BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 20 2005, pages 3905–3911
doi:10.1093/bioinformatics/bti647

Gene expression

Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data

Lei Xu[1,2,*], Aik Choon Tan[1,2], Daniel Q. Naiman[1,2,3], Donald Geman[1,2,3,4] and Raimond L. Winslow[1,2]

[1]The Whitaker Biomedical Engineering Institute, [2]Center for Cardiovascular Bioinformatics and Modeling, [3]Department of Applied Mathematics and Statistics and [4]Center for Imaging Sciences, The Johns Hopkins University, Baltimore MD 21218, USA

Received on May 19, 2005; revised on June 21, 2005; accepted on August 25, 2005
Advance Access publication August 30, 2005

ABSTRACT

Motivation: DNA microarray data analysis has been used previously to identify marker genes which discriminate cancer from normal samples. However, due to the limited sample size of each study, there are few common markers among different studies of the same cancer. With the rapid accumulation of microarray data, it is of great interest to integrate inter-study microarray data to increase sample size, which could lead to the discovery of more reliable markers.

Results: We present a novel, simple method of integrating different microarray datasets to identify marker genes and apply the method to prostate cancer datasets. In this study, by applying a new statistical method, referred to as the top-scoring pair (TSP) classifier, we have identified a pair of robust marker genes (HPN and STAT6) by integrating microarray datasets from three different prostate cancer studies. Cross-platform validation shows that the TSP classifier built from the marker gene pair, which simply compares relative expression values, achieves high accuracy, sensitivity and specificity on independent datasets generated using various array platforms. Our findings suggest a new model for the discovery of marker genes from accumulated microarray data and demonstrate how the great wealth of microarray data can be exploited to increase the power of statistical analysis.

Contact: leixu@jhu.edu

INTRODUCTION

The advent of DNA microarrays provides a powerful tool in cancer research and several studies have used this technology to identify genes that could be used as candidate markers to discriminate cancer conditions from normal conditions (Alizadeh *et al*., 2000; Bittner *et al*., 2000; Dhanasekaran *et al*., 2001; Golub *et al*., 1999; Takahashi *et al*., 2001). It is quite interesting, but perhaps not surprising, that the marker genes from different investigations involving patients with the same cancer (e.g. prostate cancer) are study-specific, that is, there are few common marker genes among different studies (Nelson, 2004). These diverse results make it difficult to identify the most important marker genes, and the corresponding decision rules, for detecting cancer. Differences in results potentially result from the relatively small sample sizes used in each study. Indeed, in a recent study performed by Mukherjee *et al*. (2003) concerning the influence of the size of the microarray dataset, it was concluded that increased sample size improves both the accuracy and significance of classification results.

The rapid accumulation of microarray gene expression data suggests that combining microarray data obtained from different studies may be a useful way to increase sample size. This could in turn lead to the discovery of more robust markers of cancer. However, several issues arise when attempting to integrate microarray data generated by disparate groups using different array technologies. Results obtained using microarray data generated from different technologies, such as spotted cDNA and Affymetrix arrays, may show poor correlation and cannot be compared directly (Kuo *et al*., 2002). Even when using the same microarray technology, different generations of microarrays have different probe sets and duplicate spots, making direct integration difficult. In addition, variation among datasets, resulting from alternative experimental protocols and parameters as well as sampling of different patient populations, poses further challenges to the integration of microarray data from independent studies.

Recently, several methods have been proposed to combine inter-study microarray data at different levels in cancer research (Choi *et al*., 2003; Jiang *et al*., 2004; Rhodes *et al*., 2002; Shen *et al*., 2004). Instead of integrating microarray gene expression values, some methods, referred to as meta-analysis, combine results (e.g. *t*-statistic) of individual studies to increase the power of identifying genes differentially expressed between normal and cancer samples (Choi *et al*., 2003; Rhodes *et al*., 2002). One limitation of these methods is that the small sample sizes typical of individual studies, coupled with variations due to differences in study protocols, will inevitably degrade the results of meta-analysis. In addition, a recent study (Mah *et al*., 2004) has demonstrated that there is only moderate overlap in gene detection on different array platforms. By applying specific data transformation and normalization procedures, other methods translate gene expression values of independent studies into a common scale and then combine inter-study data onto that scale to develop methods for building prognostic signatures or identifying marker genes (Jiang *et al*., 2004; Shen *et al*., 2004). However, there is still no consensus on how best to perform data normalization.

Here we propose a novel, simple method to integrate inter-study gene expression values in order to identify marker genes for cancer

[*]To whom correspondence should be addressed.

Table 1. Training and testing datasets

Dataset	Microarray platform	Number of probe sets	Number of normal samples	Number of cancer samples
Training set				
Singh (Singh *et al*., 2002)	Affymetrix (HG_U95Av2)	12 600	50	52
Stuart (Stuart *et al*., 2004)	Affymetrix (HG_U95Av2)	12 625	50	38
Welsh (Welsh *et al*., 2001)	Affymetrix (HG_U95Av2)	12 626	9	24
Testing set				
LaTulippe (LaTulippe *et al*., 2002)	Affymetrix (HG_U95Av2)	12 626	3	23
Lapointe[a] (Lapointe *et al*., 2004)	Spotted cDNA	44 160/43 008	41	62

[a] 22 samples (9 normal/13 cancer) have 44 160 probes and 81 samples (32 normal/49 cancer) have 43 008 probes.

classification and apply it to prostate cancer microarray datasets. In this method, there is no need to perform data normalization and transformation before data integration. We choose to study prostate cancer because of the public availability of a substantial pool of prostate cancer microarray data. The method is generally applicable to many other types of microarray data for marker gene identification. In this work, we apply a new molecular classification method, referred to as the top-scoring pair (TSP) classifier, to microarray datasets integrated from three independent studies to identify a pair of marker genes, called the marker TSP (Geman *et al*., 2004). This is feasible since this classification method is invariant to standard procedures for data normalization and transformation. Cross-platform validation shows that the classifier built from the marker TSP (HPN and STAT6), which declares prostate cancer if the expression value of HPN is greater than that of STAT6, achieves high accuracy, sensitivity and specificity on two independent datasets generated from different microarray platforms. In addition, the performance of the marker TSP is compared with that of the TSPs from the three individual studies on the same cross-platform testing datasets. Performance of the marker TSP classifier is better than all of the TSP classifiers trained from individual data. Furthermore, by reviewing the prostate cancer literature, we note that HPN has been identified as a marker gene of prostate cancer in recent studies (Dhanasekaran *et al*., 2001; Klezovitch *et al*., 2004; Luo *et al*., 2001; Nelson, 2004). STAT6 is also found to be closely related to prostate cancer (Ni *et al*., 2002). These findings suggest that we have identified robust prostate cancer marker genes by directly integrating inter-study microarray data. Upon further validation on additional independent datasets, the marker genes could be used to develop a genomic-based, more accurate diagnostic test for prostate cancer.

METHODS

Gene expression data

Five prostate microarray datasets are included in this study. Each dataset has been downloaded from publicly available gene expression repositories or supporting web sites (Lapointe *et al*., 2004; LaTulippe *et al*., 2002; Magee *et al*., 2001; Rhodes *et al*., 2004; Singh *et al*., 2002; Stuart *et al*., 2004; Welsh *et al*., 2001). The three datasets used as training samples are generated from the same Affymetrix HG_U95Av2 platform by different labs and the remaining two datasets used as testing samples are from cross-platform independent studies (Table 1). In this study, we focus on identifying marker genes which can distinguish primary prostate cancer from normal samples. Therefore, metastatic prostate cancer samples are not included in the study. The summaries of the training and testing datasets are provided in Table 1. Here, the names of the first authors of individual studies are used as the names of the datasets. Details about each dataset have been described in the corresponding literature.

TSP classifier

Recently, our group has developed a statistical molecular classification method, referred to as the TSP classifier, based on pairwise comparisons of gene expression values within each microarray (Geman *et al*., 2004). This classifier discriminates between two classes by finding pairs of genes that achieve the largest 'score' defined by a simple measure of discrimination (see below). This approach only uses the ranks (orderings) of gene expression values within each profile. Whereas information is lost using a rank-based method, the results obtained by the TSP classifier on several different microarray datasets show that rank information within each microarray is sufficient to perform molecular classification reliably. In fact, despite its simplicity with respect to other methods, the TSP classifier achieves classification rates comparable to or exceeding the best results reported in the literature (Geman *et al*., 2004). An important feature of rank-based methods is that they are invariant to monotonic transformations of the expression data, such as most data normalization methods. This property makes these methods useful for combining inter-study microarray data without the need to perform data normalization and transformation.

Detailed information about the TSP classifier can be found in (Geman *et al*., 2004). Here we give a brief and intuitive description of this rank-based approach. Assume the training microarray dataset is a $G \times N$ matrix $\boldsymbol{X} = [X_{gn}]$, $g = 1, 2, \ldots, G$ and $n = 1, 2, \ldots, N$, where G is the number of genes in each profile and N is the number of samples (i.e. profiles). Each column represents a gene expression profile of G genes and each row represents observations of a particular gene over N samples. Each sample has a class label of either 1 or 2. For simplicity, we assume that samples 1 to N_1 ($N_1 < N$) are labeled as class 1 (e.g. normal) and samples ($N_1 + 1$) to N are labeled as class 2 (e.g. cancer). For each pair of genes (i, j), $i, j = 1, 2, \ldots, G$, $i \neq j$, we calculate a score based on the training set $\boldsymbol{X}$,

$$\Delta_{ij} = |p_{ij}(1) - p_{ij}(2)|. \tag{1}$$

Here $p_{ij}(1)$ and $p_{ij}(2)$ are defined as

$$p_{ij}(1) = \frac{1}{N_1} \sum_{n=1}^{N_1} I_{(X_{in} < X_{jn})}, \quad p_{ij}(2) = \frac{1}{N - N_1} \sum_{n=N_1+1}^{N} I_{(X_{in} < X_{jn})}. \tag{2}$$

And $I_{(X_{in} < X_{jn})}$ is the indicator function defined as

$$I_{(X_{in} < X_{jn})} = \begin{cases} 1, & \text{if } X_{in} < X_{jn} \\ 0, & \text{if } X_{in} \geq X_{jn}, \end{cases} \quad n = 1, 2, \ldots, N. \tag{3}$$

Table 2. TSPs from training datasets with increased sample sizes

Training dataset	Sample size	Probe set ID of TSP (HG_U95Av2)	Gene symbol of TSP	Score of TSP	Classification accuracy (%)[b]
Welsh	33	39 608_at, 32 526_at	SIM2, JAM3	1.00	97.0
Stuart	88	41732_at, 456_at	CTNNB1, SMARCD3	0.74	69.3
Singh	102	40282_s_at, 2035_s_at	DF, ENO1	0.90	95.1
Welsh_Stuart[a]	121	31971_at, 34213_at	TP73L, KIBRA	0.79	77.7
Welsh_Singh	135	37639_at, 32198_at	HPN, COMMD4	0.88	83.7
Stuart_Singh	190	37639_at, 41222_at	HPN, STAT6	0.75	86.8
Welsh_Stuart_Singh	223	37639_at, 41222_at	HPN, STAT6	0.78	88.8

[a] Welsh_Stuart is the integrated dataset of Welsh and Stuart datasets. Other integrated datasets use similar symbols.

[b] Classification rates were estimated by leave-one-out cross-validation on the individual training sets.

Table 3. Classification accuracy of the marker TSP classifier on cross-platform testing sets

Testing dataset	Microarray platform	Number of normal sample	Number of cancer sample	Accuracy (%)	Sensitivity (%)	Specificity (%)
LaTulippe	Affymetrix (HG_U95Av2)	3	23	96.2	95.7	100.0
Lapointe[a,b,c]	Spotted cDNA	41	61	93.1	90.2	97.6
Overall	Cross-platform	44	84	93.8	91.7	97.7

[a] 'Log_2 of R/G normalized ratio (Mean)' values are used as gene expression values.

[b] The corresponding clone IDs for the gene pair (HPN, STAT6) is (IMAGE:208413, IMAGE:85541).

[c] One of the cancer samples has missing value for HPN and is removed from the testing set.

In other words, $p_{ij}(1)$ (respectively, $p_{ij}(2)$) is the estimated probability of observing X_i less than X_j in class 1 (respectively, class 2). Consequently, it is sufficient to know the ranks of gene expression values within each profile to obtain all the scores Δ_{ij}, $i, j = 1, 2, \ldots, G$, $i \neq j$. The next step is to select all pairs achieving the largest score. If there is only one pair achieving the top score, we take this pair to be the final TSP of the training set. Otherwise (i.e. if multiple pairs achieve the top score) we use a more sensitive score, called the rank-score, in order to find a unique pair. The rank-score takes into account the rank differences for each TSP and each sample, i.e. the extent to which a gene pair inverts from one class to the other. For each TSP (i, j), the rank-score, denoted by δ_{ij}, is defined as

$$\delta_{ij} = \left| \frac{1}{N_1}\sum_{n=1}^{N_1}\left(R_{in}-R_{jn}\right) - \frac{1}{N-N_1}\sum_{n=N_1+1}^{N}\left(R_{in}-R_{jn}\right)\right|. \tag{4}$$

Here R_{in} is the rank of the expression value of gene i within the n-th profile. Finally, we select the TSP with the highest rank-score.

The classification decision will be made by comparing the expression values of the two genes in the TSP(i,j) on a test sample. Suppose $p_{ij}(1) \geq p_{ij}(2)$. In this case, given a test sample with expression values $X_1, X_2, \ldots, X_G$, if we observe that $X_i < X_j$, then the TSP votes for class 1; otherwise, i.e. if $X_i \geq X_j$, the TSP votes for class 2. On the other hand, suppose $p_{ij}(1) < p_{ij}(2)$. Then, if we observe that $X_i < X_j$, the TSP votes for class 2; otherwise, i.e. if $X_i \geq X_j$, it votes for class 1.

It is noteworthy that for classification based on a single gene pair (i, j), the sum of sensitivity and specificity can be expressed as $1 + \Delta_{ij}$, which provides a natural justification for score maximization. Another method for scoring gene pairs, based on estimating means and covariances under a bivariate normality assumption, has been presented (Lai *et al*., 2004). Note, however, that in this previous work, the objective was not classification, but rather to find significantly 'co-expressed' pairs of genes.

Error estimation

In estimating the error rate of a classifier based on cross-validation, gene pair selection and the corresponding classification are performed within the cross-validation loop. With n samples and leave-one-out cross-validation, this means choosing n separate top-scoring pairs, one for each sample left out during training, then classifying the left-out sample. In particular, both the actual top score, as well as the gene pair which achieves it, may vary with the left-out sample. The estimated classification accuracy is then $1 - e/n$ where e is the number of errors observed in the cross-validation. This is the way the classification accuracies reported in Table 2, as well as the cross-validation results given in Table 5, were calculated. However, in order to associate a single TSP with each training set (as in Table 2), and a corresponding decision rule for evaluation on an independent test set (as in Tables 3 and 4), the 'final TSP' is computed using the entire training set. As a result, all error estimates are unbiased.

Table 4. Comparisons of the marker TSP with the TSPs from individual datasets

Testing dataset	TSP	Accuracy (%)	Sensitivity (%)	Specificity (%)
LaTulippe (HG_U95Av2)	Welsh	69.2	69.6	66.7
	Stuart	84.5	82.6	100.0
	Singh	88.5	87.0	100.0
	Welsh_Stuart_Singh	96.2	95.7	100.0
Lapointe (cDNA)	Welsh	70.9	95.2	34.1
	Stuart	43.6	6.7	97.6
	Singh	43.7	6.4	100.0
	Welsh_Stuart_Singh	93.1	90.2	97.6

Data integration

By applying the TSP method, no data transformation and normalization are required before integration. Among the three individual training datasets used in this study, there are 12 600 common probe sets. We directly merge individual datasets, using the 12 600 common probe sets, to form integrated datasets of increasing sample size.

Stability analysis

We have designed an experiment to analyze the stability of a TSP in response to slight perturbations of the training set resulting from changing its size. This is accomplished by randomly removing a small percentage (*K*%) of samples from the original training set and generating a TSP from the reduced training set. After repeating the experiment a large number of times with different values of *K* (e.g. 1, 2, . . .), we calculate the appearance frequency of the TSP among all TSPs generated at each sample size, for instance, 99% appearance frequency at sample size 220. If this frequency remains very high (e.g. >95%) when the sample sizes are slightly (e.g. 5%) different from the original training set size, we conclude that the TSP is stable for the original training set.

RESULTS

Identification of marker genes by directly integrating inter-study data

To investigate whether robust marker genes that distinguish primary prostate cancer from normal samples can be identified, three microarray datasets from different studies have been collected and the TSP method has been applied to analyze the individual and integrated datasets. To avoid loss of potential marker genes, we first analyze inter-study microarray data obtained using the same platform (the HG_U95Av2 array; see Table 1). Starting from individual datasets, we gradually increase the sample size by sequentially merging two and then three data sets (see Data integration in Methods). Applying the TSP method to the training sets, individual and integrated, generates a TSP for each of the training sets. The TSPs are listed in Table 2 along with scores and classification accuracy. The score refers to the absolute value of the difference between two probabilities estimated from the training data—the probability that the expression of the first gene in the pair exceeds the expression of the second gene in the pair for the cancer class and the same probability for the normal class (see ‘TSP classifier’ in ‘Methods’). Classification accuracy is estimated by leave-one-out cross-validation on each training set. In Table 2, underscores are used to join names of individual datasets to denote an integrated dataset. For example, ‘Welsh_Stuart’ is the name of the integrated data resulting from the merging of the Welsh and Stuart datasets.

Results show that when the sample size is small, different datasets generate distinct TSPs. As the sample size reaches a certain level (between 135 and 190) and continues to increase, the pair (HPN, STAT6) is consistently selected as the TSP.

Stability analysis of the marker TSP

We subsequently performed an analysis of the stability of the marker TSP (HPN, STAT6) (see Stability analysis in Methods), where ‘stability’ refers to the sensitivity of the selection procedure to perturbations of the training set. To do this, small numbers of samples are randomly removed from the integrated dataset of size 223. At each sample size, we repeat the experiment 100 times and calculate the appearance frequency of the marker TSP. The results of the analysis are shown in Figure 1. When 1–3% of the samples are removed, the appearance frequency of the marker TSP is 100%. The marker TSP appears with very high frequency when <10% of the samples are randomly removed from the original training set. From this analysis, we have shown that the marker TSP is stable for the integrated training set. We carry out the same analysis on the TSP selected from the Stuart dataset (size 88 samples). Results are shown in Figure 1. When one sample (∼1%) is removed from the training set, the appearance frequency of the Stuart TSP declines by ∼30%. With two or more samples removed from the training set, the appearance frequency declines further. Therefore, we can conclude that the Stuart TSP is not stable for the Stuart training set.

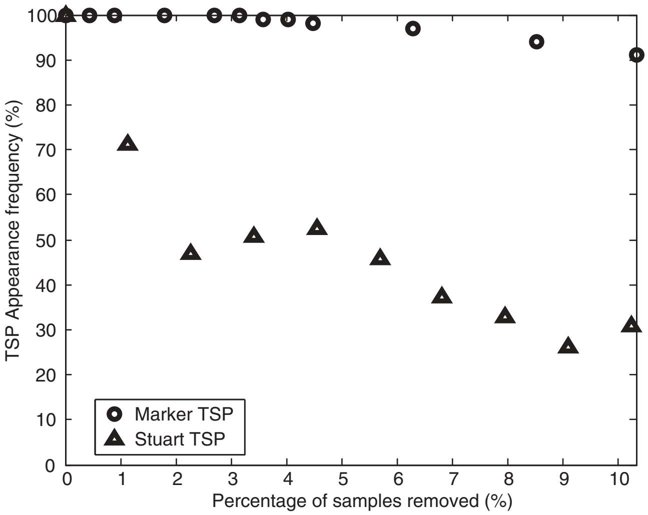

Fig. 1. Results of the stability analysis of the marker TSP and the Stuart TSP.

Validation of the marker genes and decision rule using cross-platform independent datasets

In order to further validate the reliability and robustness of the marker TSP, the marker TSP classifier is tested on independent cross-platform microarray data (Table 3). The decision rule for the maker TSP classifier is that if the expression value of HPN is greater than that of STAT6, a test sample is classified as prostate cancer; otherwise it is classified as normal. The performance of the classifier is measured by examining how well the classifier predicts the normal and cancer samples in the testing sets. Accuracy is defined as the ratio of the number of correctly predicted samples to the total number of samples. Sensitivity (respectively, specificity) is the ratio of the number of correctly predicted cancer (respectively, normal) samples to the total number of cancer (respectively, normal) samples. The marker TSP classifier consistently achieves high accuracy, sensitivity and specificity on the testing sets across different platforms (Table 3). The overall accuracy, sensitivity and specificity of the marker TSP are 93.8%, 91.7% and 97.7%, respectively.

As a comparison, the TSPs trained from individual microarray datasets are tested on the same testing sets. For the remainder of this paper, we use the name of a training dataset to represent the TSP generated from that dataset. For example, ‘Welsh’ represents the TSP from the Welsh dataset and ‘Welsh_Stuart_Singh’ is the marker TSP generated from the integrated dataset Welsh_Stuart_Singh. In each case, the decision rule is based on comparing the expression values in the marker pair for that training set (see TSP classifier in Methods). The results of the comparison are summarized in Table 4. The marker TSP outperforms all of the TSPs obtained from

Table 5. Comparison between cross-validation and independent testing results

TSP	Cross-validation results (%)			Independent testing results (%)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Welsh	97.0	95.8	100.0	70.5	88.2	36.4
Stuart	69.3	81.6	60.0	52.0	27.7	97.7
Singh	95.1	96.2	94.0	52.7	28.2	100.0
Welsh_Stuart_Singh	88.8	90.4	87.2	93.8	91.7	97.7

individual datasets on all testing sets. The results suggest that due to the limited sample size and/or artifacts of the individual datasets, the TSPs from individual datasets provide less reliable predictors of prostate cancer than the marker TSP obtained from the integrated data. Although the marker TSP is generated from integrated data obtained using a single microarray platform (Affymetrix HG_U95Av2), it can also be used for accurate classification of a novel dataset obtained using spotted cDNA microarrays.

Consistency of cross-validation and independent test results

In Table 2, we note that the classification accuracy of Welsh and Singh (again, estimated by cross validation) is higher than that of the marker TSP. However, when tested on independent testing sets, the performance of the marker TSP is considerably better than those of the TSPs from individual studies (Table 4). Table 5 summarizes the results of cross-validation and independent testing of the marker TSP and the TSPs from the three individual studies. The independent testing results reported here are the overall results on the two cross-platform independent testing sets with a total of 128 samples. For the three TSPs from individual studies, cross-validation results are inconsistent with independent testing results. This implies that the cross-validated estimates of error of individual studies have high variation and that the corresponding TSPs are somewhat study-specific and not as reliable as the marker TSP in classifying prostate samples generated from other independent studies. On the other hand, the marker TSP generates consistent results between cross-validation and testing with independent data. By integrating inter-study microarray data, the study-specific effect is reduced and more stable features of the cancer are captured by the marker TSP classifier.

DISCUSSION

The increasing availability of gene expression microarray data has been calling for methods to effectively integrate multiple, independently generated datasets targeting the same biological question. This paper presents a novel, simple method of integrating different microarray datasets to identify marker genes and illustrates the method using prostate cancer datasets. By applying a new statistical method (the TSP classifier), we have successfully identified a pair of robust prostate cancer marker genes (HPN and STAT6) from direct integration of inter-study microarray data. The TSP classifier built on the marker gene pair, which simply compares relative expression values, achieves high accuracy (93.8%), sensitivity (91.7%) and specificity (97.7%) on independent cross-platform microarray datasets.

Integration of microarray data across platforms can be achieved by using the subset of gene probes that are common to all platforms. However, the large number of genes which are not in the common set may include potential marker genes. Therefore, in this study, we use inter-study data from the same platform (Affymetrix HG_U95Av2) to identify prostate marker genes. The reason that the Affymetrix HG_U95Av2 is chosen is that there is a large amount of prostate microarray data generated on this platform among published microarray datasets. This makes it possible to increase the sample size of the integrated data to a level necessary to identify robust marker genes.

A unique pair of genes, the marker TSP (HPN and STAT6), is consistently selected as the TSP with the increase of sample sizes. However, when the TSP is computed for individual datasets, each of the datasets generates a different TSP and none of the TSPs are the same as the marker TSP. This observation implies that the dependence of the TSP on the individual dataset can be significantly diminished, and the information provided about prostate cancer can be significantly increased, by data integration. An advantage of inter-study microarray data integration is then that it increases the statistical power to capture consistent features which might be masked by the small sample size and experimental artifacts in an individual dataset. In this sense, the marker TSP is more reliable and more robust to variations in individual datasets.

In a separate paper, we present a generalization of the original TSP algorithm, referred to as k-TSP, which uses k disjoint TSPs of genes for classifying gene expression data; we also extend the k-TSP classifier from binary classification to a multi-class setting. The parameter k is determined by an inner loop of cross-validation and prediction is made by unweighted majority voting of the k TSPs. In principle, the k-TSP method can provide more reliable and reproducible results; however, the performance of the k-TSP classifier is no better than that of the TSP classifier on the prostate datasets (results reported elsewhere).

To provide a comparison with a common approach to classifying gene expression profiles, we applied the software for a popular variation of diagonal linear discriminant analysis called prediction analysis of microarrays (PAM) (Tibshirani *et al.*, 2002) which automatically selects an optimal number of genes ranked by a modification of the t-statistic. (The normalization of the standard deviation involved in 'shrunken centroids' is first performed on the three datasets separately.) We ran PAM on the integrated Welsh_Stuart_Singh dataset. This resulted in a classifier based on 135 genes (with HPN at the top of the list) and a classification accuracy of 86.1% as estimated by cross-validation (similar to the 88.8% accuracy of the TSP classifier; see Table 2). Since one of the independent test sets, LaTulippe, is generated from the same microarray platform, i.e. HG_U95Av2, we could also test the classifier learned from PAM. The accuracy of PAM and TSP are the same as on LaTulippe. However, it was not possible to evaluate PAM on the Lapointe test set, which is generated from a spotted cDNA microarray, because some of the 135 genes are not present in that dataset. Moreover, even were all these genes present in the cDNA dataset, the issue would remain of how to normalize the cDNA data to make them comparable to the HG_U95Av2 data. This comparison demonstrates the advantage of our method in the integration of inter-study microarray data.

We did not include all publicly available prostate datasets in the testing sets. For some older platforms, such as the Affymetrix Hu35kA array, there is no probe set corresponding to either of the marker genes HPN or STAT6. For cDNA microarray datasets, those which only provide the ratio of the gene expression values of normal and cancer samples cannot be used as testing sets since our analysis requires the gene expression values of both normal and cancer samples. In addition, in some cDNA microarray datasets, one of the marker genes is missing. Even with these limitations, we still obtain a reasonable number of independent testing sets (total sample size 128) on differing platforms.

An interesting and important finding of the study is that although the marker TSP is generated from integrating data obtained using a single type of microarray (the Affymetrix HG_U95Av2 short-oligo microarray), it can be used to classify microarray data obtained using a different technology (cDNA microarrays) with high accuracy. This suggests that as long as the sample size reaches a certain ‘statistically significant’ level, valid conclusions can be drawn from single-platform inter-study microarray data. Upon further confirmation from other studies, this finding might provide an alternative approach for microarray data analysis. Because cross-platform data integration is much more complex than single-platform data integration, this finding, as reported in our study, will greatly facilitate microarray data integration.

One of the TSP marker genes, HPN, has been identified as a marker gene of prostate cancer in recent studies (Dhanasekaran *et al*., 2001; Klezovitch *et al*., 2004; Luo *et al*., 2001; Nelson, 2004). HPN encodes hepsin, a cell surface transmembrane serine protease which plays an essential role in cell growth and maintenance of cell morphology. Using both cDNA and oligonucleotide microarray technologies, hepsin was shown to be significantly over-expressed in prostate cancer samples versus normal samples, and it has been identified as a potential biomarker for screening prostate cancer (Dhanasekaran *et al*., 2001; Luo *et al*., 2001; Magee *et al*., 2001; Stamey *et al*., 2001). mRNA over-expression has also been validated using RT–PCR (Luo *et al*., 2001) and protein over-expression has been verified using tissue microarrays (Dhanasekaran *et al*., 2001). Magee *et al*. (2001) also confirmed the over-expression of hepsin in prostate tumor by using the *in situ* hybridization technique on an independent panel of prostate specimens. Furthermore, the expression of hepsin has been shown to have positive correlation with prostate cancer staging (Stamey *et al*., 2001), and to promote prostate cancer progression and metastasis (Klezovitch *et al*., 2004). Thus, hepsin may be used as a diagnostic as well as prognostic marker for prostate cancer.

STAT6 encodes the signal transducer and activation of transcription 6 (Stat6), a member of the STAT transcription factors located in the cytoplasm that is involved in the Jak-Stat signaling pathway. The Jak-Stat pathway is an important signaling pathway in cellular development/survival (Calo *et al*., 2003; O’Shea *et al*., 2002). It is activated by a small number of cytokines (e.g. interleukin-4) and plays a distinct role in the development of T-cells (e.g. T-helper cell type 2) and in IFNγ signaling. The expression of STAT6 has been shown to be down-regulated in gastric cancer (Sakakura *et al*., 2002). From our study, we observe that STAT6 is slightly down-regulated in prostate cancer compared to normal samples. This down-regulation of STAT6 is necessary for the cancer cell to escape from the tumor immunosurveillance mechanism, where the tumor protects itself from being killed by the natural killer T-cells (Dunn *et al*., 2002). It has been shown that T-helper cell type 2 cytokines down-regulate anti-tumor immunity (Terabe *et al*., 2000). At the protein expression level, Ni *et al*. (2002) showed that Stat6 was selectively activated in prostate cancer using western blot analysis.

One of the goals of cancer marker gene identification is to translate inter-study microarray data analysis into clinically useful cancer markers. Prostate specific antigen (PSA), as a prostate tumor marker currently used in clinical practice has, as its major limitation, low specificity. When normal serum PSA levels are defined as 4.0 ng/mL or less, PSA testing has a sensitivity of ~67.5–80% and a specificity of ~60–70% (Brawer, 1999; Catalona *et al*., 1997). Our study has discovered robust marker genes that are sufficient to distinguish prostate cancer from normal in both training and testing sets. The high sensitivity (91.7%) and specificity (97.7%) of the marker gene pair achieved on a large number (128) of independent testing samples are encouraging, suggesting its potential clinical applicability. A possible application would be to make a simple diagnostic chip using the marker genes. A testing sample will be predicted as cancer or normal simply by comparing the expression values of the marker genes. Clearly, validation on a larger set of independent data will be required before the idea can be translated into clinical practice. Nevertheless, this study provides evidence for promising potential prostate cancer markers to improve the diagnostic accuracy of prostate cancer.

In conclusion, this work has not only established a new model for the discovery of marker genes from accumulated microarray data, but also demonstrated how the great wealth of microarray data can be exploited to increase the power of statistical analyses.

ACKNOWLEDGEMENTS

This work was supported by NHLBI RO1-HL72488, The Falk Medical Trust and The Whitaker Foundation.

Conflict of interest statement: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Bittner,M. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Brawer,M.K. (1999) Prostate-specific antigen: current status. *CA A Cancer J. Clinicians*, **49**, 264–281.

Calo,V. *et al*. (2003) STAT proteins: from normal control of cellular events to tumorigenesis. *J. Cell. Physiol*., **197**, 157–168.

Catalona,W.J. *et al*. (1997) Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: enhancement of specificity with free PSA measurement. *J. Am. Med. Assoc*., **277**, 1452–1455.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, 84i–90.

Dhanasekaran,S.M. *et al*. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.

Dunn,G.P. *et al*. (2002) Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol*., **3**, 991–998.

Geman,D. *et al*. (2004) Classifying gene expression profiles from pairwise mRNA comparison. *Stat. Appl. Genet. Mol. Biol*., **3**, 19.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Jiang,H. *et al*. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, **5**, 81.

Klezovitch,O. *et al*. (2004) Hepsin promotes prostate cancer progression and metastasis. *Cancer Cell*, **6**, 185.

Kuo,W.P. *et al*. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.

Lai,Y. *et al*. (2004) A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, **20**, 3146–3155.

Lapointe,J. *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

LaTulippe,E. *et al*. (2002) Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res*., **62**, 4499–4506.

Luo,J. *et al*. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res*., **61**, 4683–4688.

Magee,J.A. *et al*. (2001) Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res*., **61**, 5692–5696.

Mah,N. *et al*. (2004) A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol. Genomics*, **16**, 361–370.

Mukherjee,S. *et al*. (2003) Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol*., **10**, 119–142.

Nelson,P.S. (2004) Predicting prostate cancer behavior using transcript profiles. *J. Urol*., **172**, S28–S33.

Ni,Z. *et al*. (2002) Selective activation of members of the signal transducers and activators of transcription family in prostate carcinoma. *J. Urol*., **167**, 1859–1862.

O'Shea,J.J. *et al*. (2002) Cytokine signaling in 2002: new surprises in the Jak/Stat pathway. *Cell*, **109**, S121–S131.

Rhodes,D.R. *et al*. (2002) Meta-Analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*., **62**, 4427–4433.

Rhodes,D.R. *et al*. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.

Sakakura,C. *et al*. (2002) Differential gene expression profiles of gastric cancer cells established from primary tumour and malignant ascites. *Br. J. Cancer*, **87**, 1153–1161.

Shen,R. *et al*. (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, **5**, 94.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203.

Stamey,T.A. *et al*. (2001) Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. *J. Urol*., **166**, 2171–2177.

Stuart,R.O. *et al*. (2004) *In silico* dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 615–620.

Takahashi,M. *et al*. (2001) Gene expression profiling of clear cell renal cell carcinoma: Gene identification and prognostic classification. *Proc. Natl Acad. Sci. USA*, **98**, 9754–9759.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Terabe,M. *et al*. (2000) NKT cell-mediated repression of tumor immunosurveillance by IL-13 and the IL-4R-STAT6 pathway. *Nat. Immunol*., **1**, 515–520.

Welsh,J.B. *et al*. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res*., **61**, 5974–5978.