Scoring Genes for Relevance

Amir Ben-Dor[a]
Agilent Laboratories

Nir Friedman
Hebrew University

Zohar Yakhini
Agilent Laboratories

Abstract

Recent molecular level studies that compare different classes of disease conditions produce labeled gene expression data. We examine scoring methods that are useful in mining such gene expression data for genes that have biological relevance to the condition studied. Relevance information is useful in identifying genes driving the biological process, in selecting small subsets of genes with diagnostic potential, and in better understanding the condition studied and its relationship to known or hypothesized biochemical pathways. We present the scoring methods; describe a process for computing the corresponding p-values; and finally, present results from application to actual cancer gene expression data. These include applying classification techniques employing varying relevance based selected sets of genes.

1 Introduction

Recent studies on molecular level classification of cancer cells produced remarkable results, strongly indicating the usability of gene expression assays as diagnostic tools. [1,2,7,8] In such studies classified gene expression data is collected and analyzed. Such data consists of tissue samples for which the expression levels of thousands of genes were measured. The tissues ares labeled as belonging to certain classes (such as tumor or normal, particular kinds of tumors, phase, differentiation

Contact author. Email: amir_ben-dor@agilent.com

stage, etc). Some of the genes measured play a major role in the processes that underly the differences between the classes or are majorly 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.

Attaching a measure of relevance to each gene in such studies is useful in several ways. Seeking small sets of genes that can jointly serve as a classifier and as a basis for the development of diagnostic assays one can choose amongst the more informative genes found in preliminary more comprehensive studies. Highly informative genes that are parts of known biochemical pathways give insight into the processes that underly the differences between classes. Highly informative genes (or ESTs) of unknown function suggest new research directions.

In this paper we examine several ways of scoring genes for relevance. Assume that we are given a *data set* D, consisting of pairs $\langle x_i, l_i \rangle$, for $i = 1, \ldots, M$. Each *sample* x_i is a vector in $\mathbf{R}^N$ that describes the expression values of N genes/clones. The *label* l_i associated with x_i is either $-$ or $+$ (for simplicity, we focus on two-label classifications). We examine four scores for genes. The first, $TNoM$, is a combinatorially derived score that depends only on the vector of class labels that results from putting the expression levels $x_1(g), \ldots, x_m(g)$ in ascending order and permuting $l_1, \ldots, l_m$ accordingly. The second, $Info$, is an information-theoretic score that also depends only on that order. The third score is based on *logistic* regression and depends on actual expression values. Finally, we also briefly describe Gaussian based scores.[9]

For each one of these scores we start by presenting some underlying theory. We then turn to explaining the process of computing the scores. Numerical values for relevance scores mean a lot more if they come with statistical significance figures. In fact, comparison across sets of differing characteristics is essentially impossible without a uniform figure of merit. Section 2.4 addresses issues related to statistical benchmarking of relevance scores. Section 4 elaborates on the importance of statistical benchmarking for the analysis of data with missing values. Finally, we present results from applying these methods to actual gene expression data sets. We work with a colon cancer data set [2], a Leukemia data set[7] and a Lymphoma data set[1]. We apply scoring methods to assess the abundance of informative genes in the data. We also test the performance of the different methods in classification problems.

2 Rank Based Scoring Methods

2.1 TNoM Score

Ben-Dor *et al*[3] describe the *TNoM score* (*Threshold Number of Misclassification*). It is based on searching for a simple rule that uses a given expression level, for the given gene, to predict the label of an unknown. Formally, a rule is defined by two parameters a, and b. The predicted class is simply sign$(ax + b)$. (Note that since only the sign of the linear expression matters, we can limit our attention to $a \in \{-1, +1\}$.) A natural approach is to choose the values of a and b to minimize the number of errors:

$$Err(a, b \mid g) = \sum_i \mathit{1}\{l_i \neq \text{sign}(a \cdot x_i[g] + b)\},$$

where $x_i[g]$ is the expression value of gene g in the i'th sample. We can find the best values by exhaustively trying all $2(m + 1)$ possible rules. (Attention is limited to threshold values that are mid-way points between actual expression values.) The TNoM score of a gene is simply defined as:

$$TNoM(g) = \min_{a,b} Err(a, b \mid g),$$

the number of errors made by the best rule. The intuition is that this number reflects the quality of decisions made based solely on the expression levels of this gene.

2.2 Mutual Information Score

A shortcoming of the TNoM score is that it provides partial information about the quality of the predictions made by the best rule. Thus, for example, TNoM does not distinguish a rule that makes k one-sided errors (e.g., all the errors are tissues of class $+$ that are predicted as $-$) and a rule that makes $k/2$ errors of the first kind and $k/2$ error the second kind. This distinction is important, since the rule that makes only one-sided errors is performing quite badly in the cases that are above (or below) the threshold. In this case, we would expect the rule to have less confidence in the predictions it makes on this side of the threshold.

To make such finer distinctions, we'd like to measure the sample label *information* provided by a thresholded gene expression vector. For this purpose, we use the information-theoretic notion of *mutual information*[5]. Let X and Y be two random variables, and let $P(X, Y)$ be their joint distribution. The *mutual information* between X and Y is defined as

$$I(X; Y) = H(Y) - H(Y \mid X),$$

where $H(Y) = E[-\log P(Y)]$ and $H(Y \mid X) = E[-\log P(Y \mid X)]$ are the *entropy* of Y and the *conditional entropy* of Y given X, respectively. The mutual information can be interpreted as the number of bits we save in compressing values of Y if both the sender and the receiver of the compressed message know the value of X.[5]

In our setting, we measure the mutual information between labels and expression values using the empirical distribution induced by g and a threshold as follows. For $a \in \{-1, +1\}$ and any b set $t_{a,b}(x) = sign(ax + b)$. For appropriate l and x let $M(l, x)$ be the number of samples in D in which $l_i = l$ and $t_{a,b}(x[g]) = x$. We then define the empirical distribution $P_{D,g}$: $P_{D,g}(l, x) = M(l, x)/m$. We have thus defined two jointly distributed random variables L and $X_{g,a,b}$. To evaluate a choice of a and b, we compute the mutual information for these variables.

We note that for comparing different genes and thresholds, we can use the conditional entropy term, $H(L \mid X_{g,a,b})$, since the other term, $H(L)$, is the same for all genes we compare. Thus, to find the most informative threshold of a gene g, we want to find the parameters that minimize $H(L \mid X_{g,a,b})$:

$$\mathit{Info}(g) = \min_{a,b} H(L \mid X_{g,a,b}).$$

As with the TNoM score, we find the information score of a gene by exhaustively searching over all possible $2(m + 1)$ linear decision rules.

2.3 Logarithmic Loss

We now discuss a different derivation of the Info score. Suppose we want to predict a label given the expression value of g. One way to do this is to estimate the probability of labels given the expression level of g. That is, we seek a function $f(l \mid x)$ that represents our estimate of $P(L = l \mid X_g = x)$, where L denotes the sample label and X_g the expression level of gene g. Usually, we determine a parametric family of functions. The vector parameterization is realized by denoting the family member determined by θ as $f(l \mid x : \theta)$. To evaluate and compare different parameter settings we define the *logloss function*:

$$ll(\theta \mid L, X_g) = \frac{1}{m} \sum_i - \log f(l_i \mid x_i[g] : \theta).$$

Observe that $-ll(\theta \mid L, X_g)$ is the logarithm of the probability of obtaining the observed labeling on the measured X_g according to the model dictated by θ. Thus, minimizing the logloss function is equivalent to maximizing the *likelihood* of θ.

The choice of the parametric family determines the type of predictions we can make. One simple family of predictors employs linear thresholds:

$$f_t(+ \mid x : a, b, p, q) = \begin{cases} p & sign(ax + b) = + \\ q & sign(ax + b) = - \end{cases}$$

and $f_t(- \mid x : a, b, p, q) = 1 - f_t(+ \mid x : a, b, p, q)$. This predictor uses one coin (p) for labels when x is above the threshold, and another (q) when x is below the threshold.

For this parameterized family, the logloss is related to the conditional entropy:

Proposition 2.1 *If we use the parameterized family $f_t(l \mid x : a, b, p, q)$, then*

$$H(L \mid X_{g,a,b}) = \min_{p,q} ll(a, b, p, q : L, X_g)$$

and consequently $Info(g) = \min_{a,b,p,q} ll(a, b, p, q \mid L, X_g)$.

2.4 Distributions and p-Values

When analyzing actual gene expression data we do encounter many genes that are strongly indicative of the class of samples. One way to evaluate the significance of such results is to test them against random data. More explicitly: we want to estimate the probability of a gene scoring better than some fixed level s in randomly labeled data. 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. Genes with low p-values for which the latter two options can be ruled out are interesting subjects for further investigation and are expected to give deeper insight.

Let $\{-,+\}^{(n,p)}$ denote all vectors with n $'-'$ entries and p $'+'$ entries (the normal/cancer semantic is one possible interpretation). Let u be a vector of labels. Also let g be a vector of gene expression values. A scoring method $\mathcal{S}$ (e.g., $TNoM$, or $Info$) is a function that takes g and u and returns the score of g with respect to labeling u.

Let $U_{n,p}$ be a random vector drawn uniformly over $\{-,+\}^{(n,p)}$. The p-value of a score level s is then

$$pVal(s : g, n, p) = Prob(\mathcal{S}(g, U_{n,p}) \leq s). \tag{1}$$

Note that since $U_{n,p}$ is uniformly drawn, the order of expression values in g does not change the p-value of scores. Thus, we can assume, without loss of generality, that the values in g appear in ascending order. Furthermore, note that both the $TNoM$ score and the $Info$ score are insensitive to the actual distance between consecutive expression values of the gene. Thus, when we examine the p-value these scores, we do not need to examine the specifics of g.

The combinatorial character of TNoM makes it amenable to rigorous calculations. Ben-Dor *et al*[3] develop a recursive procedure that computes the exact distribution of TNoM scores in $\{-,+\}^{(n,p)}$. Due to space consideration we do not repeat the details here. Roughly speaking, the procedure estimates the number of permutations in $\{-,+\}^{(n,p)}$ for which the TNoM score is exactly k. This is developed into a recursive formula that involves the number of labels in $\{-,+\}^{(n-1,p)}$ and $\{-,+\}^{(n,p-1)}$ with TNoM score k and $k-1$.

The analysis of Ben-Dor *et al* does not directly extend for computing p-values for other scores, such as the $Info$ score. A seemingly simple alternative is to use stochastic simulations for evaluating p-values. Such a procedure generates random samples from $\{-,+\}^{(n,p)}$, and computes the score of each sampled labeling. Then, we can estimate the p-value of s by the fraction of samples with score smaller than s. Simple stochastic simulation procedures suffer from a serious drawback. To compute the p-value of a rare score, we need to generate a huge number of samples. Since we are interested in identifying rare genes, this renders the simple stochastic sampling impractical for our application.

Focusing our sampling in the "interesting" parts in $\{-,+\}^{(n,p)}$ can potentially overcome this problem. How can such focus be achieved? The intuition is that a reasonably good division of negative and positive labels above and below some threshold value exists in a labeling vector u that has a small $TNoM$ score. Thus such u is expected to score well with other methods, as well. Sampling from the rare TNoM scores will therefore enrich the occurrence of well scoring vectors.

To formalize this idea we start by rewriting the p-value term:

$$Prob\left(\mathcal{S}(U_{n,p}) \leq s\right) \;=\; \sum_t Prob\left(\mathcal{S}(U_{n,p}) \leq s | A_t\right) \cdot Prob\left(A_t\right) \quad (2)$$

where A_t denotes the event $[TNoM(U_{n,p}) = t]$.

Using the results of Ben-Dor *et al*, we can compute $Prob(A_t)$. To estimate $Prob\left(\mathcal{S}(U_{n,p}) \leq s | A_t\right)$ for different values of t we sample uniformly vectors from A_t and then compute the fraction of samples with score less than or equal to s. These estimated conditional probabilities, for different values of t, are combined using (2) to get an approximation of the p-values.

To apply this procedure, we need to sample from A_t. This is done recursively in a manner that follows the general lines of the recursive process for the calculation of the size of sets in $\{-,+\}^{(n,p)}$ with particular TNoM score. The details of this procedure are omitted here.

3 Smooth Scoring Methods

One of the shortcomings of the scores we examined in the previous sections is that they only allow very simple queries on the gene's expression value: whether it is above or below a specified threshold. In these methods, the prediction of the label is the same for expression values that are slightly above the threshold and for expression values that are significantly above the threshold. As a consequence, the score of a gene is determined only by the permutation of the class labels it defines. This later property helps us in efficiently computing p-values for genes' score.

Nonetheless, it seems more reasonable that the confidence in the predictions made close to the threshold value should be lower than the confidence in predictions made for genes that are significantly above (or below) the threshold value. We now examine two scoring methods that are based on this intuition.

3.1 Logistic Prediction

We would like that for expression values close to the decision threshold, the probability of both labels would be close to $1/2$. On the other hand, for extreme expression values, our prediction should be confident. That is, the conditional probability is either 1 or 0. One parametric family that allows for representing such conditional probabilities is the *logistic* family:

$$f_{logit}(+ \mid x : a, b) = logit(ax + b),$$

where $logit(z)$ is the *logistic* function

$$logit(z) = \frac{1}{1 + e^{-z}}.$$

In this family, the probability of $+$ is an sigmoid function that asymptotes to 0 and 1 at the extreme values of x. As we can easily check, the value $-b/a$ determines the point at which the probability of both labels is equal. The sign of the a parameter determines whether higher expression values are assigned higher probability of $+$ or $-$. Finally, the magnitude of a determines the slope at the threshold point.

Thus, a larger value of a implies a narrower region of uncertainty about the label.

To score a gene we need to find the parameters a and b that minimize the logloss function. We do so by gradient based non-linear optimization.[4] Although there is no analytic solution for the best parameters, we can efficiently compute the gradient of the logloss function with respect to a and b and use (conjugate) gradient descent methods to optimize the parameters.

Using logistic functions for representing conditional distributions has theoretical roots that we briefly touch on now. Suppose that the gene expression values X_g are normally distributed around a mean that depends on the type of tissue. Also suppose that the variance of the gene expression values is the same for both types of tissues. Thus,

$$P(X_g \mid l) \sim N(\mu_l, \sigma^2)$$

If we know the parameters of this distribution (μ_+, μ_-, and σ) and also the prior probability of tissue types, we can compute the probability of a tissue type given the expression value:

$$P(+ \mid X_g) = \frac{P(X_g \mid +)P(+)}{P(X_g \mid +)P(+) + P(X_g \mid -)P(-)} = \frac{1}{1 + \frac{P(X_g \mid -)}{P(X_g \mid +)} \cdot \frac{P(-)}{P(+)}}$$

Since $P(X_g \mid l)$ is a Gaussian distribution, we can write the *likelihood ratio* in an exponential form:

$$\frac{P(X_g \mid -)}{P(X_g \mid +)} = e^{-\frac{1}{2\sigma^2}(2X_g(\mu_- - \mu_+) - (\mu_-^2 - \mu_+^2))} \qquad (3)$$

Thus, the conditional probability of the label $+$ is logistic function of the expression with parameter $a = \frac{\mu_- - \mu_+}{2\sigma^2}$. As we can see, when the means are far a part in proportion to the variance, then there is a sharp transition from high confidence in one label to high confidence in the other. On the other hand, if the means are relatively close, then the transition is gradual.

This well known derivation shows that in learning logistic conditional distribution we are choosing from a parametric family that includes the ones we would have seen if we learned a normal distribution for X_g given each label. In this sense, we are learning from a class of distributions that are as rich, or richer, than the Gaussian model. Note, however, that the parameters that minimize the logloss do not allow us to reconstruct a Gaussian model of the distributions $P(X_g \mid l)$.

Also note that in learning the logistic function we do not assume that the data is distributed in a Gaussian manner. In fact, the optimal parameters are mainly determined by the expression levels at the "boundary" between positively sampled examples and negative ones. The exact expression values of at extreme tails of the distribution, that is, the examples far away from the threshold, have negligible effect on the best threshold value.

3.2 Gaussian Separation Score

The final score we evaluate is proposed by Slonim *et al* [9]. Their approach is motivated by the Gaussian model we discussed in the previous section. Given the data, they estimate the mean μ_l and standard deviation σ_l of the expression values of g among the samples labeled l. (Note, that unlike our analysis above, there can be different variances for each tissue type.) Using this Gaussian approximation, they attempt to measure to what extent the positive and negative classes are separated. This score is defined as

$$Sep(g) = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}.$$

The intuition is that the separation between the two group of expression values is proportional to distance between their mean (i.e., center points). However, this distance has to be normalized by the standard deviation of the groups. A large standard deviation implies that we expect to find points in the group far away from the mean value and thus the separation would not be strong.

Such a score is expected to work well when the data is normally distributed in each class of samples. In this case, the estimate of the standard deviation takes in to account all the data points given to us. On the other hand, if the data is not normally distributed, this score can fail. For example, an asymmetric distribution of values in one of the classes can skew the estimation of the variance a lead to misleading score.

4 Handling Missing Values

In actual gene expression data it is often the case that expression levels for some genes are not reported for some samples. This is typically due to technical measurement problems. The result is that the mixture of labels that needs to be considered is dependent on the gene in question. Obviously, a TNoM score of 0 has a different meaning for a $n = 20$ and $p = 20$ mixture then it does for $n = 20$ and $p = 5$ mixture. When selecting a subset of genes as a classification platform or when looking for insight into the studied biological process we should therefore consider the relevance of each gene in the *context* of the appropriate mixture. Absolute score values do not provide a uniform figure of merit in this context. We use p-values as a uniform platform for such comparisons, as they do depend on the mixture that defines the model. This emphasizes the importance of statistical benchmarking of relevance scores. The process is examplified in the analysis of the lymphoma data [1] below.

5 Empirical Evaluation

5.1 Description of the Data Sets

We evaluate the gene selection methods on three data sets.

Colon cancer data set. A collection of 62 expression measurements from colon biopsy samples reported by Alon et al. [2]. Of these samples, 38 are labeled “tumor” biopsies and 20 are labeled “normal”. Gene expression levels in these 62 samples were measured using high density oligonucleotide microarrays. Of the ≈ 6000 genes detected in these microarray, 2000 genes were selected based on the confidence in the measured expression levels.

Leukemia data set. A collection of 72 expression measurements reported by Golub *et al.* [7] These samples are divided to two variants of leukemia: 25 samples of *acute myeloid leukemia* (AML) and 47 samples of *acute lymphoblastic leukemia* (ALL). The source of the gene expression were taken from 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays that report on the expression levels of 7129 genes. The data is available at `http://www.genome.wi.mit.edu/MPR`.

Lymphoma data set. A collection of 96 expression measurements reported by Alizadeh *et al.* [1] Of these, 46 samples of *diffused large b-cell lymphoma* (DLBCL). The remaining 50 samples are of 8 types of tissues. Alizadeh *et al* used clustering techniques to separate the DLBCL into two classes *Germinal centre B-like DLBCL*, and *Activated B-like DLBCL*. In our experiment we used gene expression measurements of 4096 genes. This data can be found at `http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt`.

In this data set we examined two labeling systems. In the first, we want to distinguish DLBCL samples from the remaining samples. In the second, we focus only on the 46 DLBCL samples and try to distinguish thet two variants of DLBCL identified by Alizadeh *et al*.

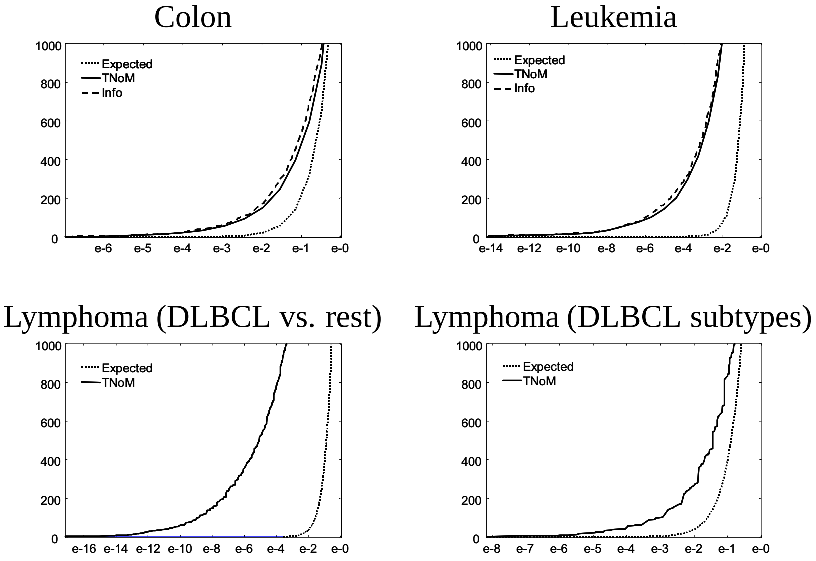
5.2 Abundance of Highly Informative Genes

Consider a set of actual labeled gene expression data, such as the sets studied by Alon *et al*[2] and in Golub *et al*[7]. It is beneficial to give some quantitative score to the abundance of highly informative genes, with respect to the given labeling. A tool for doing this may also be useful in class discovery, where an un-labeled set of data is mined for a strongly distinguishable class.

Figure 1 depicts a comparison between the expected number of genes scoring better than a given p-value score and the actual number found in the data. As we can see in all three data sets there is a large number of abundant genes. In particular, in both the leukemia data set and the DLBCL vs. rest classification problem there are many significant genes even at p-values smaller than 10^{-8}. We also see that number of significant genes for TNoM and Info is roughly similar. Although note that the significance is slightly larger for Info.

5.3 Classification

One way of evaluating the usefulness of the genes selection method is test its effect on classification accuracy. The intuition is that if we restrict ourselves to “relevant” genes, then the ability to classify examples



Colon

Leukemia

Lymphoma (DLBCL vs. rest)

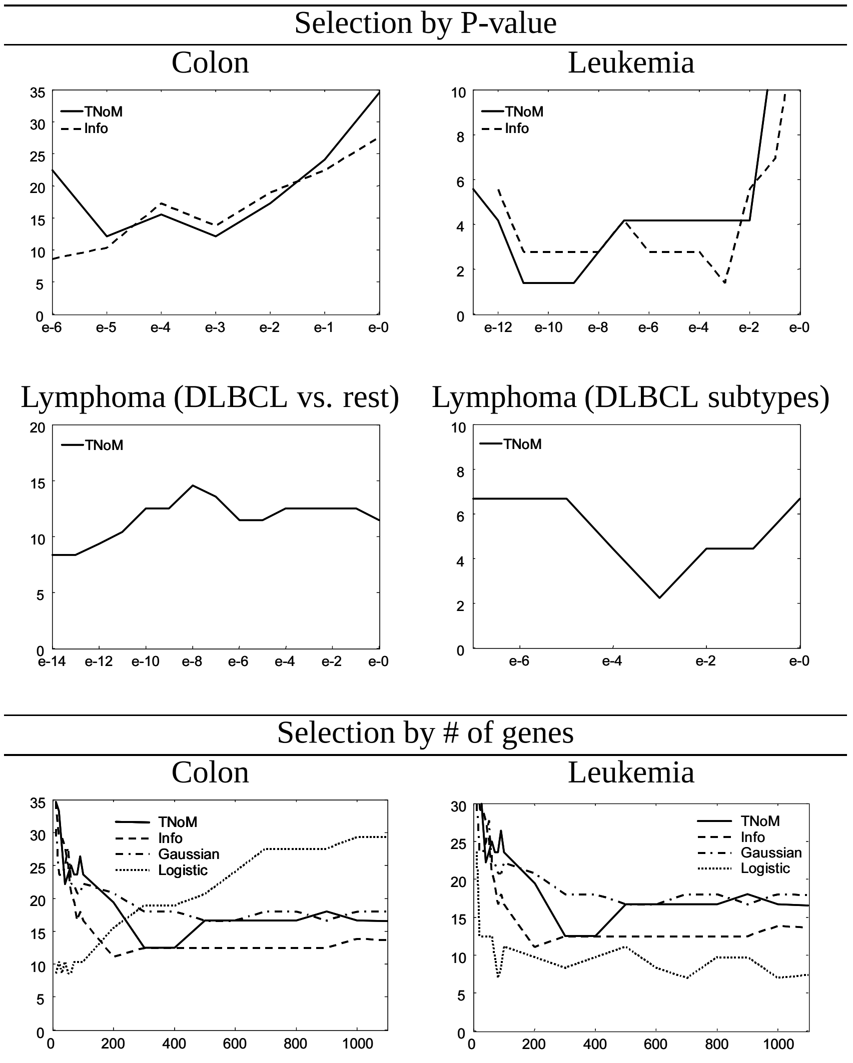
Lymphoma (DLBCL subtypes)

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 p-value and the y-axis the number of genes. The expected number of genes is the p-value multipled by the number of genes in the data set.

should not deteriorate, and might even improve. [3] In this work we use the *naive Bayesian* classifier[6] that evaluates the log-odd for the labels using the formula:

$$
\begin{aligned}
\log \frac{P(+ \mid x)}{P(- \mid x)} &\approx \log \frac{P(+)}{P(-)} + \sum_i \log \frac{P(X_{g_i} \mid +)}{P(X_{g_i} \mid -)} \\
&= \log \frac{P(+)}{P(-)} + \sum_i \left(\log \frac{f(+ \mid X_g)}{f(- \mid X_g)} - \log \frac{P(+)}{P(-)} \right).
\end{aligned}
$$

We want to evaluate the accuracy of a classification method applied to subsets of the data. In here we follow Ben-Dor *et al* 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 samples. The trained classifier is then applied to the held-out sample and



Selection by P-value

Colon

Leukemia

Lymphoma (DLBCL vs. rest)

Lymphoma (DLBCL subtypes)

Selection by # of genes

Colon

Leukemia

Figure 2: The effect of gene selection on classification. Each curve corresponds to a gene selection method. In the top figures, the x-axis denotes the p-value threshold for selection. In the bottom figures, the x-axis denotes the number of genes selected. In all figures, the y-axis denotes the percentage of incorrect classifications based on LOOCV. The classification is performed by a *naive Bayesian classifier* (see text). The gene selection methods are based on TNoM, Information score, Logistic score and Gaussian Separation score. The latter two methods were only applied when using # of genes, since we did not compute p-values for genes.

the predicted label is compared to the true label. The fraction of errors when we repeat this for all samples is our estimate of the error rate of the classification procedure.

When using gene selection, we need to pre-process the training data to select genes. Then, the classification procedure is applied using the training data restricted to the subset of selected genes. To evaluate performance with gene selection, we have to be careful to jointly evaluate both stages of the process: gene selection and classification. To do so, we have to apply gene selection in each cross-validation trial on on the training examples of that trial. Note, that since the training examples are different in different cross validation trials, we expect the number of selected genes to depend on the trial.

After running this procedure for each sample, report are the fraction of successful predictions in all LOOCV trials. We run this LOOCV procedure using several p-value thresholds for both the $TNoM$ score and the $Info$ score. For the logistic and Gaussian Separation score we select genes by number. Thus, in each LOOCV iteration, we selected the k-best scoring genes. Figure 2 show how the the performance of the classification procedures changes with the p-value threshold/number of selected genes and the scoring method.

These results show several interesting trends. First, as we can see, selecting relevant genes does, in general, improve classification performance. Moreover, there is a wide range of p-values in which we obtain good classification accuracy. Thus, the process is not too sensitive to the exact p-value employed. In almost all tested cases, setting the threshold p-value in the range 10^{-3} to 10^{-5} seems to give good performance. For the Leukemia data set, setting p-value of 10^{-9}, we select, on average, 16 genes per LOOCV trial with TNoM. For this threshold, the voting classifier makes only 1 error out of 72 samples (%98.5 accuracy). We get similar results on the DLBCL subtype classification problem with p-value of 10^{-3}. In this case, the TNoM score selects, on average, 102 genes per LOOCV trial. It is interesting to note that the performance of TNoM and Info score on the leukemia data set degrades when we select a fixed number of genes rather than by p-value. This examplify the claim p-values provide is a more robust approach for selecting genes.

6 Conclusion

In this paper, we examined the problem of identifying relevant genes in labeled gene expression data. We described several approaches for selecting genes and a general procedure for efficiently estimating p-values.

Our analysis shows that relevant genes are significantly abundant in actual gene expression data. We also demonstrate that by restricting classification rules to examine these genes, performance improves, often dramatically. We are currently extending our analysis to other labeled gene-expression data as it becomes available. We are currently studying more direct approaches to the selection of informative *sets* of genes. Identifying sets of genes that give rise to efficient learned classifiers might reveal previously unknown disease related genes and guide further biological research.

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, 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, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 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. Nat. Acad. Sci. USA*, 96:6745–6750, 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*, 2000. To appear. A preliminary version appears in RECOMB 2000. See http://www.cs.huji.ac.il/∼nir/publications.html.
4. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford

University Press, Oxford, U.K., 1995.

5. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
6. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
7. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, J.P. Mesirov M. Caasenbeek, 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.
8. J. Khan, R. Simon, M. Bittner, Y. Chen, S. B. Leighton, T. Pohida, P. D. Smith, Y. Jiang, G. C. Gooden, J. M. Trent, and P. S. Meltzer. Gene expression profiling of *Alveolar rhabdomyosarcoma* with cDNA microarrays. *Cancer Reasearch*, 1998.
9. 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.