Class Discovery in Gene Expression Data

Amir Ben-Dor
Agilent Laboratories
and
University of Washington
amirbd@cs.washington.edu

Nir Friedman[*]
Hebrew University
nir@cs.huji.ac.il

Zohar Yakhini
Agilent Laboratories
and
Technion
zohary@labs.agilent.com

ABSTRACT

Recent studies (Alizadeh et al, [1]; Bittner et al,[5]; Golub et al, [11]) demonstrate the discovery of putative disease subtypes from gene expression data. The underlying computational problem is to partition the set of sample tissues into statistically meaningful classes. In this paper we present a novel approach to class discovery and develop automatic analysis methods. Our approach is based on statistically scoring candidate partitions according to the overabundance of genes that separate the different classes. Indeed, in biological datasets, an overabundance of genes separating known classes is typically observed. we measure overabundance against a stochastic null model. This allows for highlighting subtle, yet meaningful, partitions that are supported on a small subset of the genes.

Using simulated annealing we explore the space of all possible partitions of the set of samples, seeking partitions with statistically significant overabundance of differentially expressed genes. We demonstrate the performance of our methods on synthetic data, where we recover planted partitions. Finally, we turn to tumor expression datasets, and show that we find several highly pronounced partitions.

1. INTRODUCTION

An important application of gene expression profiling technologies, such as array-based hybridization assays, is to improve our understanding of cancer related processes. Recent studies [1, 5, 11, 16] demonstrate the discovery of putative disease sub-types from gene expression data. For example, Alizadeh *et al* [1] discover a putative sub-class of a certain type of lymphoma, and Bittner *et al* [5] discover a putative sub-class of cutaneous melanoma. In both cases the findings were further biologically validated.

In this paper we discuss algorithmic approaches to the unsupervised inference of novel putative tissue subclasses in gene expression data. Current approaches to this problem [1, 5, 11] are a combination of clustering driven methods and human intervention. A clustering approach starts by computing similarity values (e.g. Pearson correlation) of the expression profiles of each pair of tissue samples, and then uses these values to partition the data. This approach suffers from several shortcomings. First, the measure of similarity depends uniformly on the entire set of measured genes. In practice, dramatic phenotypical differences might effect only a relatively small subset of the mRNA transcripts.[1] Such differences are "washed out" by uniform measures of similarity. For example, the classification proposed by Alizadeh *et al* [1] does not appear when tissues are clustered using all the genes. Instead, the authors hand-pick a small subset of genes and cluster tissue samples with respect to this subset. Second, clustering methods return a single clustering of the data. In actual data we expect multiple significant partitions where different tissue classes are separated by different sets of genes (e.g., treatment success for two drugs, each targeting a different pathway).

We take a direct approach to class discovery. The process we develop consists of two components. We start by defining a figure of merit to putative partitions of the set of samples. We then apply heuristic search methods, such as *simulated annealing*, to find the best partition in the space of all possible partitions of the set of samples. Finally, we iterate the process to find additional, different, partitions. To develop an effective figure of merit as a basis for comparing different putative classes we are guided by the fact that an overabundance of significantly differentially expressed genes is typically observed in data that contains meaningful biological partitions. That is, the number of genes that sharply separate the two classes is extremely higher than expected in the *null* model where partitions of the data are uniformly drawn. Therefore, reasoning in reverse, we seek putative partitions for which we observe an overabundance of informative genes.

In Section 2 we describe a measure for how well genes separate different classes and define the null-model. In Section 3 we use these to derive a *surprise score* for putative classifications of the data. In Section 4 we consider an alternative score that is based on the ability to predict putative class membership based on the gene expression profile of a sample (this can be thought of as an in-silico classification assay). In Section 5, we describe search procedures. In Section 6 we present a synthetic model of gene expression data, and use it to evaluate the performance of different methods. In Section 7 we apply our methods to actual biological data. We conclude with a discussion in Section 8.

2. INFORMATIVE GENES

[*] Contact author.

[1] Two cells with dramatically different biological characteristics (such as a normal cell versus a tumor cell from the same tissue) are expected to also have different gene expression profiles. It is important, however, to realize that the majority of the active cellular mRNA is not effected by the differences. That is, a dramatic biological difference does have a gene expression level manifestation, but the set of genes that is involved can be rather small.

2.1 Scoring Informative Genes

We start with some definitions. Assume that we are given a *data set* D, consisting of vectors $\langle x_1, \ldots, x_M \rangle$. Each *sample* or *expression pattern*, x_i, is a vector in $\mathbf{R}^N$ that describes the expression values of N genes/clones in a particular biological sample. A (binary) *labeling* for D is a vector $l = \langle l_1, \ldots, l_M \rangle$, where the *label* l_i associated with x_i is either -1 (negative example), $+1$ (positive example), or 0 (control example).

Consider a set of expression data with a known classification of tissues. This is typically based on auxiliary information, such as histological measurements, pathological analysis, or genetic level information. We want to understand molecular level differences between the different classes of tissues. Such data often spans thousands of genes. Some of these genes play major roles in the processes that underly the differences between the classes or are dramatically effected by the differences. Such genes are highly relevant to the studied phenomenon. On the other hand, the expression levels of many other genes may be irrelevant to the distinction between tissue classes. Identifying highly relevant genes from the data is a basic problem in the analysis of expression data.

The literature discusses several methods for scoring genes for relevance. These include parametric measures, such as the standard *t-test* score [15], the *separation score* of Golub *et al* [11], and non-parametric measures [3, 4].

We now briefly describe the $TNoM$ (*Threshold Number of Misclassification*) score of Ben-Dor *et al.* [3] which we use for scoring genes in this work. We emphasize that the ideas we present can be easily applied with other relevance scores.

Let k denote the number of tissues, consisting of a tissues from class A, and b tissues of class B. Assume we want to score a gene g for relevance with respect to the A:B partition of the tissues. Intuitively, g is relevant to the tissue partition if it is either over-expressed in class A tissues (compared to class B tissues) or vice-versa.

To formalize the notion of relevance, we consider how g expression levels in class A tissues interlace with its expression levels in class B tissues. Denote by t_i the i-th tissue ranked according to the expression level of g (that is, g express minimally in t_1 and maximally in t_k). We define the rank vector, v, of g to be a $-,+$ vector of length k, as follows:

$$v_i = \begin{cases} + & \text{if } t_i \in A \\ - & \text{if } t_i \in B \end{cases}$$

Note that the rank vector v captures the essence of the differential expression profile of g. If g is under-expressed in class A, then the positive entries of v are concentrated in the left hand side of the vector, and the negative entries are concentrated at the right hand side. Similarly, for the opposite situation. Therefore, the relevance of g increases as the homogeneity within the left hand side of v, and the homogeneity within the right hand side of v increase.

A natural way to define the homogeneity on the two sides, and to combine them into one score, leads to the TNoM Scoring method. The score of v corresponds to the maximal combined homogeneity over all possible ways to break v to two parts. Define the *Min-Cardinality*, of a $-,+$ vector x, to be the cardinality of the minority symbol in x. That is,

$$MC(x) = \min(\#_-(x), \#_+(x)).$$

The TNoM score of a rank vector v is defined as

$$TNoM(v) = \min_{x;y=v}(MC(x) + MX(y)).$$

2.2 p-Values

When scoring a gene for how relevant it is to a given partition of the set of samples it is important to evaluate the result against a null model - what is the probability of this gene (with the given expression values) appearing so relevant for a randomly drawn partition of the samples. This number is the *p-value* corresponding to the scoring method in effect and the given level s. Genes with very low p-values are very rare in random data and their relevance to the studied phenomenon is therefore likely to have biological, mechanistic or protocol reasons.

Let $\{-1, +1, 0\}^{(n,p,c)}$ denote all labelings with $n\ '-1'$ entries, $p\ '+1'$ entries, and $c\ '0'$ entries. Let l be a labeling, and let g be a vector of gene expression values. A scoring method $\mathcal{S}$ (e.g., $TNoM$) is a function that takes g and l and returns the score of g with respect to the labeling l. Let L be a random labeling drawn uniformly over $\{-1, +1, 0\}^{(n,p,c)}$. The p-value of a score level s is then

$$pVal(s) = Prob(\mathcal{S}(g, L) \leq s) \qquad (1)$$
$$\text{where} \quad L \sim \mathrm{Unif}\left(\{-1, +1, 0\}^{(n,p,c)}\right)$$

The combinatorial character of the TNoM score makes its distribution over $\{-1, +1, 0\}^{(n,p,c)}$ amenable to rigorous calculations. Ben-Dor *et al* [3] develop efficient procedures for computing the exact distribution of TNoM scores in $\{-1, +1, 0\}^{(n,p,c)}$.

Access to p-values allows us to compute the expected number of genes with score s or better in the null model. Examining data sets with biologically meaningful classifications, we found an over-abundance of significantly informative genes [4, 5]. For example, Figure 1 contrasts the expected number of genes with particular p-value to the actual number of genes for the $TNoM$ score, in various previously published datasets. This overabundance analysis is instrumental in evaluating the statistical significance of putative previously unknown classes, as in [5]. Biological significance can then be experimentally established as described therein.

3. SURPRISE SCORES

Recall that our aim is to apply statistical methods to the discovery of sample classifications. Clearly, the ultimate test for a putative classification is a biological validation test. Statistics here is a tool and not an end by itself. Our approach has two components: a statistical score measuring the significance of a suggested partition and a procedure that attempts to find the labeling with the highest score. The significance score has to be well correlated with biological meaning. Only more data will help us learn more about the advantages and disadvantages of the various scores. In this section we discuss candidate scores, termed *surprise scores*.

The significant overabundance of informative genes in biologically meaningful pre-classified data (e.g., see Figure 1) suggests that biologically meaningful classifications of the sample set can be characterized by such overabundance. Biological class differences manifest themselves as dramatic differences in the expression levels of a (not very large) set of genes, resulting in the observed overabundance. Therefore, we will choose candidate putative sample classes amongst those label vectors that show a significant overabundance of informative genes when applied to the data.

To formalize this we use a stochastic model, as follows. Suppose we want to evaluate a labeling in $\{-1, +1, 0\}^{(n,p,c)}$. For any score s set $p_s = pVal(s)$, where the parameters n, p, c characterize the composition of the putative labeling we want to evaluate, and $L \sim \mathrm{Unif}\left(\{-1, +1, 0\}^{(n,p,c)}\right)$. Let $i_{\leq s}(g)(L) = \mathbf{1}\{\mathcal{S}(g, L) \leq s\}$ indicate the event that the gene g received a score s or better (i.e., smaller or equal s), under L. Clearly,

$$i_{\leq s}(g) \sim Bernoulli(p_s), \text{ for } g = g_1, \ldots, g_N.$$

The number of genes with score s or better, under L, is simply $N_{\leq s}(L) = \sum_g i_{\leq s}(g)(L)$. We have thus defined a set of random variables $N_{\leq s}$ on $\{-1, +1, 0\}^{(n,p,c)}$. Given a putative labeling, l we can compute the score of each gene and summarize these scores in a set of numbers $n_{\leq s}(l)$ (one number for each value s in the range of scores of the scoring function). We now ask how rare it is to observe such values for $N_{\leq s}(L)$ when $L \sim \mathrm{U}nif\left(\{-1, +1, 0\}^{(n,p,c)}\right)$.

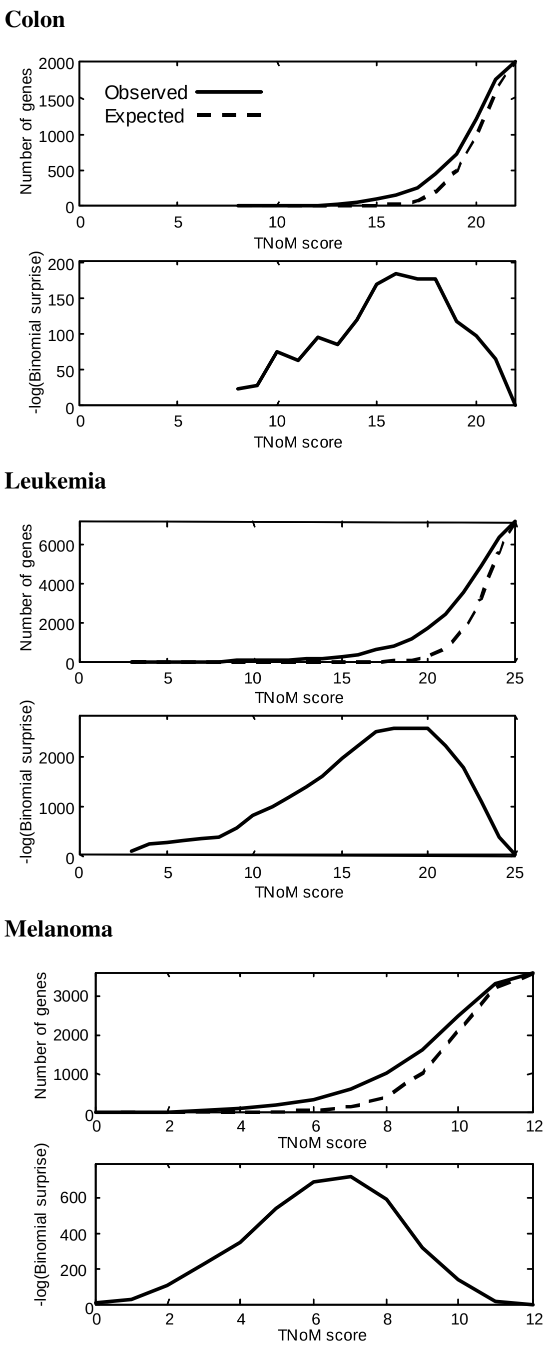

Figure 1: Comparison of the number of significant genes in actual dataset to expected number under the null-hypothesis (random labels). The x-axis denotes TNoM score. In the top part of each graph, the y-axis is the number of genes, in the bottom part, the y-axis is the negative logarithm of the probability of the expected number of genes given the binomial model, which is exactly the surprise score (see below). Data sets: Colon from Alon *et al* [2]; leukemia from Golub *et al* [11]; melanoma from Bittner *et al* [5].

3.1 Max-Surprise Score

One of the simplest modeling assumptions we can make is the following

> Ind-Genes: The indicator variables $i_{\leq s}(g),\ g = g_1, \ldots, g_N$ are independent.

This assumption implies that knowing the scores of any subset of $\{g_i : i \neq j\}$ under a random labeling L adds no information about the score of g_j under the same labeling. It is clear that this assumption oversimplifies things: expression levels of genes can be very correlated, and thus their scores on random labeling are correlated. Nonetheless, as in many modeling situations, this simplifying assumption allows for efficient computations, and might not change the essence of the results.

Under this independence assumption $N_{\leq s}$ is a sum of independent Bernoulli variables, and thus has a binomial distribution: $N_{\leq s} \sim \mathrm{B}in(p_s, N)$ (N is the number of genes in he dataset). We define the *surprise* at observing n genes at score s or better, to be

$$\text{Surprise}(s, n) = -\log(P(N_{\leq s} \geq n))$$

Where P is the appropriate binomial probability measure. The smaller the probability of observing n, the larger the surprise. We can compute the surprise score using the tail probability of the binomial distribution.

Of course, at different score values we get different surprise values. However, we can find the score s at which the observed number is most surprising:

$$\text{Max-Surprise}(l) = \max_s \text{Surprise}(s, n_{\leq s}(l)).$$

This quantity is employed to evaluate a labeling vector. This matches our intuition that a good labeling should have a surprisingly large number of informative genes.

3.2 Sanov Score

The Max-Surprise score computes the score at which the largest overabundance of informative genes is observed. However, it ignores the distribution of scores. For example, suppose that the maximum score is at s such that $p_s = 10^{-3}$ with $n_{\leq s}(L) = 100$. The Max-Surprise may be the same if all 100 genes have score exactly s or if a non-negligible fraction of these have better scores (with much smaller p-values).

Suppose $s_1, \ldots, s_k$ is the range of possible scores in the given set-up. Under the null hypothesis and under the gene independence assumption, the set $\{N_{s_i} : i = 1, \ldots, k\}$ is multinomially distributed with $p(i) = P(\mathcal{S}(g, L) = s_i)$. The frequency at which we observe the score s_i in the actual data is $q(i) = \frac{n_{s_i}}{N}$. The distribution q describes the *type* of this empirical sample. Under the null model, the probability of drawing this type with N samples is exactly

$$P^N(q) = \binom{N}{Nq_1, \ldots, Nq_k} \prod_i p_i^{Nq_i}$$

We can approximate this probability using an information theoretic measure of the distance.

THEOREM 3.1. *[7, Theorem 12.1.4]:*

$$-ND(p\|q) - \log(k(N+1)) \leq \log P^N(q) \leq -ND(p\|q)$$

where

$$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$$

is the KL divergence *between* p *and* q.

Thus, $2^{-ND(p\|q)}$ is a first-order exponential approximation to the probability of the type. Moreover, using this approximation we can bound the probability of observing types that are more skewed than q.

THEOREM 3.2. *(Sanov's Theorem) [7]:*
$\log P^N(\{q : D(p\|q) > \epsilon\}) \leq -N\epsilon + k\log(N+1)$

This motivates the definition of a surprise measure by

$$\text{Sanov-Surprise}(l) = ND(p\|q_l) - k\log(N+1)$$

where q_l is the type of the observed distribution of gene scores under the labeling vector l.

To summarize, we have defined a measure of surprise based on the entire vector of observed scores frequencies. This quantity is a bound on the probability of observing all types q with $D(p\|q) \geq D(p\|q_l)$. Observe that the Max-Surprise score measures the probability of having a labeling vector at least as informative as the candidate l, in terms of overabundance of informative genes. On an intuitive level, when the entire distribution is considered, a labeling vector l is more informative than K if the type q_l has more weight on the low (better) scores than does q_K. We are not formalizing this here and are using the Sanov bound on the probability of at least as improbable types rather than at least as informative types.

4. CLASSIFICATION SCORES

In the previous sections we studied scores that evaluate the quality of putative label vectors by measuring significant deviations from the distribution of scores we expect under null-hypothesis models. An alternative approach is to seek classes which are predictable based on the gene expression measurements.

4.1 Classification of Gene Expression Patterns

In [3, 16] the problem of predicting tissue classification is examined. As demonstrated there, for actual labelings in real-life data sets, it is possible to *train* a classifier that has good predictive accuracy. More precisely, a classification algorithm is a function f_D that depends on a data set D of patterns and sample labels. Given a new query $x \in \mathbf{R}^N$, this function returns a predicted label $\hat{l} = f_D(x)$. Good predictive accuracy means that predicted labels match the "true" label of the query. Several classification methods were applied to gene expression data [3, 4, 16]. For completeness we briefly review the method used here.

The *naive Bayesian classifier* [8, 9, 10] is based on a probabilistic approach to the problem. We start by estimating the probability of each label (e.g., -1 or $+1$) given gene g's expression level. We model this distribution by a *decision stump*: we learn a threshold t, and make one prediction of $x_g < t$ and another if $x_g > t$. The threshold t is chosen as in the $TNoM$ score, and the conditional distribution for x_g above (resp. below) the threshold is estimated from the proportions of $+1$ and -1 labels for samples where g's expression level is above (resp. below) t. Then, assuming that expression patterns of genes are independent given the labeling (this is the "naive" assumption) and using Bayes rule, we get that

$$\log \frac{P(+1 \mid x)}{P(-1 \mid x)} = \log \frac{P(+1)}{P(-1)} + \sum_g \left(\log \frac{P(+1 \mid X_g)}{P(-1 \mid X_g)} - \log \frac{P(+1)}{P(-1)} \right).$$

If this quantity is positive we predict $+1$, otherwise we predict -1. See [4] for more details.

A key issue we need to address is how to evaluate the accuracy of a classification method applied to a given labeled data set and labeling. We follow standard methodology and use *leave one out cross validation* (LOOCV) to estimate the prediction accuracy of a classification method on new examples. This procedure iterates on the samples in the data set. In each iteration it removes a single sample and trains the classification procedure on the remaining data. The trained classifier is then applied to the held-out sample and the predicted label is compared to the true label. The fraction of errors thus committed in the entire process, is our estimate of the error rate of the classification procedure.

A final issue is *feature selection*. As [3, 16] show, predictions based on an informative subset of genes are more accurate than these that are based on all genes. In our procedure here we employ a simple, but surprisingly effective, procedure to select genes. Given a training data, we compute the score s that attains max surprise. We then focus on genes that have this score or better. The learned classifier is then based on these genes only. Note that in each LOOCV iteration this procedure is applied on the data set without the held-out sample. Thus, each time a different score will attains maximum surprise and a different set of genes is selected.

4.2 Classification Score for Putative Labeling Vectors

Suppose we are given a putative labeling of samples in our training data. The intuition we outlined above suggests that if the labeling captures a "true" phenomenon in the data, then a LOOCV evaluation of a classification procedure (e.g., the naive Bayesian classifier) would lead to accurate predictions. In other words, we can score a labeling by the accuracy reported by LOOCV evaluation of classification with respect to this labeling. This suggests that the distinctions made by the labelings are inherent in the data and not an artifact. From a different point of view: the suggested classes can be successfully in-silico diagnosed.

There are several shortcomings to the classification score. First, it is computationally intensive, since we need to perform LOOCV iterations, and in each of these perform gene selection (which in turn, requires computing scores for all genes). Second, since the number of samples is small, the range of the score is quite limited. This implies that the classification score gives little guidance during search for high scoring labeling vectors. Third, if we use the classification score during search we are going to perform large number of LOOCV evaluations; statistical considerations show that even in random data, enough repetitions of this test will find high scoring artifactual labeling vectors [13, 14]. To avoid these problems, we mainly use the classification score to evaluate candidate labeling vectors that are found using surprise scores we discussed in the previous section.

5. SEARCH METHODS

Once we choose a score for putative labeling vectors we need to find a labeling that maximizes the score. This is a discrete optimization problem and we use heuristic search techniques to find high-scoring labelings. We can formalize our problem as a search over a graph, where our goal is to find a vertex with maximum score, where one assumes some locality in the scores (i.e., neighboring vertices have similar scores). Presently, the vertices correspond to potential labelings of the samples, and the score we attempt to maximize is the max-surprise score. The edges in the graph are between pairs of labelings that agree on all labels except one sample, the labeling of which changed from "0" to either "-1" to "+1". Thus, we can move from one labeling to another by modifying the label of exactly one sample from classified to unclassified, or vice versa. Note that over a set of M samples, each vertex has at most $2M$ neighbors.

A common search method is the *first ascend hill climb*. In this procedure we consider all the neighbors of our current labelings in some random order. We evaluate the score of each neighbor, and once we find a neighbor with a better score, we "move" there and continue. If all neighbors have worse scores than the current candidate, we are at a local maximum, which is returned.

This procedure is straight forward and has the intuitive aspect of climbing up-hill toward better solution. However, it can get "stuck" at local maxima. Unfortunately, local maxima are common in the class discovery optimization problem. A common method to escape local maxima is *simulated annealing* [12]. This method resembles the first ascend procedure. However, now the search procedure maintains a temperature parameter t. This parameter is updated during the search by an exponentially decreasing *cooling schedule*: at the kth step of the process the temperature is $t_0\alpha^{\lceil k/d \rceil}$ where $\alpha < 1$ and d is an integer. Now, if the score of the current labeling is s, and a random neighbor labeling scores s', we move to that neighbor with probability $\min(1, e^{\frac{s'-s}{t}})$. Thus, at very high temperatures the probability of taking a score-decreasing step is close to 1. It gets closer to 0 as the temperature decreases. The procedure terminates after a fixed number of steps (or equivalently, after t reaches a pre-specified temperature) and returns the best scoring labeling it encountered.

Recall that we want to construct *several* different partitions of the data. Toward this end, we employ a simple strategy of *peeling* the data set. First, we perform a heuristic search and find a high scoring putative labeling. Then, we "peel off" the genes that support this labeling from the data set. More precisely, we remove all genes with score smaller than or equal to the score that attains maximum surprise. We then reiterate the search on the remaining genes until either we exhausted all genes, or the score of the best labeling on these genes falls below a pre-specified threshold. By iteratively peeling the data set we discover a set of partitions, each supported by a disjoint set of genes. Once we finalize the search, we reevaluate each of the labeling vectors with respect to the original data set (since some previously removed genes can be relevant to a putative partition and effect its score).

6. MODEL AND SIMULATIONS

All attempts to stochastically model gene expression data are intrinsically problematic. It is impossible to make a reasonable set of model assumptions that is universally valid for a complicated system such as the living cell. Modeling approaches are, however, successful in highlighting biological phenomena that do follow the model and thus allow for selective inference of knowledge from data.

The purpose of the stochastic simulation exercise we describe in this section is threefold: to validate our computational class discovery methods on a general stochastic model; to identify the mode of convergence to planted classes; to compare performance across methods and test the effects of various parameter changes.

6.1 The Stochastic Model: Planted Classes

We assume that the gene expression dataset stochastically depends on a hidden, biologically significant, classification $\mathcal{C}$ of the tissues into subclasses. As in real datasets, we further assume that the classification $\mathcal{C}$ effects only a small fraction of the genes, called the $\mathcal{C}$*-genes*, while the other genes, called *random genes*, express independently of $\mathcal{C}$.

For simplicity we describe a binary classification model. Modelling data with more classes in the same manner is, however, straight forward. We assume that there exists a hidden classification $\mathcal{C}$, that partitions the M tissues into a class A tissues and b class B tissues. We denote by m the total number of genes, and by $1-e$ the fraction of $\mathcal{C}$-gene. That is, there are $(1-e)m$ $\mathcal{C}$-genes, and em random genes.

For each $\mathcal{C}$-gene, g, we model its expression levels in the different tissues using two distributions - $\mathcal{D}_A$ for class A tissues, and $\mathcal{D}_B$ for class B tissues. We assume that $\mathcal{D}_A$ and $\mathcal{D}_B$ are normal distributions with a constant coefficient of variation[2], s. That is,

$$\mathcal{D}_A = N(\mu_A, \mu_A s), \quad \mathcal{D}_B = N(\mu_B, \mu_B s).$$

The means of the distribution, μ_A, and μ_B are uniformly chosen from the interval $[-1.5d, 1.5d]$. Thus, the expected distance between the two means is d (a parameter of the model).

The expression level of the random genes in all tissues, independent of their class, is assumed to be normally distributed with zero mean and standard deviation of one. Note that any classification of the tissues, $\mathcal{C}'$, might be supported by some random genes. However, the true classification $\mathcal{C}$, will be supported by a statistically significant number of genes because of the $\mathcal{C}$-genes.

In summary, the planted classification model is fully specified by a classification $\mathcal{C}$, of M tissues into two classes of sizes a and b respectively, and by the model parameters:

- m - the total number of genes.
- e - the fraction of random genes in the data.
- d - the expected distance between the two mean expression levels, pertaining to the two planted classes.
- s - the coefficient of variation for expression level distributions.

6.2 Results on Synthetic data

In this section we report a simulation based evaluation of our discovery process. We varied the model parameters (a, b, m, e, d, s), and employed Max-Surprise and Sanov (defined in Section 3) in a simulated annealing local search. In initial simulations we observed that max-surprise based searches perform better than Sanov score based searches, so we chose to use Max-Surprise in the rest of the study, and for real datasets.

Simultaneously varying d, s, and e, we observed that the search results were relatively insensitive to the parameter d (compared to s and e). Hence, we concentrate on s and e in the rest of the simulations.

In order to choose realistic parameter values we examined the leukemia data set [11] and best fit the model parameters to it. Omitting the details of the fitting process, the resulting values are: $m = 7129$, $a = 25$, $b = 47$, $e = 0.72$, $d = 555$, $s = 0.75$.

In our stochastic model we implicitly assume that all genes are independently distributed. However, in biological dataset, there are complicated dependencies among genes. Therefore, the effective number of independent genes in the real data set is much smaller

[2]This assumption is supported, for example, when the expression levels are logs of red to green signal ratios in a two dye expression profiling measurement [6].

Data set	Labeling	p/n/c	Max-Surprise Score	p-Value	#	Sanov score	LOOCV acc. (%)	Jackard Coeff.
Leukemia	original	47/25/0	2601	0.0154	1173	2057	91.7	1
	1	43/29/0	13784	0.0007	1890	7733	98.6	0.469
	2	32/40/0	7541	0.0126	2182	5465	91.7	0.344
	3	43/29/0	11524	0.0054	2400	7781	93.0	0.469
	4	46/26/0	2690	0.0558	2014	2235	87.5	0.949
Lymphoma	Original	50/46/0	8259	0.0010	1188	-	87.5	1
	1	24/37/35	14514	$1.9 * 10^{-5}$	1148	-	100	0.780
	2	53/23/20	8342	0.0049	1598	-	100	0.382
	3	34/36/26	7728	0.0012	1148	-	91.4	0.485
	4	33/29/34	6674	0.0013	1046	-	88.7	0.539
Lymphoma	Original	23/22/51	545	0.0139	359	-	97.8	1
DLBCL	1	33/12/51	2669	0.0668	1625	-	88.9	0.362
	2	23/22/51	2005	0.0139	776	-	95.5	0.324
	3	25/20/51	917	0.0460	815	-	88.9	0.354
	4	31/14/51	2171	0.1318	1975	-	82.2	0.350

Table 1: Evaluation of the best 4 discovered labelings and the original labelings in three data sets. The table reports the composition of the labeling; the max-surprise score, the p-value at the point of max surprise and the number of genes with that p-value; the Sanov score; LOOCV accuracy of predictions the labeling (ignoring control samples); and the *Jackard coefficient* that measures the similarity of the labeling to the original labeling.

than 7129. One way to choose a better model value for m is to choose it such that the Max-Surprise score of the hidden classification $\mathcal{C}$ (fixing the other parameters to the above values) would resemble the Max-Surprise score of the AML/ALL classification in the leukemia data (which is 2603). Using this approach we derive $m = 600$. Therefore, set the *leukemia parameters* to be

$$m = 600,\ \ a = 25,\ \ b = 47,\ \ e = 0.72,\ \ d = 555,\ \ s = 0.75$$

To test the performance of our methods on leukemia parameters, we generated 10 synthetic datasets according to the planted classification model, and compared the returned classification of the tissues to the original, planted classification. In all 10 cases, the original class was recovered perfectly.

To better study the effect of the model parameters on the algorithm performance, and to learn our algorithm limits we have varied each of the parameters (m, d, e, s) in turn, while fixing the others to their leukemia value. In the reported results below, we use $\mathcal{C}$ to denote the planted classification (that has proportions of 25 class A tissues vs. 47 class B tissues), and by $\mathcal{A}$ the classification returned by our algorithm. Recall that our algorithm searches for the tissue classification with the maximal Max-Surprise score. As we vary the model parameters, the Max-Surprise score advantage of $\mathcal{C}$ (compared to other classifications) changes, and thus the algorithm performance is accordingly effected:

- m - Increasing m, the number of genes, increases Max-Surprise$(\mathcal{C})$, and thus makes it easier for the search heuristic to find it. We have found that $m \approx 250$ is the phase transition point. If m is larger, then the algorithm consistently recovers the hidden classes. However, for smaller m's, $\mathcal{C}$ is *not* the optimal classification (with respect to the Max-Surprise score), and thus a different classification, $\mathcal{A}$, is recovered. The difference between $\mathcal{A}$ and $\mathcal{C}$ depends on m; the smaller m is, the larger is the difference. For example, setting $m = 100$, we get that on average Max-Surprise$(\mathcal{C}) \approx 360$, while Max-Surprise$(\mathcal{A}) \approx 390$. Still, $\mathcal{A}$ and $\mathcal{C}$ are very close (differ on average only on 3 tissues).
- d - We have found that d, the expected distance between the two means has very little effect on Max-Surprise$(\mathcal{C})$, and thus has very limited effect on the algorithm performance. In particular, we have varied d in the range 1 through 1000, and in all cases the algorithm recovered $\mathcal{C}$ perfectly.
- e - In our model, e represented the fraction of random genes in the data, genes that express independently from the planted classification. Stated differently, we are trying to recover planted classifications that are supported on a $1 - e$ fraction of the genes. As the Max-Surprise score of a classification reflects the over-abundance of informative genes, we expect Max-Surprise based methods to perform well even for high values of e. Indeed, in this study by simulations we have varied e in the range $[0, .99]$, and observed that the algorithm consistently recovered $\mathcal{C}$, up to $e = 0.95$. For higher values of e, we typically get Max-Surprise$(\mathcal{A}) >$ Max-Surprise$(\mathcal{C})$.
- s - The coefficient of variation, s, plays a major role in our model. It represents the inherent random nature of the expression profile of a gene within tissues of the same class. For large values of s we get very spread distributions, contributing to higher TNoM scores, and thus a lower Max-Surprise$(\mathcal{C})$. In this study we varied s in the range $[0.5, 5]$. The transition point was found at around $s = 2$. For smaller s, the planted classification $\mathcal{C}$ is recovered, for larger $\mathcal{C}$, we typically recover classification $\mathcal{A}$ with larger Max-Surprise score.

Our simulation study can be summed up as follows. First, the algorithm is very robust, performing under high levels of noise, either in the form of random genes ($e \approx .95$), or in form of high coefficient of variation ($s \approx 2$). Second, for a wide range of parameters, even much more pessimistic than those that correspond to the leukemia dataset, the algorithm consistently recovers the planted classification. Finally, if either there ar too few genes ($m < 250$), or too high noise level ($e > 0.95$, or $s > 2$), than the planted classification is no longer the optimal classification, and we cannot hope to perfectly recover it.

7. CLASS DISCOVERY IN GENE EXPRESSION DATA

To evaluate the usefulness of our approach, we applied it to several gene expression data sets. They all come with a known classification that is either based on pathological considerations, or was discovered using manual analysis of gene expression data. The data sets are: **Leukemia:** 72 expression profiles reported by Golub *et al* [11]. These samples are divided to two variants of leukemia:

25 samples of *acute myeloid leukemia* (AML) and 47 samples of *acute lymphoblastic leukemia* (ALL). mRNA was extracted from 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays spanning 7129 genes. **Lymphoma:** 96 expression profiles reported by Alizadeh *et al* [1]. 46 of these are of *diffused large b-cell lymphoma* (DLBCL) samples. The remaining 50 samples are of 8 types of tissues. In our analysis we used gene expression measurements of 4096 genes shown in [1, Figure 1]. **Lymphoma-DLBCL:** This data set is the subset of 46 DLBCL samples from the lymphoma data set. Alizadeh *et al* separated these samples into two classes *Germinal centre B-like DLBCL*, and *Activated B-like DLBCL*.

In each of these data sets we run the peeling procedure using the maximum surprise score of Section 3.1. Table 1 summarizes the scores of the top discovered classifications using the various scoring mechanisms we discussed above and their difference from the original classification of the data.

On the leukemia data set we run our search procedure with the additional constraint that it should only examine labeling without control tissues. Peeling found six labelings, the first four of which are shown in Table 1. All six labelings score better than the original labeling by both the max-surprise and Sanov scores, and by the number of "significant" genes. The first three labelings also have better LOOCV accuracy than the original score. Thus, we believe that each of these captures a significant distinction. Note that the first three labelings are quite different from the original one (the Jackard coefficient is low). Of the next three labelings, two (4 and 6) are very similar to the original labeling, yet receive slightly lower LOOCV scores.

In the Lymphoma data set, peeling found 7 labelings. The top 2 labelings score better than the original labelings both in terms of max-surprise score and LOOCV accuracy. The first labeling contains a large group which contains mostly DLBCL samples (34 out of 37), and another group consisting mostly of samples of other types of lymphoma (Fl and CLL) . We note though that additional 12 DLBCL are set as controls. Thus, we suspect that this classification is based on genes whose expression separates DLBCL samples from the types we mentioned above.

When we focused on the DLBCL samples (constraining all others samples to be controls), peeling found 4 labelings. These labelings are all quite different than the one reported by Alizadeh *et al* [1]. All three score higher in terms of max-surprise and are supported by larger number of genes. The classification of Alizadeh *et al*, however, has higher LOOCV accuracy.

For some of the DLBCL samples, Alizadeh *et al* also report survival data [1]. They show that the classification they discover in the data is a good predictor of patient survival chances. They also show, that this distinction is informative even if they focus only on low clinical risk patients. (Clinical risk is evaluated using *international prognostic index*, a standard medical index, evaluated at the time the sample was taken.) In Figure 2 we plot survival rates for patients for the four putative DLBCL classifications described in Table 1. As we can see, some of the classifications, such as the forth one, are not predictive about patient survival. On the other hand, the second and third classifications are predictive about the survival chances of patients with good prognostic evaluation, and the third classification is also predictive for the whole patient population. This shows that the classifications we discover are potentially relevant to the development of the disease.

In conclusion, in two of these data sets we manage to recover close approximations to known biologically meaningful classifications. In addition, in all three data sets we uncovered classifications that are as strongly pronounced in the data (large number of genes

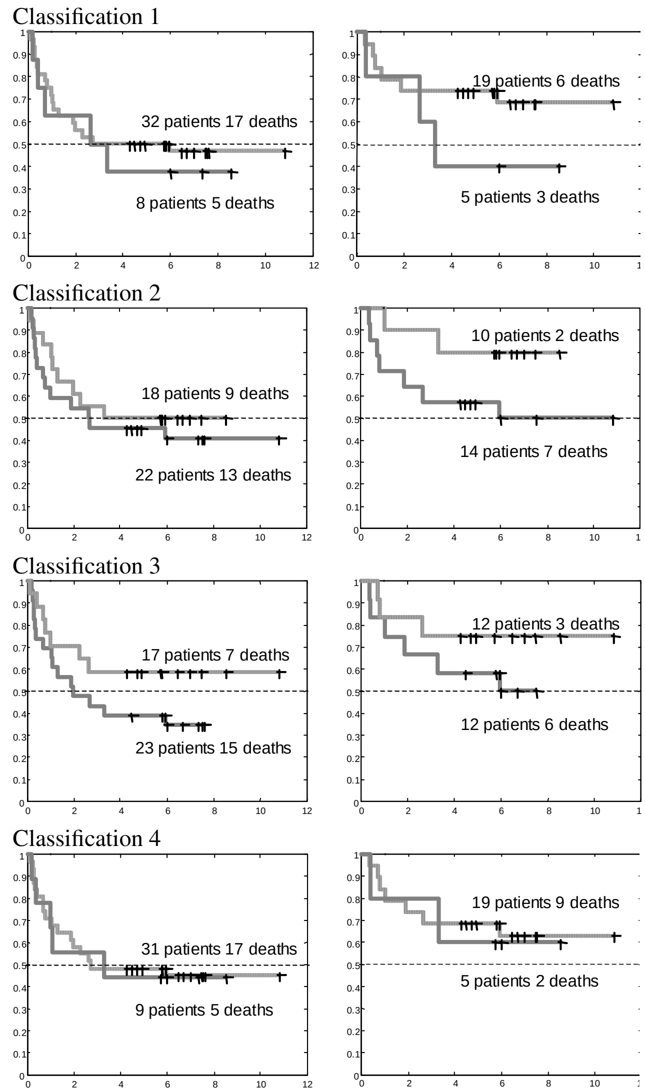

Figure 2: Kaplan-Meier survival plots for the 4 DLBCL classifications described in Table 1. The x-axis is the number of years after the samples were taken, and the y-axis is the fraction of patients survived so far. Each plot shows the survival rate for the two classes defined by a putative classification. The plots on the left column show the survival rate of all 40 patients for whom survival data is available. The plots on the right column, show the survival rate of the 24 patients with low clinical risk (see [1] for details).

at significant p-value). These classifications might be biologically meaningful or artifacts of the sample preparation, or hybridization procedures. In either case, it is important that the analysis of the results take in to account such strong signals in the expression data.

8. CONCLUSIONS

The contribution of this paper is threefold. First, we put forth the problem of class discovery and distinguish it from standard clustering problems. Second, we propose several criteria for evaluating putative classifications for significance. The central idea is to quantify the overabundance of genes that are informative with respect to a putative classification. Finally, we develop an efficient search procedure for finding multiple significant classifications in data sets.

The main criterion we use in searching for new classifications is the max-surprise score. This score is appealing both because of its definition is clear and can be easily mapped to biological counterparts, and because it can be efficiently evaluated. Our synthetic evaluation shows that searching using the max-surprise score can recover a "true" classification in synthetic data under a wide range of operating parameters including the number of relevant and irrelevant genes, the amount of variance in the expression level, and the difference between the expression of genes in two classes.

When we applied this procedure to real-life cancer related gene expression data sets, we found multiple highly pronounced classifications that were supported by independent evaluation methods that measure the predictiveness of the classifications. Our procedure managed to recover close approximations to known classification in two of these data sets.

The work reported here opens several intriguing research questions. First, both the max-surprise and the Sanov score exploit a strong independence assumption. This assumption can potentially overstate the surprise of the scores we observe in the data. Thus, although our procedures performed well in practice, we still might be able to improve upon them by relaxing this independence assumption. A potential direction of work is estimating the distribution of $N_{\leq s}(L)$ under the null hypothesis without assuming independence. A first cut approach is based on stochastic simulation. Unfortunately, simple stochastic simulation is useful only for estimating the distribution of scores with relatively large p-value. For scores with small p-values, we will need massive repetitions of the simulation to get a single case where such a score is attained. We are currently working on developing more sophisticated methods for estimating the distribution of $N_{\leq s}(L)$ under the null-hypothesis, and using this estimates to get a better assessment of the surprise.

Another issue is the search procedure. In this work we mainly focused on the criteria for evaluating putative classifications, and used simulated annealing, a fairly generic search method, with parameters that ensure a wide search. In addition we used peeling for finding multiple classifications. In the future, we plan to study the theoretical properties of this optimization problem, aiming at developing principled methods for this task.

Acknowledgements

We thank Ash Alizadeh and Izidore Losses for useful discussions relating to this work, and for making the DLBCL survival data available. Nir Friedman was supported by ISF grant 244/99, Israeli Ministry of Science grant 2008-1-99, and an Alon fellowship.

9. REFERENCES

[1] 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, Jr. Hudson, J., 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.

[2] 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.

[3] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–584, 2000.

[4] A. Ben-Dor, N. Friedman, and Z. Yakhini. Scoring genes for relevance. Technical Report 2000-38, School of Computer Science & Engineering, Hebrew University, Jerusalem, 2000. http://www.cs.huji.ac.il/∼nir/Abstracts/BFY1.html, and Technical Report AGL-2000-13, Agilent Labs, Agilent Technologies, 2000, http://www.labs.agilent.com/resources/techreports.html.

[5] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, 2000.

[6] Y. Chen, E. Dougherty, and M. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4):364–374, October 1997.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[9] J. Friedman. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1, 1997.

[10] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[11] 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(5439):531–7, 1999.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[13] A. J. Klockars and G. Sax. *Multiple Comparisons*. Sage Publications, 1986.

[14] A. Y. Ng. Preventing "overfitting" of cross-validation data. In *Proc. 14th Inter. Conf. Machine Learning*, pp. 245–253. 1997.

[15] M. Schummer, W. V. Ng, R. E. Bumgarner, P. S. Nelson, B. Schummer, D. W. Bednarski, L. Hassell, R. L. Baldwin, B. Y. Karlan, and L. Hood. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene*, 238(2):375–85, 1999.

[16] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class prediction and discovery using gene expression data. In *RECOMB*, 2000.