidate data matrix: $A_\alpha = \sum_{i=1}^{n} \alpha_i \mathbf{m}_i \mathbf{m}_i^\top$ (sum of rank-1 matrices). From algebraic graph theory, if the columns of the candidate data matrix are coherently grouped into k clusters, we should expect the leading k eigenvalues of A_α to be of high magnitude [18, 19, 20, 21, 13]. The resulting scheme should therefore be to maximize the sum of eigenvalues of the candidate data matrix over all possible settings of the indicator variables α_i.

What we do in practice, in order to avoid the exponential growth of assigning binary values to n indicator variables, is to allow α_i to receive real values in an unconstrained manner. A least-squares energy function over the variables α_i is formed and its optimal value is sought after. What makes this approach different from the "garden variety" soft-decision-type algorithms is that this particular setup of optimizing over spectral properties guarantees that the α_i *always come out positive and sparse* over all local maxima of the energy function. This property is intrinsic rather than being the result of explicit constraints in the form of regularizers, priors or inequality constraints. The energy function takes the following

form:

$$\max_{Q,\alpha_i} \ \text{trace}(Q^\top A_\alpha^\top A_\alpha Q) \qquad (1)$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i^2 = 1, \ \ Q^\top Q = I$$

Note that the matrix Q holds the first k eigenvectors of A_α and that trace$(Q^\top A_\alpha^\top A_\alpha Q)$ is equal to the sum of squares of the leading k eigenvalues: $\sum_{j=1}^{k} \lambda_j^2$. A local maximum of the energy function is achieved by interleaving the "orthogonal iteration" scheme [22] within the computation of α as follows:

DEFINITION 1 ($Q - \alpha$ METHOD). *Let M be an $n \times q$ input matrix with rows $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$, and some orthonormal $q \times k$ matrix $Q^{(0)}$, i.e., ${Q^{(0)}}^\top Q^{(0)} = I$. Perform the following steps through a cycle of iterations with index $r = 1, 2, ...$*

1. *Let $G^{(r)}$ be a matrix whose (i, j) components are*

$$(\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top Q^{(r-1)} {Q^{(r-1)}}^\top \mathbf{m}_j$$

   .

2. *Let $\alpha^{(r)}$ be the leading eigenvector of $G^{(r)}$.*
3. *Let $A^{(r)} = \sum_{i=1}^{n} \alpha_i^{(r)} \mathbf{m}_i \mathbf{m}_i^\top$.*
4. *Let $Z^{(r)} = A^{(r)} Q^{(r-1)}$.*
5. *$Z^{(r)} \xrightarrow{QR} Q^{(r)} R^{(r)}$, that is, $Q^{(r)}$ is determined by the "QR" factorization of $Z^{(r)}$.*
6. *Increment index r and go to step 1.*

Note that steps 4,5 of the algorithm consist of the "orthogonal iteration" module, i.e., if we were to repeat steps 4,5 *only* we would converge onto the eigenvectors of $A^{(r)}$. However, the algorithm does not repeat steps 4,5 in isolation and instead recomputes the weight vector α (steps 1,2,3) before applying another cycle of steps 4,5.

The algorithm would be meaningful provided that three conditions are met:

1. the algorithm converges to a local maximum,
2. at the local maximum $\alpha_i \geq 0$ (because negative weights are not admissible), and
3. the weight vector α is *sparse* (because without it the soft decision does not easily translate into a hard gene selection).

Conditions (2) and (3) are not readily apparent in the formulation of the algorithm (the energy function lacks the explicit inequality constraint $\alpha_i \geq 0$ and an explicit term to "encourage" sparse solutions) but are nevertheless satisfied. The key for having sparse and non-negative (same sign) weights is buried in the matrix G (step 1). Generally, the entries of G are not necessarily positive (otherwise α would have been non-negative due to the Perron-Frobenious theorem) — nevertheless due its makeup it can be shown that in a probabilistic manner the leading eigenvector of G is positive with probability $1 - o(1)$. In other words, as the number of genes n grows larger the chances that the leading eigenvector of G is positive increases rapidly to unity. Fig. 2 shows the (sorted) α values on the Hematopoietic differentiation cell lines (details about this dataset are found below). The details of why the makeup of G induces such a property, the convergence proof and the proof of the "Probabilistic Perron-Frobenious" claim can be found in [17].

Finally, it is worth noting that the scheme can be extended to handle the supervised situation (when class labels are provided); that the scheme can be applied also to the Laplacian affinity matrix; and that the scheme readily applies when the spectral gap $\sum_{i=1}^{k} \lambda_i^2 - \sum_{j=k+1}^{q} \lambda_j^2$ is maximized rather than $\sum_{i=1}^{k} \lambda_i^2$ alone. Details can be found in [17].

3. DATASETS

We evaluated our proposed approach for gene selection on five datasets — one of which is a time course dataset and the remaining four datasets with outcome or status labels. With the four datasets with label information we applied supervised approaches to compare with our unsupervised gene selection algorithm.

The first dataset consisted of time course data from four Hematopoietic cell lines [8]: HL-60, Jurkat, NB4, and U937. The dimensionality of the expression data was $7,229$ genes. The HL-60, U937, and Jurkat cell lines were stimulated with phorbol 12-myristate 13-acetate (PMA) for $(0, .5, 4, 24)$ hours. The NB4 cell line was stimulated with with all trans-retinoic acid (ATRA) for $(0, 6, 24, 48, 72)$ hours.

The remaining four datasets were treatment outcome or status studies. The first was a study of treatment outcome of patients with diffuse large cell lymphoma (DLCL), reffered to as "lymphoma" [23]. The dimensionality of this dataset was $7,129$ and there were 32 samples with good successful outcome and 26 with unsuccessful outcome. The second was a study of treatment outcome of patients with childhood medulloblastomas [24], referred to as "brain". The dimensionality of this dataset was $7,129$ and there were 39 samples with good successful outcome and 21 with unsuccessful outcome. The third was a study of the metastatis status of patients with breast tumors [25], reffered to as "breast met". The dimensionality of this dataset was $24,624$ and there were 44 samples where the patients were disease free for 5 years after onset and 34 samples where the tumors metastasized within five years. The fourth is an unpublished study of of breast tumors [26] for which corresponding lymph nodes either were cancerous or not, referred to as "lymph status". The dimensionality of this dataset is $12,600$ with 47 samples positive for lymph status and 43 negative for lymph status.

4. RESULTS

The datasets, as detailed above, used for our experiments consist of thousands of genes (in the order of 10^4). Many of the techniques presented in the past begin with a pre-filtering step aiming at reducing the number of genes from thousands to hundreds. For example, [8] passes the gene expression vectors through a variation filter before applying the SOM code for clustering the remaining gene expression vectors. The variation filter eliminates those genes with no significant change across the samples.

One of the strengths of our approach is the ability to handle large amounts of data. Any preprocessing filtering step of the data imposes a prior which very likely has a dramatic effect on the final results. In many cases, the final results depend not so much on the strength of the main algorithm but on the type and care placed on the pre-filtering step. Therefore, we applied our algorithm on the original data set without performing pre-filtering steps for pruning the data. The results reported below start with data matrices consisting of thousands of genes and produces a few tens of relevant genes.

4.1 Comparison with SOM on Time Course Data

A significant amount of expression data is time course data. Finding relevant genes in these types of datasets is an open problem.

PCA is a reasonable approach when the underlying factor of the study is cyclical for example cell cycle [6] or circadian rhythms [27]. However, for many studies the underlying process of interest is not cyclical. One may want to find genes that increase in expression over time in one cell line but decrease in expression for another cell line. A standard approach to address this is to cluster genes and use the clusters as expression profiles. Using our unsupervised gene selection procedure we can find the relevant genes in time course data directly without having to cluster.

In [8] Hematopoietic differentiation was studied across four cell lines. Two myeloid cell lines HL-60 and U937 were examined, a T cell line called Jurkat was examined, and an acute promyelocytic leukemia cell line was examined. Time course data for these four cell lines was concatenated into a dataset with 17 samples and 7229 genes. A 6×4 self-organizanizing map (SOM) was used to cluster this dataset after preprocessing with a variation filter. The 24 clusters are displayed in figure 1. We applied the $Q - \alpha$ to this dataset and found that the set of relevant genes was sparse (meaning it contains a small number of relevant genes) as shown in Fig. 2. Of the genes corresponding to the top 40 α values we display the time course signatures of 6 genes if figure 3 for brevity. The signature of all 40 genes can be found at www.ai.mit.edu/people/sayan/hemat. The time course of these 6 genes correspond to clusters 20, 1, $22/23$, 4, 15, 21 in figure 1. For a biological explanation of these genes or coresponding clusters see [8]. Using the $Q - \alpha$ algorithm we were able to recapitulate the time courses of [8] with *individual genes* rather than gene clusters and also find those genes that are relevant.

4.2 Comparison with Other Supervised and Unsupervised Methods using Labeled Data

For the four datasets with label information classification accuracy was used as a measure of the goodness of the (unsupervised) $Q - \alpha$ algorithm. We compared the leave-one-out error on these datasets with that achieved by both supervised and unsupervised methods of gene selection. The supervised methods used were signal-to-noise (SNR) [1], radius-margin bounds (RMB) [28, 29], and recursive feature elimination (RFE) [3]. The unsupervised methods used were PCA and gene shaving (GS) [5]. In the unsupervised mode the class labels were ignored — and thus in general one should expect the supervised approaches to produce superior results than the unsupervised ones. A linear support vector machine classifier was used for all the gene selection methods [30]. Parameters for SNR, RFE, and RMB were chosen to minimize the leave-one-out error. A summary of the results appear in table 1.

The $Q - \alpha$ algorithm considerably out-performs all other unsupervised methods. Furthermore, and somewhat intriguing, is that the $Q - \alpha$ algorithm is competitive with the other supervised algorithm (despite the fact that the labels were not taken into account in the course of running the algorithm) and performs *significantly better* on the lymph status of breast tumors as compared to all other gene selection approaches — including the supervised methods.

5. DISCUSSION

The advent of array technologies make it possible to collect data on thousands of genes simultaneously while recording both static information (in which tissues the gene is expressed) and dynamic information (how the expression pattern of one gene relates to those of others). A typical microarray contains several to tens of thousands of genes over a relatively few samples (few scores). In many cases it has been observed that among the many genes only a small fraction are really relevant for providing discriminatory information over the tissue samples or providing other non-discriminatory

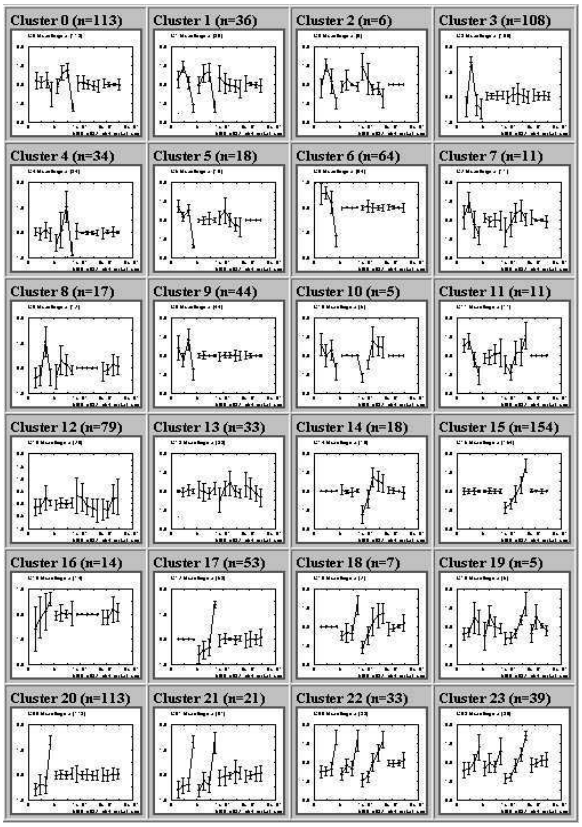

Figure 1: A plot of the 24 SOM clusters from the Hematopoeitic differentiation cell lines. In each of the 24 clusters the time courses of all four cell lines are shown (left to right) HL-60+PMA, U937 + PMA, NB4+ATRA, Jurkat+PMA. This is Figure 4 from Tamayo et al.

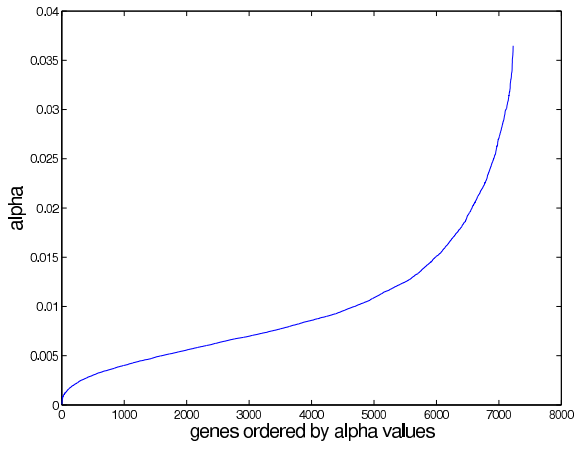

Figure 2: A plot of the sorted α-values for the Hematopoietic differentiation cell lines. As noted, all values come out positive despite the fact that positivity is not explicitly constrained in the energy function. The profile of the values indicates sparsity meaning that around 95% of the values are of an order of magnitude smaller than the remaining 5%.

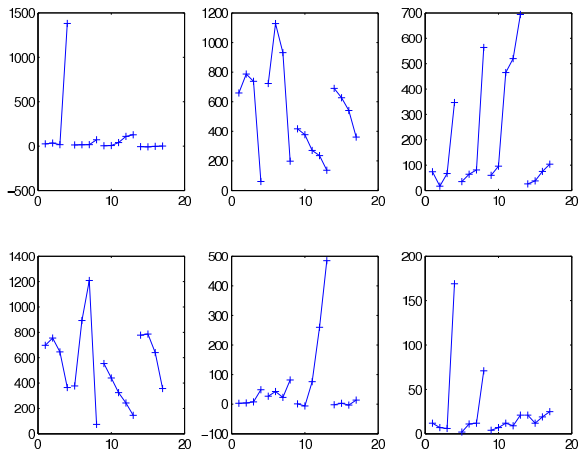

Figure 3: A plot of 6 of the top 40 genes that correspond to clusters 20, 1, 22/23, 4, 15, 21 in Tamayo et al. In each of the six panels time courses of all four cell lines are shown (left to right) HL-60+PMA, U937 + PMA, NB4+ATRA, Jurkat+PMA.

Method	brain	lymph status[1]	breast met.[1]	lymphoma
RAW	32	44	34	27
PCA5	22	47	33	40
PCA10	26	47	26	27
PCA20	25	47	25	29
PCA30	31	47	31	33
PCA40	31	47	31	33
PCA50	30	47	30	33
GS5	20	45	32	33
GS10	24	43	31	30
GS20	28	47	32	31
GS30	30	44	33	33
$Q-\alpha$	15	19	22	15
SNR	16	42	29	18
RFE	14	38	26	14
RMB	13	39	24	14

Table 1: Leave-one-out classification results for the supervised and unsupervised algorithms on the various datasets. In both PCAN and GSN the number N the number of components used. Parameters for SNR, RFE, and RMB were chosen to minimize the leave-one-out error. The $Q-\alpha$ algorithm considerably out-performs all other unsupervised methods. Furthermore, and somewhat intriguing, is that the $Q-\alpha$ algorithm is competitive with the other supervised algorithm (despite the fact that the labels were not taken into account in the course of running the algorithm) and performs significantly better on the lymph status of breast tumors as compared to all other gene selection approaches — including the supervised methods. [1] Only the first $7,000$ genes were used.

information about class discovery or cell line analysis. The task of selecting a small subset of genes is particularly challenging from an information theoretic point of view in light of the few samples. Another challenge in gene selection (or generally in feature selection) is the combinatorial explosion introduced when all possible gene subsets are to be scored.

In this work we focused on the unsupervised version of gene selection. The selection is unsupervised when class labels are either absent (as in class discovery) or when the selection is required in the context of cellular process studied in experiments — such as cell cycle or heat shock in yeast experiments and chemical or genetic perturbations of mammalian cell lines. The algorithm we proposed was also applied to supervised settings (where class labels are available) and was competitive and in some cases performed significantly better than well studied supervised gene selection methods. In unsupervised settings, our algorithm significantly and consistently outperformed other well studied approaches.

The principle of our method is based on scoring gene subsets by means of measuring the coherence of the cluster arrangements of the sample vectors induced by the selection. The cluster coherence can be indirectly evaluated by the magnitude of the leading eigenvalues of the corresponding affinity matrix. The combinatorial explosion problem is avoided by the special makeup of a key matrix in the algorithm which makes possible to use a soft-selection type of approach yet guarantee sparse solutions (i.e. discrete solution).

We have first illustrated the value of our approach on a problem that is inherently unsupervised – that of finding relevant genes in time course data. Instead of directly selecting relevant genes most algorithms cluster all genes and explain the time courses in terms of gene clusters and then look for genes in the various clusters to try and understand the underlying biology. We directly find the relevant genes in the time course data. We compared the two approaches on four well studied Hematopoietic cell lines (HL-60, Jurkat, NB4, and U937). Using our approach we were able to find individual genes with time courses very similar to those of gene clusters found using SOMs on this dataset. We then applied our algorithm to four (labeled) treatment outcome datasets. Cmparisons with other supervised and non-supervised approaches showed a consistent superiority over other unsupervised approaches we tested in our studies and comparable performance to supervised approaches (despite the fact that our algorithm did not make use of the available class labels). In one case, the performance of our algorithm on the lymph nodes dataset for breast tumor study was significantly superior compared to all the methods we compared against — including the supervised methods.

The simplicty of our approach to gene selection (see Matlab code in Section 6) and the success of the $Q-\alpha$ algorithm in finding relevant genes in the datasets explored in this paper suggests that it would be a useful addition to expression analysis tools being developed. Coupling this approach with tools such as Gene Set Enirchment Analysis [31, 32] which take the list of relevant genes and measures whether genes in known or computed pathways fall into this list can provide valuable insights into biological processes at the molecular level.

6. CODE

To illustrate the simplicity of the $Q-\alpha$ algorithm we attach a vanilla implementation of it below:

function [alpha,A] = Qalpha(M,nclusters,numIter);
%M is the measurement matrix
%nclusters is the number of expected clusters
%numIter is the total number of iterations

```
%normalize data to have zero mean unit norm features vectors
[M,normXparams] = normalizeData(M);
mimj = M*M';

%init A,Q,alpha
alpha = rand(nf,1);
alpha = alpha./norm(alpha);
A = diag(alpha)*M'*M*diag(alpha);
[Q,R] = qr(A,0);
Q = Q(:,1:nclusters);

for i = 1:numIter,
   %compute G
   G = zeros ( size(M,1) )
   miq = M*Q;
   for j = 1:nclusters,
      D = diag(miq(:,j) );
      G = G + D*mimj*D;
   end
   %compute the vector of weights alpha
   [alpha,S] = svds(G,1);

   A = diag(alpha)*M'*M*diag(alpha);

   %Orthogonal Iteration steps
   Z = A*Q;
   [Q,R] = qr(Z,0);
   Q = Q(:,1:ne);
end
```

7. REFERENCES

[1] T. R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. *Science*, 286:531-537, (1999).

[2] S. Dudoit, J. Fridlyand and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 2001.

[3] I. Guyon, J. Weston, S. Barnhill and V. Vapnik Gene selection for cancer classification using support vector machines. *Machine Learning*, Vol. 46, pp. 389–422, (2002).

[4] L. Miller et al. Optimal gene expression analysis by microarrays *Cancer Cell.*, Vol. 2, pp. 353–361, (2002).

[5] T. Hastie et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns *Genome Biology*, 1(2), pp. 1–21, (2000).

[6] O. Atler et al. Singular value decomposition for genome-wide expression data processing and modeling. *P.N.A.S.* 97:18, pp. 10101-10106, (2000).

[7] T .R. Hughes et al. Functional Discovery via a Compendium of Expression Profiles. *Cell* , 102:109-126, (2000).

[8] P. Tamayo et. al. Interpreting patterns of gene expression with self-organizing maps P.N.A.S., 96:6, pp. 2907-2912, (1999).

[9] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. *Pacific Symposium on Biocomputing*, Honolulu, Hawaii, 452-463, (2000).

[10] A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data, *RECOMB*, pp. 31-38, (2001).

[11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. *P.N.A.S.*, 95, pp. 14863–14868, (1998).

[12] R. J. Cho et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 95(5), pp. 65-73, (1998).

[13] A.Y. Ng, M.I. Jordan and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. *NIPS*, 2001.

[14] M. Belkin, P. Niyogi Semi-supervised Learning on Manifolds *Machine Learning Journal*, Special Issue of Clustering, to appear.

[15] A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, (2000).

[16] N. Alon Spectral Techniques in Graph Algorithms *LATIN*, 206–214, (1998).

[17] L. Wolf and A. Sashua Feature Selection for Unsupervised and Supervised Inference: the Emergence of Sparsity in a Weighted-based Approach. *Submitted to J. of Mach. Learn. Research.* Note: preliminary version to appear in ICCV, Oct. 2003.

[18] T.S. Motzkin and E.G. Straus. Maxima for graphs and a new proof of a theorem by turan. *Canadian Journal of Math.*, 17:533–540, 1965.

[19] L. E. Gibbons, D. W. Hearn, P. M. Pardalos, and M. V. Ramana. Continuous characterizations of the maximum clique problem. *Math. Oper. Res.*, 22:754–768, 1997.

[20] M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[21] F. R. K. Chung. *Spectral Graph Theory*. AMS, 1998.

[22] G. Golub and C.V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[23] M. A. Shipp et al. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning *Nature Medicine*, 8:1, pp. 68–74, (2002).

[24] S. L. Pomeroy et al Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors *Nature*, 415:24, pp. 436–442, (2002).

[25] L. J. van't Veer et al. Expression profiling predicts cliniacl outcome of breast cancer *Nature*, 415, pp. 530–536, (2002).

[26] S. Ramaswamy Personal communication.

[27] R. Schaffer et al Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell*, 13 (1), pp. 113-123, (2001).

[28] O. Chapelle et al. Choosing Multiple Parameters for Support Vector Machines *Machine Learning*, 2001.

[29] J. Weston et al. Feature Selection for SVMs *NIPS*, 13, pp. 668–764, (2001).

[30] V. N. Vapnik. *The nature of statistical learning*. Springer, 2nd edition, 1998.

[31] V. Mootha et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34, pp. 267–273, (2003).

[32] A. Subramanian et al. Gene Set Enrichment Analysis. in preparation.