Published in final edited form as:

Stat Appl Genet Mol Biol. 2014 August ; 13(4): 477–496. doi:10.1515/sagmb-2013-0053.

Multiclass cancer classification based on gene expression comparison

Sitan Yang^{*} and Daniel Q. Naiman[†]

^{*}Department of Applied Mathematics and Statistics, Johns Hopkins University, syang45@jhu.edu

[†]Department of Applied Mathematics and Statistics, Johns Hopkins University, daniel.naiman@jhu.edu

Abstract

As the complexity and heterogeneity of cancer is being increasingly appreciated through genomic analyses, microarray-based cancer classification comprising multiple discriminatory molecular markers is an emerging trend. Such multiclass classification problems pose new methodological and computational challenges for developing novel and effective statistical approaches. In this paper, we introduce a new approach for classifying multiple disease states associated with cancer based on gene expression profiles. Our method focuses on detecting small sets of genes in which the relative comparison of their expression values leads to class discrimination. For an m -class problem, the classification rule typically depends on a small number of m -gene sets, which provide transparent decision boundaries and allow for potential biological interpretations. We first test our approach on seven common gene expression datasets and compare it with popular classification methods including support vector machines and random forests. We then consider an extremely large cohort of leukemia cancer to further assess its effectiveness. In both experiments, our method yields comparable or even better results to benchmark classifiers. In addition, we demonstrate that our approach can integrate pathway analysis of gene expression to provide accurate and biological meaningful classification.

Keywords

Multiclass cancer classification; Biomarker discovery; Gene expression analysis

1 Introduction

In recent years, microarray-based gene expression profiling has become a widespread approach for identifying biomarkers associated with cancer. In particular, expression patterns of genes have been sought extensively through statistical learning techniques to classify molecular subtypes, predict clinical outcomes and chemotherapy responses (Quackenbush, 2006). This has resulted in a proliferation of such methods introduced and developed in the literature (Statnikov et al., 2008). While many early studies have focused on distinguishing two disease classes, relatively few approaches have been designed specifically for classification in presence of multiple tumor classes (see e.g., Statnikov et al., 2005). In fact, as the heterogeneity of cancer has become more clear in recent studies (Burgess, 2011), new cancer subtypes are expected to continue to be discovered, leading to a

growing number of multiclass problems. Moreover, clinical experiments for investigating tumors in stage, grade, survival time, and drug sensitivity are also likely to produce multiclass microarray datasets, see e.g., Dyrskjot et al. (2007) and Shah et al. (2011). Therefore, there is an increasing need for developing multiclass methods.

However, the methodological development of gene expression classifiers often suffers from two limitations. First, the sample size available for most microarray datasets is small, but at the same time, these involve a large number of gene transcripts, leading to what is commonly referred to as the “small n , large p ” dilemma. As a result, classification performance of complex models can be degraded by the large variance resulting from parameter estimation, and it is often necessary to restrict attention to classifiers of limited complexity. Second, although well-established and advanced machine learning techniques such as support vector machines (Cortes and Vapnik, 1995) and neural networks (Khan et al., 2001) can be immediately introduced as gene expression classifiers, their decision rules in most cases behave as “black boxes” that do not lend easily to biological mechanistic understanding.

To address these limitations, many statistical methods have been proposed. Tibshirani et al. (2002) developed “Prediction Analysis of Microarrays” (PAM) that modifies the diagonal linear discriminant analysis method by introducing a shrinkage parameter, which creates a “de-noised” version of the class centroids (i.e., mean expression levels of classes) used in the discriminant function. Also, Grate (2005) investigated the discriminatory power of small gene subsets. Each gene set with size three or less is analyzed as a candidate for constructing a parameterized linear hyper-plane for distinguishing cancer classes, which is similar to the traditional separating hyper-plane classifiers. In addition, Leban et al. (2005) proposed the “VizRank” method that focuses on visualizing different cancer classes through data projections, which was further used by Mramor et al. (2007) for cancer classification. In general, these methods provide simplified decision rules that achieve comparable classification performance to traditional techniques.

One approach attempting to take both classifier complexity and biological interpretability into consideration was proposed by Geman et al. (2004). Here, a concept (later called “Relative Expression Analysis” in Eddy et al., 2010) was introduced to construct classifiers using the relative orderings (instead of raw values) of gene expression within each sample. In view of extensive preprocessing required for gene expression data, these relative orderings seem to be reliable pieces of information: they are likely to be preserved under slight perturbations of gene expression values and are robust against effects that shift expression values in the same direction. For example, they have been proved (Lin, 2008) to be invariant under commonly used preprocessing techniques such as convolution and quantile normalization of RMA (Irizarry et al., 2003). Based on this concept, “Top Scoring Pair” (TSP) was introduced as a new binary classification approach by simply comparing expression levels in one or more pairs of genes (i.e., top scoring pairs) for class prediction (Figure 1). As shown by Geman et al., the TSP approach provides transparent but powerful decision rules that compete with many sophisticated machine learning methods. In addition, gene pairs selected by TSP in various subsequent studies have been found to be biologically informative, see, e.g., Edelman et al. (2009), Zhao et al. (2010) and Patnaik et al. (2010).

In the multiclass case, cancer classification becomes considerably more challenging. First, the “small n , large p ” problem is especially compounded when subdivision of an already small set of samples into subclasses leads to dramatically smaller sample sizes for subclasses. Second, multiclass methods typically require significantly more computation, and decision rules generated can become substantially more complex (Shen and Tan, 2006). Many classifiers developed in binary problems do not naturally apply to the multiclass case and have to rely on decomposition of the problem into many binary sub-classification problems, together with an aggregation scheme for combining various sub-classifications, which is likely to increase the computation time and decrease the interpretability of the final decision rule.

Motivated by these considerations, in this paper, we introduce a new approach called “Top Scoring Set” (TSS) for multiclass cancer classification based on gene expression microarrays. TSS is a generalization of the TSP classifier in the multiclass case. It is parameter-free, purely data driven and robust to some common microarray preprocessing transformations. For an m -class problem, the class prediction is determined by a relatively small number of m -gene sets, namely, top scoring sets. Each top scoring set votes for a class based on the ordering of expression levels of its genes. The final prediction is the class that receives the majority of votes. In principle, TSS makes specific statistical hypotheses about gene expression comparison that could have biological interpretations, and even without the potential interpretability, the decision rule itself can be easily appreciated by non-specialists.

An example of a TSS classifier is illustrated in Figure 2, here we consider a more difficult task than that in Figure 1 to distinguish three cancer subtypes of the leukemia data in Golub et al. (1999). Figure 2 depicts a top scoring set that has been found consisting of gene *PLCB2*, *MB-1* and *LCK*. Class prediction for a particular sample is determined by the gene whose expression level in this set is the highest. As shown later, TSS yields 95.83% prediction accuracy on this dataset using leave-one-out cross-validation (see Methods). To demonstrate the effectiveness of our approach, we evaluate its predictive performance on seven common human cancer gene expression datasets and compare it with popular benchmark classifiers including PAM, support vector machines (SVMs) and random forests (Breiman, 2001). In most cases, TSS has achieved comparable or better classification performance. Moreover, we validate TSS on an extremely large multiclass cohort of leukemia cancer (Haferlach et al., 2010) containing 14 subclasses, and its predictive ability is demonstrated to compete with that of an large ensemble of SVMs.

There is no question that some gene expression patterns that are potentially useful for classification may be dismissed by TSS, and the assumptions under which TSS would likely prove the most useful might seem overly simplistic to reflect biological conditions in complex diseases. However, TSS provides a practical attempt for modeling the statistical dependency structure among genes given the amount of data. Also, as multiple top scoring sets are often found on a particular dataset, our results demonstrate that the information in the ordering of gene expressions is sufficient to reliably perform classification. Furthermore, we will show that TSS can also integrate biological information from functional pathway analysis of genes. Publicly available databases such as the Kyoto Encyclopedia of Genes

and Genomes (KEGG) in Kanehisa et al. (2004) can be used by TSS to provide accurate and biologically meaningful classification.

2 Methods

Let us consider G genes measured using DNA microarrays and their expression levels $\mathbf{X} = \{X_1, X_2, \dots, X_G\}$ regarded as a random vector. Each observed gene profile x is a realization of $\mathbf{X}$ and has a true label y representing its class. A microarray dataset is a collection of many, say N , observed gene profiles and can be represented as a matrix $\{x_{ij}\}$ with G rows of genes and N columns of samples (typically $G \gg N$). In this section, we first provide a brief review of the TSP method. Then the TSS approach is introduced in the multiclass setting. In addition, we discuss some implementation details of TSS.

2.1 A short review of TSP

As discussed in Geman et al. (2004), for a two-class problem (with classes denoted by 1 and 2), TSP aims to find each “marker” gene pair (i, j) ($i, j \in \{1, 2, \dots, G\}$) that has a simple relation whose probability distribution changes significantly from one class to the other. The simple relation considered here is the comparison between the expression levels of gene i and j , and a highly relevant quantity of interest is the conditional probability $P(X_i > X_j | y)$ where y is the class variable, $y \in \{1, 2\}$. So if $P(X_i > X_j | y = 1)$ is high while $P(X_i > X_j | y = 2)$ is low, it will be very likely to observe $X_i > X_j$ in class 1 but not in class 2 where $X_i < X_j$ is more likely to happen. As a result, this property of (i, j) leads to the ability to distinguish between two classes simply by determining the gene having the higher expression value, a simple decision rule for predicting class labels. In TSP, a score is defined for each distinct gene pair (i, j) as $|P(\hat{X}_i > \hat{X}_j | y = 1) - P(\hat{X}_i > \hat{X}_j | y = 2)|$ in order to estimate the probability change from class to class where $P(\hat{X}_i > \hat{X}_j | y)$ is the frequency observed from the data. The ones that achieve the highest score among all possible gene pairs (i.e., top scoring pairs) are involved in the decision rule. For a top scoring pair that has $P(\hat{X}_i > \hat{X}_j | y = 1) > P(\hat{X}_i > \hat{X}_j | y = 2)$, it predicts the class label $\hat{y}$ of a new sample x as

$$\hat{y} = \begin{cases} \text{class 1, if } x_i > x_j, \\ \text{class 2, if } x_i < x_j. \end{cases} \quad (1)$$

Then the predictions for each class are summed up over all top scoring gene pairs, and the majority rule is applied to produce the final prediction. From (1), we can see that the decision rule of TSP is only based on simple comparisons of gene pairs. However, as mentioned earlier, it has been shown as an effective classifier on many cancer datasets, and some top gene pairs from these studies are shown to be informative. In addition, several extensions of TSP have also been developed. Xu et al. (2007) considered the average ranks in two groups of genes (rather than a pair of genes) for constructing the decision rule. Tan et al. (2005) introduced the k -TSP classifier where the top k scoring pairs are involved using the majority rule in the decision process. Also, Lin et al. (2009) proposed the “Top Scoring Triplet” method in which relative orderings in each triplet (i.e., three genes) are investigated using a similar approach in TSP. Recently, Kaur et al. (2012) introduced the “ProtPair”

method that uses TSP for human disease prognosis based on protein expression data. Thus far, all of these derivations have been aimed at the binary classification problem.

2.2 Top Scoring Set

In this section, we introduce TSS as a new multiclass classification approach. The motivation of TSS comes from the relative comparison idea used in TSP. As discussed in Geman et al. (2004), such relative comparison of mRNA concentrations indicated by gene expression levels provides a natural link with biochemical activity, and proposes concrete hypotheses for a small list of genes. Therefore, our goal here is to discover valuable information for multiclass separation by comparing expression patterns of a few genes. In particular, for an m -class problem (with classes denoted by $1, 2, \dots, m$), we are interested in finding m “marker” genes $\mathcal{S} = \{i_1, i_2, \dots, i_m\} \subset \{1, 2, \dots, G\}$ where the presence of some simple relations among these genes with high conditional probability depending on the class leads to class separability. Specifically, we aim for the high expression level of gene i_c relative to the other $m - 1$ genes in $\mathcal{S}$ being indicative of a sample coming from class c . To be precise, the desired statistical property for $\mathcal{S}$ is that $\forall c \in \{1, 2, \dots, m\}$

$$P[\arg \max \{X_r, r \in \mathcal{S}\} = i_c \mid y=c] \gg P[\arg \max \{X_r, r \in \mathcal{S}\} = i_c \mid y \neq c]. \quad (2)$$

In other words, gene i_c is much more likely to be the gene among the m genes in the set that has the maximum expression level in $\mathcal{S}$ for class c than for any other class. In this case, a classification rule can be constructed by determining which gene is most expressed in $\mathcal{S}$ with a simple “arg max” function. Therefore, it is essential to find gene sets satisfying (2) with the greatest possibility. For this purpose, we define a score for each m -gene set to estimate its probability of holding (2). The sets with the highest score are hence referred to as the top scoring sets and will be used for classification.

In general, TSS searches for gene sets exhibiting a particular pattern that may be suitable for classification. There are, of course, many other patterns that one might consider with the potential for effective classification. Still, it is important to be mindful that increasing the size of the pattern search space would result in significant increases in already substantial computational costs, and is more likely to produce over-fitting.

2.2.1 Gene set score—To illustrate the score calculation for gene sets, we start with a previous example of the leukemia data in Golub et al. (1999), which consists of 7,129 genes and three leukemia subtypes identified as AML, B-ALL and T-ALL, with 25, 38 and 9 samples respectively. For this three-class problem, we score a particular gene set consisting of *PLCB2*, *MB-1* and *LCK*. As described in Figure 2, this is a top scoring set identified by TSS. Here, we denote three genes as i_1 , i_2 and i_3 respectively, and we calculate the observed class conditional frequencies of their expression comparison in Table 1.

Interestingly, we observe that for AML, gene i_1 has the highest expression level among three genes in 100% of the samples, as indicated by the first column of Table 1. Similarly, for B-ALL, gene i_2 has the highest expression level in 97.4% of the samples, and for T-ALL gene i_3 has the highest expression level for 100% of the samples. Therefore, based on the

information provided by these three genes, a natural way to classify a sample with expression levels x , would be to predict AML, B-ALL, and T-ALL respectively by determining which of the three expression levels x_{i_1} , x_{i_2} , and x_{i_3} is highest. Accordingly, we define a score for $\{i_1, i_2, i_3\}$ based on Table 1 as the sum of the row maxima, i.e.

$$\max\{1, 0, 0\} + \max\{0, 0.9737, 0\} + \max\{0, 0.0263, 1\} = 2.9737.$$

If the score is 3, clearly, the rule described above obtains a zero apparent error rate on the dataset. Furthermore, if the underlying probability distributions of gene expression comparison are well reflected by the observed frequencies, a higher score can indicate that the rule is more likely to be effective for new samples. Therefore, our goal is to search for three-gene sets with the highest possible score.

In general, for an m -class problem, each m -gene set $\mathcal{S} = \{i_1, i_2, \dots, i_m\}$ can produce a similar table (Table 2) where p_{rj} is the frequency that given class $y = j$, gene i_r has the highest expression level in $\mathcal{S}$. The score for $\mathcal{S}$ is then defined as

$$\sum_{r=1}^m \max_{j=1,2,\dots,m} (\hat{p}_{rj}). \quad (3)$$

Based on some assumptions, the equation (3) has a Bayesian decision-theoretic interpretation, where a Bayes optimal rule is chosen among a set of possible decision rules by minimizing the Bayes risk. To define the Bayes risk, one must introduce a prior distribution for the classes and a loss function specifying the penalty for each misclassification. In the absence of class priors and information about relative losses associated with various types of misclassification, it is natural to use 0-1 loss, and assume equal prior probabilities for classes. The resulting Bayes rule turns out to be (3). However, different loss functions or class priors can also be considered. For example, the empirical class prior n_c/N , where n_c is the sample size of class c and N is the total sample size. Details of the interpretation and the optimum Bayesian classifier with a general loss function and class priors can be found in Appendix A.

In practice, for further breaking ties among gene sets with the highest score, we also considered a secondary score based on Table 2 as

$$\sum_{j=1}^m \sum_{r=1}^m (-\hat{p}_{rj}) \ln \hat{p}_{rj}, \quad (4)$$

i.e., the sum of estimated class conditional entropies. As a result, each top scoring set that is finally chosen is required to minimize the secondary score (4) as well.

2.2.2 Decision rule—As mentioned earlier, the TSS classifier is built based on top scoring sets. For each top scoring set $\tilde{\mathcal{S}}$, the prediction for sample x is

$$\hat{y} = \arg \max_{c=1,2,\dots,m} \{x_{i_c}, i_c \in \hat{\mathcal{S}}\}. \quad (5)$$

Here we used the same notations as in (1). We can see that when $m = 2$, (5) turns out to be (1). Therefore, TSS is essentially a generalization of TSP in the multiclass case.

Although it rarely happens, due to expression level ties the decision rule for a single top scoring set can produce multiple classes associated with genes whose expression is the highest. In this situation, we consider a randomized decision where each associated class is assigned a vote of $1/T$, where T is the number of genes producing the tie. For the final prediction, these votes are summed over the top scoring sets and the majority rule is applied.

2.2.3 Greedy search—In theory, TSS finds top scoring sets among all possible gene sets. For a dataset with G genes and m classes, an exhaustive search has the complexity $O(G^m)$, which grows exponentially with respect to G . Hence, it is quite necessary to relax the global optimality requirement, which can be done in various ways. One idea would be to a priori reduce the search space. This can be typically done by pre-selecting a small number of genes based on a univariate multiclass criterion such as one-way ANOVA or Kruskal-Wallis. Also, as illustrated in the “Results” section, it is possible to use pathway information to restrict the search within naturally defined groups of genes. In this section, however, we propose a different idea that adopts a greedy search algorithm to select gene sets that are top scoring in each of several stages, leading to what could be called *locally optimal* scoring sets.

For an m -class problem, the greedy search algorithm takes $m - 1$ steps to form m -gene sets that are used in the final decision rule. It is initialized by finding the collection of gene pairs

with the highest score for each of possible $\binom{m}{2}$ two-class (1-vs-1) sub-problems. Next, each possible two-class sub-problem is augmented by a single class, and for every such augmentation, a collection of three-gene sets with the highest score is found based on top scoring gene pairs obtained in that two-class sub-problem. In particular, a distinct gene is added to each of such gene pairs to yield a group of three-gene sets, among which the ones with the highest score are sought. Then, the algorithm is performed iteratively until the size of sub-problems reaches m , and a collection of m -gene sets with the highest score is obtained from each sub-problem of size m . Finally, the (locally optimal) top scoring sets are found among all such collections for building a TSS classifier.

The greedy search process is illustrated in Figure 3. Since the first step involves two classes and each subsequent step deals with one more class, the formation of an m -gene set requires $m - 1$ steps. Importantly, all possible sequences of classification problems in which we start with a two-class problem and augmenting by one class at a time until we reach m classes are considered, so that ultimately, we arrive at $m!/2$ collections of m -gene sets to be compared. The complexity of the first step is $O(G^2)$, and each following step only requires $O(G \cdot l)$ additional computations where l is the maximum size of collections of gene sets with highest scores generated in the previous step. Because the number such sets for a certain sub-

problem is expected to be small, l is typically small so that the fully implemented algorithm has $O(G^2)$ complexity, which is significantly lower than the $O(G^m)$ complexity of an exhaustive search for $m > 2$.

The TSS classifier built by the greedy algorithm is typically validated by cross-validation. Normally, the greedy algorithm is assumed to be performed in each iteration of the cross-validation loop, which can lead to relatively extensive computation. To address this difficulty, we have further developed an acceleration algorithm that extends the pruning algorithm introduced by Tan et al. (2005) for the TSP classifier to the multiclass case (see Appendix B). The acceleration algorithm applies the greedy search method only one time on the entire dataset and generates a small list of gene sets. Then, the top scoring sets identified from the list in each iteration of the cross-validation are guaranteed to be the same as those obtained by applying greedy search on the reduced training set.

The greedy algorithm is not likely to find gene sets having the globally maximum score, but since it is based on iterating through gene sets with highest scores in each possible sub-classification problem. The resulting efficiency gain is substantial enough to compensate for the fact that the space in which the search is carried out is limited, and produces high-scoring gene sets for the original problem. Also, since all possible sizes of sub-problems are investigated sequentially and only gene sets with the highest score are kept in each iteration, the final top scoring sets can be globally optimal if the global solution is also optimal in each of these sub-problems. In particular, this happens when the top scoring sets obtained by the algorithm have a perfect score.

3 Results

3.1 Gene expression data

The TSS classifier introduced in this paper has been tested on seven human gene expression datasets retrieved from public databases or authors' websites (Table 3). They are related to human cancers including leukemia (Leukemia), mixed lineage leukemia (MLL), lung adenocarcinoma (Lung), small round blue cell tumors (SRBCT), bladder carcinoma (Bladder), childhood acute lymphoblastic leukemia (ChildALL) and non-small cell lung cancer (NSCLC). The types of problems range from classification of cancer subtypes (Leukemia, SRBCT and NSsCLC), tumor stages (Lung and Bladder) and treatment responses (ChildALL). Some of these datasets have been investigated previously (Tan et al., 2005 and Mramor et al., 2007) for evaluating gene expression classifiers. Additional information can be obtained from the references included in Table 3.

3.2 Top scoring gene sets

The top scoring gene sets found on seven datasets are summarized in Table 4. For each dataset, top scoring sets have been obtained by applying the greedy search algorithm on all samples. The corresponding top score and re-substitution error have also been calculated. The algorithm has been implemented in R 3.0.0 (<http://www.r-project.org/>) using the package *Rcpp* (<http://cran.r-project.org/web/packages/Rcpp/>) and the codes are available at our website (http://jshare.johnshopkins.edu/dnaiman1/public_html/tss).

We have identified 73 top scoring sets for Leukemia, seven sets for MLL, three sets for Bladder, two sets for SRBCT and NSCLC, and one for Lung and ChildALL. Only a few genes are actually involved in these sets and most genes appear in multiple sets. It is interesting to note that some of these genes would not be regarded as differentially expressed based on their individual expression values, but the relative comparison of expression levels in each top scoring set produces enhanced class separability. The observed frequencies of some top scoring sets are displayed in Table 5. For each set, the table gives the relative frequency at which the maximum expression value appears among genes in the set for every class. In each case, these relative frequencies provide good evidence for discriminability of the set, which indicates the potential for class prediction.

3.3 Classification accuracy

To validate the greedy search-based TSS (G-TSS) in the previous section, we assessed its classification accuracy on seven microarray datasets. As a comparison to the greedy search algorithm, we considered a common differential expression technique based on the Mann-Whitney test and the “1-vs-all” strategy to select top n genes for separating each class from the union of other classes. To save computation time, n was chosen to be 50 for three-class and 25 for four-class problems. The resulting TSS classifier (denoted as “MW-TSS”) is compared to G-TSS in terms of classification accuracy. Furthermore, we considered five popular machine learning techniques as benchmarks to the TSS approach: k-nearest neighbors (kNN), naive Bayes (NB), random forests (RF), support vector machines with a linear kernel (l-SVM) and PAM. All analyses have been performed using packages in R 3.0.0. LIBSVM (Chih-Chung and Chih-Jen, 2011) was used as the implementation for SVMs. There are a variety of model choices provided by LIBSVM and the linear kernel SVM is suggested for microarray data. In particular, multiclass problems are handled by LIBSVM using the “1-vs-1” approach.

The accuracy of a classifier has been estimated using leave-one-out cross-validation (LOOCV), a common procedure to evaluate classifiers on datasets with a small sample size (see e.g., Geman et al., 2004 and Lin et al., 2009). For classifiers with parameters (e.g., the number of nearest neighbors k in kNN and the cost factor C in l-SVM), the performance evaluation was realized by a double LOOCV loop, in which the inner loop is responsible for model optimization that usually involves parameter tuning, and the outer loop is used for calculating accuracy by averaging classification results. To avoid over-optimistic evaluation results, each step of the outer loop is carried out so that the training data on which the model optimization is performed is fully independent of the left out testing sample.

Table 6 provides a comparison of classification accuracies for different methods on seven datasets. In general, G-TSS has achieved comparable or better performance on most datasets. For Leukemia and MLL, it competes with the highest accuracies obtained by PAM and NB respectively. For Bladder, ChildALL and NSCLC, it turns out to be the most accurate classifier. In contrast, MW-TSS only yields comparable results on Leukemia and NSCLC and has the lowest accuracies on four out of seven datasets, which seems to indicate the inappropriateness of using traditional differential expression methods that focus on

individual expression values to search for relative expression patterns. These results demonstrate the superiority of the greedy search algorithm for building the TSS classifier.

In our study, kNN, NB, RF and l-SVM make use of all available genes for classification. Although their performances could often be improved by feature selection or ensemble approaches, investigation of such improvements is beyond the scope of this paper as our goal is not to merely develop a more accurate classifier. Instead, the competitive performance of TSS across all datasets demonstrates its stability. More importantly, our approach is able to discover small informative subsets of genes, and the TSS decision rule, compared with those of benchmark classifiers, proves to be much simpler, hence is more likely to provide for improved biological interpretability without a concomitant sacrifice in performance.

3.4 Leukemia study

While the predictive ability of the TSS classifier has been demonstrated across seven gene expression datasets, these datasets generally have a very limited sample size and a small number of classes. To address these limitations, we applied the TSS approach to one extreme case: the Microarray Innovations in Leukemia (MILE) study program (Haferlach et al., 2010). MILE is claimed as one of the largest gene expression microarray profiling study in hematology and oncology. The expression profiles are collected from 11 laboratories in seven countries across three continents and consist of leukemia subtypes of myeloid and lymphoid malignancies. MILE is a two-stage study where a retrospective stage I generated expression profiles for 2,143 patients and was designed for biomarker discovery. A prospective stage II produced an independent cohort of 1,152 patients and was used for validation. Stage I used commercially available whole-genome microarrays (Affymetrix HG-U133 Plus 2.0) and stage II was performed using a newly designed custom chip (Roche AmpliChip). The microarray data have been deposited in Gene Expression Omnibus database (<http://www.ncbi.nlm.nih.gov/geo/>) under series accession number GSE13204.

MILE provides an unique opportunity for validating microarray-based classification models, especially for multiclass approaches. Each of 2,143 samples in stage I contains 54,675 gene expression measurements (45 missing values). Samples are classified into 18 diagnostic *gold standard* categories including eight ALL subtypes, six AML subtypes, two chronic leukemia subtypes, myelodysplastic syndromes and normal bone marrow. Stage II contains only 1,480 (1,457 disease-related and 23 housekeeping) genes and samples are also classified into 18 classes as defined in stage I. In the initial MILE study, a classification model was trained and tested for distinguishing all 18 classes. The multiclass model consists of binary classifiers formed by support vector machines with a linear kernel (l-SVM), each of which separates a pair of classes. In the results, high accuracies were observed for most classes, indicating the robustness of microarray-based classification. To compare the predictive performance, we trained a classification model using the TSS approach on stage I samples and tested it independently on samples from stage II. Since in the original MILE paper, the independent validation results were only shown for an acute leukemia diagnostic classifier, all 14 acute leukemia subtypes (Table 7) are considered in our study. Also, both training and

test set contain only 1,457 genes that are in common for microarray datasets from two stages.

Although in principle the TSS approach can be applied to any number of classes, the predictive power of top scoring sets through relative comparison is expected to decrease as the number of classes under study increases. Therefore, for this large multiclass problem, a two-step decision tree as shown in Figure 4 was introduced based on three TSS classifiers. The hierarchy of the tree is generated from the structure of the data where 14 acute leukemia subtypes can be grouped into three major lineage leukemias (B-ALL, T-ALL and AML). The B-ALL class is further divided into seven subtypes while the AML class contains six subtypes. As a result, three TSS classifiers were built for a three-class, six-class and seven-class problem respectively. The final prediction for a sample would follow the decision tree. In addition, to further improve the predictive performance, we considered a similar procedure as introduced by Tan et al. (2005) to construct an ensemble of TSS classifiers for each of three multiclass problems. Specifically, the top k scoring gene sets were selected at each step of the greedy search process and the final prediction was the class receiving the majority of votes from k chosen gene sets. k was considered as the model parameter and the best $k \in \{1, 2, \dots, 50\}$ was determined by LOOCV on the training set. Also, the acceleration algorithm was used to expedite the cross-validation process (the case for top k scoring sets is discussed in Appendix B).

Prediction accuracies of G-TSS are shown in Table 8 and are compared to those achieved by l-SVM as presented in Haferlach et al. (2010). The optimal k for the ensemble of TSS classifiers in the three-, six- and seven-class problem (Figure 4) is 12, 7 and 42 respectively. G-TSS has achieved 100% correct predictions for two classes (C1 and C6), and $> 90\%$ accuracies for three classes (C4, C5 and C10). It outperforms l-SVM in three classes, and yields equal results in four classes. In general, there are at least 10 out of 14 classes in which comparable accuracies have been observed for both methods. For G-TSS, low accuracies are mainly observed for C8, C12 and C13. For C12, its intrinsically heterogeneous nature has been discussed in Haferlach et al. (2010). For C8 and C13, this could be due to the imbalanced training sample sizes that may violate the equal prior probabilities for classes assumed by TSS. In addition, as suggested in the confusion matrices (see Appendix C) for these three classification problems, some of the poorly scoring classes are actually being classified as closely related subclasses. For example, 28 samples in C8 are misclassified as the closely related C3. In general, G-TSS uses only three TSS classifiers as compared to the

multiclass model that contains $\binom{14}{2} = 91$ SVM classifiers in the MILE study. Many fewer genes are involved in making predictions through G-TSS and the decision process is transparent and potentially interpretable.

3.5 Pathway-based classification

As it is well recognized that many functional related genes are typically involved in the mechanism of complex diseases such as cancers, one popular approach of gene expression analysis is to investigate these naturally defined sets of genes rather than all the genes at once. Pathway-based classification using expression profiles has been shown in recent

studies (Gatza et al., 2010 and Kim et al., 2012) to provide results that are more biologically meaningful. In this section, we will demonstrate the ability of TSS to integrate biological information from traditional pathway analysis for cancer classification.

Pathways are collections of related genes and they can be ranked according to their discriminability for phenotypes under study, a process which is often referred to as enrichment analysis. The pathways identified through enrichment analysis can be useful in a variety of ways. In particular, a natural and efficient feature selection for TSS would be to restrict the search for top scoring sets to genes within those significant pathways recognized in the enrichment analysis step. One advantage of integrating TSS with pathway analysis is to detect subtle but consistent changes in expression of a small group of functional related genes, which may not lead to statistically significant results in a conventional univariate analysis based on all genes.

The pathway-based TSS was tested on Leukemia using the same protocol for training (38 samples) and testing (34 samples) in the original paper (Golub et al., 1999). Pathway information was collected from KEGG (Kanehisa et al., 2004), but other databases such as BioCarta (<http://www.biocarta.com/>) or Broad Institute (<http://www.broad.mit.edu/gsea/>) are also applicable. The enrichment method used was the GSA approach proposed by Efron and Tibshirani (2006) that is based on the well-known GSEA procedure (Subramanian et al., 2005) and also enables multiclass analysis.

The Bioconductor package (Gentleman et al., 2004) in R was used and 226 KEGG pathways were found on Leukemia. Table 9 lists all differentially expressed pathways identified by GSA when the false discovery rate threshold is set to default value of 0.2. We built a TSS classifier by searching within each of these pathways for three-gene sets with the maximum score. The final classifier contained five top scoring sets with the score of 2.94. The classification result on the independent testing set was compared to that of benchmark classifiers in Table 10.

The pathway-based TSS achieved the highest accuracy, and it used the smallest number of genes. The final decision rule consists of five top scoring sets where genes come from two pathways, namely, primary immunodeficiency and hematopoietic cell lineage, both of which are statistically significant with zero p -value using GSA, and they also have interesting biological relevance. The former one relates to disruption of the cellular immunity observed in patients with defects in T cells or both T and B cells, and the latter one is linked to blood-cell development progresses from a hematopoietic stem cell.

4 Discussion

The study of cancer via microarray analysis is producing an ever-growing set of multiclass classification problems. As limitations (e.g., reproducibility and interpretability) of traditional machine learning techniques have been increasingly appreciated, we anticipate the need for new and effective methodologies to address these problems. In this article, we have introduced a new approach for classification of multiple cancer classes based on microarray data. The main advantage of our method lies in the simplicity and power of the

top scoring gene sets, which provide clear decision boundaries and allows for potential biological interpretations. Such simple models have begun to gain favor in recent studies, see e.g., Kaur et al. (2012) and Haibe-Kains et al. (2012).

Our classification approach makes specific hypotheses about the predictive significance of relative gene expression comparison within top scoring sets. Although such comparison may not represent the actual mechanisms of complex diseases, it does not diminish the usefulness of our method for identifying diagnostic or prognostic biomarkers associated with cancer. In fact, given the amount of data typically available for microarray analysis (especially for some types of cancer), the empirical distribution of relative comparison seems to be one of statistics that can be robustly estimated. In addition, such comparison within a group of genes can be viewed as a highly simplified model of genetic network, which is well-recognized to be involved in various diseases.

The predictive ability of our approach has been demonstrated on a variety of gene expression datasets. The seven common and publicly available microarray datasets provide a good opportunity to test our method in typical small-sample learning situations. In this case, we have shown the robust performance of our classifier across different datasets. Moreover, we have explored the ability of our method to perform classification on an extremely large dataset from the MILE study, which is quite challenging due to the large number of classes and unbalanced sample sizes. In this situation, our approach has also produced comparable predictive accuracies to a large ensemble of SVMs on an independent test set. We used only three TSS classifiers for class prediction as compared to 91 SVM classifiers used in the MILE study. This result demonstrates the potential of our method for handling large multiclass problems.

It is also possible to extend our method when sufficient data are available. Considering gene sets with the “max” gene changing over classes is one of many ways to investigate possible perturbations in genetic network through expression relative comparison. In fact, the number of complete orderings (permutations) of even a few genes is large (e.g., $4! = 24$, $5! = 120$), and it is impractical to estimate of the distribution of the orderings. Our approach provides a way to combine some of these orderings to gain statistical significance. As the sample size grows, more orderings can be considered in the modeling process, which would give a more accurate estimate of the statistical dependency structures among genes.

As recent studies on pathway-based classification have achieved good progress, combining TSS with gene pathway analysis can be a promising approach for the incorporation of biological information while reducing the data dimension. Our result indicates that subtle but reliable changes in expression among genes within pathways do exist and can be useful for classification, while such signals are often dismissed when the whole set of genes is being studied. At the same time, pathway-based TSS allows one to identify components of the complex genetic networks where genes differentially contribute to phenotypes of interest, and these components can serve as important targets for further investigation.

Acknowledgements

The authors would like to thank Donald Geman and the reviewers for their insightful comments and suggestions. This work was partially supported by NIH-NCRR Grant UL1 RR 025005.

Appendix

A Bayesian decision-theoretic interpretation

In this section, we will provide an interpretation for the scoring equation of the TSS classifier (see (3) in Methods) using the Bayesian decision theory, and derive the optimum Bayesian classifier with a general loss function and class priors. Consider an m -class classification problem, TSS aims to find m -gene set $\mathcal{S} = \{i_1, i_2, \dots, i_m\}$ such that

$$P[\arg \max \{X_r, r \in \mathcal{S}\} = i_c | y = c] > P[\arg \max \{X_r, r \in \mathcal{S}\} = i_c | y \neq c], \quad c \in \{1, 2, \dots, m\}.$$

Now suppose the class conditional probability distribution associated with gene expression comparisons in $\mathcal{S}$ is given by

	Class			
	$y = 1$	$y = 2$	...	$y = m$
$X_{i_1} > \max\{X_r, r \in \mathcal{S} \setminus i_1\}$	p_{11}	p_{12}	...	p_{1m}
$X_{i_2} > \max\{X_r, r \in \mathcal{S} \setminus i_2\}$	p_{21}	p_{22}	...	p_{2m}
.....				
$X_{i_m} > \max\{X_r, r \in \mathcal{S} \setminus i_m\}$	p_{m1}	p_{m2}	...	p_{mm}

A decision procedure δ can be constructed where the result of each comparison in the table above is considered being indicative of a sample from a distinct class. In this situation, m classes lead to $m!$ possible decision procedures for a given gene set where one of those decision procedures is illustrated below.

X	$\delta(X)$
$X_{i_1} > \max\{X_r, r \in \mathcal{S} \setminus i_1\}$	3
$X_{i_2} > \max\{X_r, r \in \mathcal{S} \setminus i_2\}$	5
.....	
$X_{i_m} > \max\{X_r, r \in \mathcal{S} \setminus i_m\}$	2

Next, a loss function can be introduced for δ by specifying the penalties for misclassification as follows

	$\delta = 1$	$\delta = 2$	...	$\delta = m$
$y = 1$	l_{11}	l_{12}	...	l_{1m}
$y = 2$	l_{21}	l_{22}	...	l_{2m}
...				
$y = m$	l_{m1}	l_{m2}	...	l_{mm}

Based on the tables above, $R(i, \delta)$, the *risk function* of δ for class $y = i$ can be written as

$$R(i, \delta) = \sum_{j=1}^m l_{ij} \cdot p(\delta=j|y=i).$$

Consequently, $r(\delta)$, the *Bayes risk* of δ is given by

$$r(\delta) = \sum_{i=1}^m \pi_i \cdot R(i, \delta)$$

where π_i is the prior probability for class $y = i$. Therefore, the Bayes risk associated with δ is given by

$$r(\delta) = \sum_{i=1}^m \pi_i \cdot \sum_{j=1}^m l_{ij} \cdot p(\delta=j|y=i), \quad (\text{A-1})$$

and the decision rule δ^* satisfying

$$\delta^* = \arg \min_{\delta} r(\delta)$$

is the optimal rule referred to as the *Bayes Rule*. As mentioned earlier, for a given gene set, there are $m!$ possible decision procedures, and since the number of possible gene sets is also finite, the optimal Bayes rule δ^* for the problem can be found by searching for the rule that minimizes $r(\delta)$ among all gene sets.

It is important to note that the Bayesian optimality of the decision rule described above only applies when the gene set used for classification has been determined. Otherwise, the development of the Bayes rule requires the joint probability distribution of all genes. In fact, no Bayesian theory is directly related to the choice of gene set for classification.

Equation (A-1) uses the general form of loss function and class prior probabilities. In practice, the choice of loss function and class priors depends on the problem. For example, the empirical estimation n_c/N can be considered, where n_c is the sample size of class c and N is the total sample size. In the context of this paper, the sample size of microarray datasets is quite limited and the sample proportions in a particular dataset may not reflect the actual distribution in the population. Therefore, we used equal class priors in our approach. In

addition, without further information about the relative importance of various misclassifications, we assumed a 0-1 loss function, e.g.,

$$l_{ij} = \begin{cases} 0, & \text{if } i=j, \\ 1, & \text{otherwise.} \end{cases}$$

In this case, $R(i, \delta)$ becomes

$$R(i, \delta) = \sum_{j=1}^m l_{ij} \cdot p(\delta=j|y=i) = \sum_{j \neq i} p(\delta=j|y=i) = 1 - p(\delta=i|y=i),$$

and $r(\delta)$ is

$$r(\delta) = \sum_{i=1}^m \pi_i R(i, \delta) = 1 - \frac{1}{m} \sum_{i=1}^m p(\delta=i|y=i).$$

Then minimizing $r(\delta)$ is equivalent to finding

$$\max_{\delta} \sum_{i=1}^m p(\delta=i|y=i).$$

The optimal rule for the equation above can be found heuristically using the equation (3) in Section 2.2, and it turns out to be the top scoring sets.

B The acceleration algorithm

The acceleration algorithm described here generalizes the pruning algorithm introduced by Tan et al. (2005) for the TSP classifier to the multiclass case. Similar to the binary TSP method, an important step for the multiclass TSS approach is to search for top scoring gene sets. Once the search process is completed, the decision rule can be immediately derived. However, given the large number of genes for microarray data, the search process is often computationally expensive. We have previously introduced two methods for searching gene sets with high scores. They are significantly more efficient than the exhaustive search, but may not be fast when combined with schemes such as cross-validation. Therefore, the algorithm here aims to accelerate the search process in the cross-validation loop.

In TSS, different methods can be employed in the search process, but only top scoring sets are kept. Therefore, gene sets that are impossible to achieve the top score can be excluded in the search process, which obviously requires a “complete” comparison among all gene sets. In a typically cross-validation loop, one such comparison is needed for each iteration. However, we will demonstrate that the acceleration algorithm can produce a small list of gene sets so that only a comparison among these gene sets is sufficient to find top scoring sets.

Let $r_g(n)$ denote the score obtained for a given gene set g when a subset of n samples is left out from N training samples in the cross-validation. The lower bound $L_g(n)$ and the upper bound $U_g(n)$ are defined as

$$\begin{aligned} L_g(n) &\leq \min \{r_g(n) : \text{any size } n \text{ subset}\} \\ U_g(n) &\geq \max \{r_g(n) : \text{any size } n \text{ subset}\}. \end{aligned}$$

Now suppose the lower and upper bounds are obtained for all possible gene sets $\{g_i, i = 1, 2, \dots\}$. Rank all lower bounds from largest to smallest and set the largest lower bound to L . Without loss of generality, assume $L = L_{g_1}(n)$. Then the following claim holds:

Claim

If $U_{g_i}(n) < L$ then the gene set g_i can not be a top scoring set on $N - n$ samples for any size n subset.

Proof

According to the definition of $U_{g_i}(n)$, $r_{g_i}(n) \leq U_{g_i}(n)$. If $r_{g_i}(n) < L$, the following inequalities satisfy for any size n subset

$$r_{g_i}(n) \leq U_{g_i}(n) < L \leq r_{g_1}(n).$$

Therefore, there is at least one gene set g_1 scored higher than g_i regardless of the choice of the size n subset. This claim follows immediately.

The reduced list Ω typically contains only a few gene sets. The identification of top scoring sets from Ω is extremely fast. The significant improvement in efficiency is hence achieved by repeatedly using Ω in each iteration of the cross-validation. The lower and upper bound for a given gene set can be obtained by calculating all possible scores when any size n subset is left out. Unless a large n and a large number of classes are considered simultaneously, this process is also efficient.

In practice, $\mathcal{G}$ in Table A.1 can be any gene set collection considered in the search process. For example, in the greedy search process, there are a number of sub-classification problems generated from the original problem. In each of these sub-problems, only top scoring sets are stored. Therefore, the acceleration algorithm can be applied in each step of the greedy search to yield a reduced list of gene sets that can possibly be identified as top scoring sets in the cross-validation. As a result, the greedy search process only needs to be applied one time on the training set.

The acceleration algorithm here can be immediately extended for the k -TSS classifier that uses top k scoring sets as the decision rule. In this situation, only step 2 in Table A.1 needs to be changed so that L is set to the k -th largest lower bound. This is because any gene set whose upper bound is less than L clearly can not be one of the top k scoring sets during the cross-validation. As a result, the search process for top k scoring sets can also be quite efficient.

C Confusion matrices for acute leukemia subtypes

Section 3.4 provides the classification results of applying the greedy search-based TSS classifier on a large cohort of acute leukemia samples from the MILE project. The classification uses a two-step decision tree (Figure 4) to decompose the original problem into three sub-classification problems: a three-class problem at the top level, and a seven- and six-class problem at the bottom level. The following confusion matrices are for these three problems respectively. True/gold standard (GS) classifications of samples are presented in rows and predictions are in columns.

Table A.1

Description of the acceleration algorithm.

Acceleration Algorithm

Input: N training samples, gene set collection $G = \{g_1, g_2, \dots\}$

Output: The reduced gene set list Ω .

1. For each gene set g_i , compute the lower bound $L_{g_i}(n)$ and the upper bound $U_{g_i}(n)$ under all possible situations that n training samples are left out.
2. Rank all $L_{g_i}(n)$ in descending order and take $L = \max\{L_{g_i}(n)\}$
3. Generate the list Ω consisting of all g_i for which $U_{g_i}(n) \geq L$

Table A.2

Confusion matrices for classification of acute leukemia subtypes.

	Predicted		
GS	B-ALL	T-ALL	AML
B-ALL	352	2	3
T-ALL	2	75	2
AML	3	16	238

	Predicted						
GS	C1	C2	C3	C5	C6	C7	C8
C1	5	0	0	0	0	0	0
C2	0	20	0	0	0	2	1
C3	0	0	51	0	0	2	9
C5	0	0	0	62	0	1	1
C6	0	0	0	0	10	0	0
C7	0	0	2	1	0	30	2
C8	2	2	28	9	3	33	76

	Predicted					
GS	C9	C10	C11	C12	C13	C14
C9	14	0	0	0	0	0
C10	0	19	0	0	0	0
C11	0	0	17	0	1	0

	Predicted					
GS	C9	C10	C11	C12	C13	C14
C12	0	0	0	11	4	0
C13	0	1	1	3	127	16
C14	1	0	0	1	5	17

References

- Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*. 2002; 30(1):41–7. [PubMed: 11731795]
- Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adeno-carcinoma. *Nature Medicine*. 2002; 8(8):816–24.
- Breiman L. Random forests. *Machine Learning*. 2001; 45(1):5–32.
- Burgess DJ. Cancer genetics: Initially complex, always heterogeneous. *Nature Reviews Cancer*. 2011; 11:153.
- Cheok MH, Yang W, Pui CH, Downing JR, Cheng C, Naeve CW, Relling MV, Evans WE. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nature Genetics*. 2003; 34(1):85–90. [PubMed: 12704389]
- Chih-Chung C, Chih-Jen L. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2011; 2(3)
- Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995; 20(3):273–297.
- Dehan E, Ben-Dor A, Liao W, Lipson D, Frimer H, Rienstein S, Simansky D, Krupsky M, Yaron P, Friedman E, Rechavi G, Perlman M, Aviram-Goldring A, Izraeli S, Bittner M, Yakhini Z, Kaminski N. Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. *Lung Cancer*. 2007; 56(2):157–84.
- Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton DS, Wolf H, Orntoft TF. Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics*. 2003; 33(1): 90–6. [PubMed: 12469123]
- Dyrskjot L, Zieger K, Real FX, Malats N, Carrato A, Hurst C, Kotwal S, Knowles M, Malmstrom PU, de la Torre M, Wester K, Allory Y, Vordos D, Caillault A, Radvanyi F, Hein AM, Jensen JL, Jensen KM, Marcussen N, Orntoft TF. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. *Clinical Cancer Research*. 2007; 13(12):3545–51. [PubMed: 17575217]
- Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technology in Cancer Research and Treatment*. 2010; 9(2):149–59. [PubMed: 20218737]
- Edelman LB, Toia G, Geman D, Zhang W, Price ND. Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics*. 2009; 10:583. [PubMed: 19961616]
- Efron, B.; Tibshirani, R. Technical report. Stanford University; 2006. On testing the significance of sets of genes. <http://www-stat.stanford.edu/~tibs/GSA/>
- Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A, Nevins JR. A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(15):6994–9. [PubMed: 20335537]
- Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mrna comparisons. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3
- Gentleman RC, Carey VJ, Bates DM. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*. 2004; 5:R80. others. [PubMed: 15461798]

- Golub T, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Collier H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999; 286(5439):531–7. [PubMed: 10521349]
- Grate LR. Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. *BMC Bioinformatics*. 2005; 6:97. [PubMed: 15826317]
- Haferlach T, Kohlmann A, Wiecek L, Basso G, Kronnie GT, Bene MC, De Vos J, Hernandez JM, Hofmann WK, Mills KI, Gilkes A, Chiaretti S, Shurtle SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Liu WM, Williams PM, Foa R. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. *Journal of Clinical Oncology*. 2010; 28(15):2529–37. [PubMed: 20406941]
- Haibe-Kains B, Desmedt C, Loi S, Culhane AC, Bontempi G, Quackenbush J, Sotiriou C. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*. 2012; 104(4):311–25. [PubMed: 22262870]
- Irizarry R, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed T. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4(2):249–64. [PubMed: 12925520]
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The kegg resource for deciphering the genome. *Nucleic Acids Research*. 2004; 32:D277–80. [PubMed: 14681412]
- Kaur P, Schlatter D, Cooke K, Chance MR. Pairwise protein expression classifier for candidate biomarker discovery for early detection of human disease prognosis. *BMC Bioinformatics*. 2012; 13:191. [PubMed: 22870920]
- Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. 2001; 7(6):673–9.
- Kim S, Kon M, DeLisi C. A pathway-based classification of human breast cancer. *Biology Direct*. 2012; 7:21. [PubMed: 22759382]
- Leban G, Bratko I, Petrovic U, Curk T, Zupan B. Vizrank: finding informative data projections in functional genomics by machine learning. *Bioinformatics*. 2005; 21(3):413–4. [PubMed: 15358614]
- Lin, X. Ph.D. thesis. The Johns Hopkins University; 2008. Rank-based methods for statistical analysis of gene expression microarray data.
- Lin X, Afsari B, Marchionni L, Cope L, Parmigiani G, Naiman DQ, Geman D. The ordering of expression among a few genes can provide simple cancer biomarkers and signal *brca1* mutations. *BMC Bioinformatics*. 2009; 10:256. [PubMed: 19695104]
- Mramor M, Leban G, Demsar J, Zupan B. Visualization-based cancer microarray data classification analysis. *Bioinformatics*. 2007; 23(16):2147–54. [PubMed: 17586552]
- Patnaik SK, Kannisto E, Knudsen S, Yendamuri S. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. *Cancer Research*. 2010; 70(1):36–45. [PubMed: 20028859]
- Quackenbush J. Microarray analysis and tumor classification. *New England Journal of Medicine*. 2006; 354(23):2463–72. [PubMed: 16760446]
- Shah MA, Khanin R, Tang L, Janjigian YY, Klimstra DS, Gerdes H, Kelsen DP. Molecular classification of gastric cancer: a new paradigm. *Clinical Cancer Research*. 2011; 17(9):2693–701. [PubMed: 21430069]
- Shen L, Tan EC. Reducing multiclass cancer classification to binary by output coding and svm. *Computational Biology and Chemistry*. 2006; 30(1):63–71. [PubMed: 16321568]
- Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multiclassification methods for microarray gene expression cancer diagnosis. *Bioinformatics*. 2005; 21(5):631–43. [PubMed: 15374862]
- Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*. 2008; 9:319. [PubMed: 18647401]

- Subramanian A, T. P. Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(43):15545–50. [PubMed: 16199517]
- Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005; 21(20):3896–904. [PubMed: 16105897]
- Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(10):6567–72. [PubMed: 12011421]
- Xu L, Geman D, W. R. L. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*. 2007; 8:275. [PubMed: 17663766]
- Zhao H, Logothetis CJ, Gorlov IP. Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. *Prostate Cancer and Prostatic Diseases*. 2010; 13(3):252–9. [PubMed: 20386565]

Figure 1.

Gene expression patterns for a top scoring pair of genes. The figure displays the expression levels of gene *SPTAN1* and *CD33* on 72 patient samples in Golub et al. (1999), which are grouped according to two types of leukemia cancer: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), with 25 and 47 samples respectively. For classifying a sample, the decision rule predicts ALL if *SPTAN1* has higher expression level than that of *CD33* in the sample, and AML otherwise.

Figure 2.

Gene expression patterns for a top scoring set of genes. The set consists of three genes: *MB-1*, *LCK* and *PLCB2*. The figure shows the expression levels of these genes on 72 patient samples, which are ordered according to three subtypes of leukemia cancer: B-cell ALL (B-ALL), T-cell ALL (T-ALL) and AML, with the size of 25, 38 and 9 respectively.

Figure 3. Schematic diagram of the greedy search algorithm. The workflow of the algorithm is illustrated for a four-class problem. Blue arrows represent the initialization step where each possible two-class sub-problems are considered. Each red arrow denotes an augmentation of the current problem by a single class. One possible sequence of augmentations is shown in the graph.

Figure 4.
Two-step decision tree for classification of acute leukemia samples.

Table 1

Observed frequencies of expression comparison within a three-gene set.

	Leukemia		
	AML	B-ALL	T-ALL
$X_{i1} > \max(X_{i2}, X_{i3})$	1	0	0
$X_{i2} > \max(X_{i1}, X_{i3})$	0	0.9737	0
$X_{i3} > \max(X_{i1}, X_{i2})$	0	0.0263	1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Observed frequencies of expression comparison associated with $\mathcal{S}$.

	Class			
	$y = 1$	$y = 2$	...	$y = m$
$X_{i_1} > \max\{X_r, r \in S \setminus i_1\}$	$p_{11}^\wedge$	$p_{12}^\wedge$	...	$p_{1m}^\wedge$
$X_{i_2} > \max\{X_r, r \in S \setminus i_2\}$	$p_{21}^\wedge$	$p_{22}^\wedge$	...	$p_{2m}^\wedge$
.....				
$X_{i_m} > \max\{X_r, r \in S \setminus i_m\}$	$p_{m1}^\wedge$	$p_{m2}^\wedge$	...	$p_{mm}^\wedge$

Table 3

Seven gene expression datasets for evaluating classification performance.

Dataset	No. of classes	No. of genes	No. of samples	Reference
Leukemia	3	7129	72	Golub et al. (1999)
MLL	3	12582	72	Armstrong et al. (2002)
Lung	3	7129	96	Beer et al. (2002)
SRBCT	4	2308	83	Tibshirani et al. (2002)
Bladder	3	7129	40	Dyrskjot et al. (2003)
ChildALL	4	12625	60	Cheok et al. (2003)
NSCLC	3	12599	33	Dehan et al. (2007)

Table 4

Top scoring gene sets identified on seven gene expression datasets. The table includes the number of top scoring sets and genes involved, the top score and the re-substitution error for each dataset.

Dataset	No. of sets	No. of genes	Score	Error
Leukemia	73	36	2.97/3.00	1 /72
MLL	7	12	2.96/3.00	1 /72
Lung	1	3	2.72/3.00	16/96
SRBCT	2	5	3.93/4.00	2/83
Bladder	3	5	3.00/3.00	0/40
ChildALL	1	4	3.09/4.00	13/60
NSCLC	2	6	2.88/3.00	1/33

Table 5

Observed frequencies of gene expression comparison in two top scoring sets from (a) NSCLC and (b) SRBCT respectively. Each class conditional frequency is equal to the proportion of times a certain gene achieves the maximum expression value in the class.

(a)

Class	Max gene		
	<i>KRT14</i>	<i>CNGB1</i>	<i>GDF10</i>
SCC	1	0	0
ADCA	0	1	0
N	0	0.12	0.88

(b)

Class	Max gene			
	<i>GYG2</i>	<i>EST</i>	<i>CDH2</i>	<i>HCLS1</i>
EWS	0.93	0	0.07	0
RMS	0	1	0	0
NB	0	0	1	0
BL	0	0	0	1

Table 6

Comparison of classification accuracies estimated using LOOCV. The highest accuracy for each dataset is highlighted in boldface.

Method	Leukemia	MLL	Lung	SRBCT	Bladder	ChildALL	NSCLC
G-TSS	95.83	94.44	70.83	90.36	100.00	48.33	90.91
MW-TSS	95.83	76.39	62.50	89.15	57.50	36.67	87.87
kNN	81.94	91.67	75.00	95.18	77.50	45.00	72.72
NB	94.44	95.83	75.00	98.80	82.50	46.67	66.67
RF	93.06	94.44	78.13	100.00	90.00	48.33	72.72
PAM	97.22	93.06	70.83	100.00	85.00	31.67	69.70
l-SVM	93.06	94.44	83.33	100.00	92.50	41.67	81.82

Table 7

Samples of acute leukemia subtypes used for classification. Three major leukemia classes consist of 14 subtypes. The class labels (C1 to C14) are the same as defined in the MILE study.

Class	Diagnosis	No. of samples	
		Training	Test
-	B-ALL	576	357
C1	Mature B-ALL with t(8;14)	13	5
C2	Pro-B-ALL with t(11q23)/ <i>MLL</i>	70	23
C3	c-ALL/Pre-B-ALL with t(9;22)	122	62
C5	ALL with t(12;21)	58	64
C6	ALL with t(1;19)	36	10
C7	ALL with hyperdiploid karyotype	40	35
C8	c-ALL/Pre-B-ALL without t(9;22)	237	158
C4	T-ALL	174	79
-	AML	542	257
C9	AML with t(8;21)	40	16
C10	AML with t(15;17)	37	20
C11	AML with inv(16)/t(16;16)	28	20
C12	AML with t(11q23)/ <i>MLL</i>	38	17
C13	AML with normal kt./other abn.	351	160
C14	AML complex aberrant karyotype	48	24

Table 8

Comparison of acute leukemia classification methods. The number of correct classifications is followed by the corresponding accuracy (in percentage) for each class.

Class	G-TSS	I-SVM
C1	5 (100.0)	4 (80.0)
C2	20 (87.0)	23 (100.0)
C3	51 (82.3)	53 (85.5)
C4	75 (94.9)	75 (94.9)
C5	62 (96.9)	59 (92.2)
C6	10 (100.0)	10 (100.0)
C7	30 (85.7)	22 (62.9)
C8	76 (48.1)	141 (89.2)
C9	14 (87.5)	16 (100.0)
C10	19 (95.0)	19 (95.0)
C11	17 (85.0)	20 (100.0)
C12	11 (64.7)	15 (88.2)
C13	127 (79.4)	148 (92.5)
C14	17 (70.8)	17 (70.8)

Table 9

Significant KEGG pathways identified on Leukemia. The (default) false discovery rate (FDR) is 0.2 and the (default) size (i.e., the number of genes) of a pathway is limited to between 15 and 500.

Gene set	Size	<i>p</i> -value	FDR
Primary immunodeficiency	32	< 0.001	< 0.001
T cell receptor signaling pathway	96	< 0.001	< 0.001
B cell receptor signaling pathway	68	< 0.001	< 0.001
Hematopoietic cell lineage	104	< 0.001	< 0.001
Rheumatoid arthritis	88	0.005	0.178

Table 10

Comparison of classification methods on Leukemia.

Method	Test error	No. of genes
Golub et al. (1999)	4/34	50
kNN	4/34	7129
NB	4/34	7129
RF	4/34	1273
PAM	1/34	47
l-SVM	5/34	7129
Pathway TSS	1/34	10