BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 13 2007, pages 1599–1606
doi:10.1093/bioinformatics/btm149

Gene expression

Meta-analysis of gene expression data: a predictor-based approach

Irit Fishel[1], Alon Kaufman[2] and Eytan Ruppin[1,3,*]

[1]School of Medicine, Tel-Aviv University, Tel-Aviv 69978, [2]Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem 91904 and [3]School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel

Received on November 2, 2006; revised on March 18, 2007; accepted on April 14, 2007

Advance Access publication April 26, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: With the increasing availability of cancer microarray data sets there is a growing need for integrative computational methods that evaluate multiple independent microarray data sets investigating a common theme or disorder. Meta-analysis techniques are designed to overcome the low sample size typical to microarray experiments and yield more valid and informative results than each experiment separately.

Results: We propose a new meta-analysis technique that aims at finding a set of classifying genes, whose expression level may be used to answering the classification question in hand. Specifically, we apply our method to two independent lung cancer microarray data sets and identify a joint core subset of genes which putatively play an important role in tumor genesis of the lung. The robustness of the identified joint core set is demonstrated on a third unseen lung cancer data set, where it leads to successful classification using very few top-ranked genes. Identifying such a set of genes is of significant importance when searching for biologically meaningful biomarkers.

Contact: ruppin@post.tau.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray technology has provided researchers with the ability to measure the expression levels of thousands of genes simultaneously. The development of high-throughput screening techniques has been used with great success for molecular profiling in diverse biological systems, including cancer research (Dopazo *et al*., 2001). Supervised machine learning approaches for the analysis of gene expression profiling have proven to be a powerful tool in the prediction of cancer diagnosis (Golub *et al*., 1999; Nguyen and Rocke, 2002), prognosis (van't Veer *et al*., 2002) and treatment outcome (Shipp *et al*., 2002). Since most of the genes are not informative for the prediction task, feature selection methods, also known as gene selection, are applied prior to prediction. Such gene selection techniques aim to identify a small subset of genes that can best serve to correctly predict the class membership of unseen samples (e.g. normal versus cancerous tissues). A common step in gene selection methods is to rank the genes according to some importance measure and then select the genes with the highest score for further analysis (Golub *et al*., 1999; Guyon *et al*., 2002). By excluding irrelevant genes it is hoped that prediction accuracy is enhanced and cancer-related genes are highlighted.

However, several microarray studies addressing similar prediction tasks report different sets of predictive genes (Ein-Dor *et al*., 2006; Lossos *et al*., 2004). For example, two prominent studies have aimed to predict development of distant metastases within 5 years, van't Veer *et al*. (2002) and Wang *et al*. (2005). Both studies came up with successful predictive gene sets (70 and 76 genes, respectively), yet with only three common overlapping genes. These findings raise the obvious question: What is the reason for this discordance between independent experiments? The trivial answer attributes this lack of agreement to biological differences among samples of different studies (e.g. age, disease stage), heterogeneous microarray platforms (spotted cDNA arrays versus synthesized oligonucleotide arrays), differences in equipment and protocols for obtaining gene expression measurements (e.g. washing, scanning, image analysis) and differences in the analysis methods (Kuo *et al*., 2002; Warnat *et al*., 2005).

Recently, Ein-Dor *et al*. (2005) argued that even if the differences mentioned earlier are eliminated, the discrepancies between studies remain. They limited themselves to a single data set (van't Veer *et al*., 2002) and showed that random divisions of the data into training and test sets yield unstable ranked gene lists and consequently, different predictive genes sets are produced. Michiels *et al*. (2005), by reanalyzing data from seven published studies that attempted to predict prognosis of cancer patients, observed that within each data set there are many optimal predictive gene sets which are strongly dependent on the subset of samples chosen for training. These findings indicate that low reproducibility occurs even within a microarray data set (and not only among multiple data sets) and thus the disparity between data sets is not surprising.

For those interested primarily in high accuracy predictive results it is acceptable to have several different predictors. Yet, from a biological perspective, the inconsistency,

*To whom correspondence should be addressed.

or instability, of predictive gene sets may lead to disturbing interpretation difficulties. Moreover, the lack of transferability of these predictors (i.e. when one predictor generated by one study suffers from a marked decrease in its performance when tested on data of another study), as reported by Ein-Dor *et al.* (2006), implies a lack of reliability in terms of robustness, undermining the generalization power of the predictor in hand.

The reason for this instability phenomenon, according to Somorjai *et al.* (2003) is the combination of the *'curse of data set sparsity'* (the limited number of samples) with the *'curse of dimensionality'* (the number of genes is very large). Microarray data sets are sensitive to both 'curses' since a typical microarray experiment include thousands of genes but only limited number of samples. Ein-Dor *et al.* (2006) assessed that several thousands of patients are required, for the data set of van't Veer, to obtain an overlap of 50% between two predictive gene sets. Unfortunately, obtaining such a large number of samples is currently prohibitive due to limited tissue availability and financial constraints.

A more readily way to increase sample size is to integrate microarray data sets obtained from different studies addressing the same biological question. Several transformation methods have been proposed to translate gene expression measurements from different studies into a common scale and thus allow the unification of these studies (Jiang *et al.*, 2004; Warnat *et al.*, 2005). Nevertheless, there is no consensus or clear guidelines as to the best way to perform such a data transformation. An alternative approach for integrating gene expression values into one large data set is to combine the analysis results of different studies that address similar goals. In principle, the utilization of such meta-analysis methods can lead to the identification of reproducible biomarkers, eliminating study-specific biases. Such a comparison can reduce false positives (i.e. genes that are differentially expressed but do not underlie the observed phenomenon) and lead to more valid and more reliable results. Following this line, previous studies have applied meta-analysis methods to the analysis of cancer microarray data. These methods aimed at both identifying robust signatures of differentially expressed genes in a single cancer type (Choi *et al.*, 2003; Rhodes *et al.*, 2002) and finding commonly expressed gene signatures in different types of cancer, across multiple data sets (Rhodes and Chinnaiyan, 2005).

This study presents a meta-analysis of two publicly available cancer microarray data sets of normal and cancerous lung tissues (Beer *et al.*, 2002; Bhattacharjee *et al.*, 2001). The analysis identifies a robust predictive gene set by jointly analyzing the two data sets and produces a transferable accurate classifier. From a methodological perspective, we propose a new predictor-based approach to overcome the instability of ranked gene lists. Based on these stable lists we demonstrate that the subset of genes identified by our meta-analysis method is superior in terms of transferability to a third unseen data set (Garber *et al.*, 2001), compared with the outcome of analyzing each data set separately. The end result is hence a predictive gene set which is able to better distinguish normal from cancerous lung tissues.

2 METHODS

2.1 Overview and definitions

A common task in gene expression analysis usually involves the selection of relevant genes for sample classification. Since most of the genes are not related to the classification problem, gene selection methods are used to rank the genes according to their importance to the biological question underlying the experiment and generate *ranked genes lists*. The genes eventually selected for classification are small subsets of the genes at the top of the ranked gene lists which we refer to as *predictive gene sets*. It has been previously shown (Ein-Dor *et al.*, 2005) that the ranked gene lists are unstable and strongly depend on the training samples from which they were produced. We refer to the latter as the *instability phenomenon*, which leads to an inconsistency of these predictive sets.

The meta-analysis method presented in this work aims to identify a robust predictive gene set by jointly analyzing two independent gene expression data sets. The first stage of our method is to create stable ranked gene lists for each of the data sets separately. This is achieved by producing many different predictive gene sets (using different random partitions of the data and cross validation) and ranking the genes according to their *repeatability frequency* in the ensemble of predictive gene sets (i.e. the frequency of appearance of each gene in the different predictive gene sets). The resulting aggregated ranked gene list is denoted the *repeatability-based gene list* (RGL). The *gene core-set* of the data set includes all genes with a non-zero repeatability score (i.e. appearing in at least one predictive gene set). The core-set genes are ranked based on their repeatability frequency in the RGL.

The second stage of our method addresses the integration of two microarray experiments originating from different studies. This stage generates the *joint core* of genes which includes genes that appear in the intersection of the gene core-sets of both data sets. The genes in the joint core are ranked such that genes with relatively high repeatability frequencies in both data sets are positioned at the top of the *ranked joint core*.

2.2 Data sets

The study includes three lung cancer microarray data sets (Beer *et al.*, 2002; Bhattacharjee *et al.*, 2001; Garber *et al.*, 2001). All data sets were downloaded from publicly available websites. Table 1 summarizes the content of the data sets, naming them according to the university in which they were constructed. Only adenocarcinoma tumors and normal

Table 1. The data sets used in the analysis

Data set	Microarray platform	Probe sets	Cancer samples	Normal samples
Michigan (Beer *et al.*, 2002)	Affymetrix (Hu6800)	7127	86	10
Harvard (Bhattacharjee *et al.*, 2001)	Affymetrix (HG_U95Av2)	12 600	139	17
Stanford (Garber *et al.*, 2001)	Spotted cDNA	24 000	41	5

lung samples are included in the analysis. Additional information regarding the adenocarcinoma samples used in the analysis is given in the Supplementary information Table S1. A detailed description of data preprocessing and probe-set filtering is also provided in the Supplementary information.

2.3 Constructing a predictive model

The data is randomly divided into two sets: 80% of the samples are assigned into a 'working' set and 20% of the samples are assigned into a validation set. The proportion of normal and cancer samples in the working and validation sets is adjusted to the proportion in the complete data set. The working set is used to identify a predictive gene set (as described in the subsequent subsection) and based on it a predictive model is constructed by training a support vector machine (SVM) classifier. The classification performance of the model is then evaluated on the validation set.

2.4 Constructing predictive gene sets

Predictive gene sets are produced by two main stages: defining the number of genes required for classification and selecting the genes involved.

In the first stage, the optimal number of genes used for classification is tuned by a 5-fold cross validation technique which incorporates a gene selection procedure: the working set, defined in the previous subsection, is further divided into five random disjoint subsets of equal size. In an iterative procedure each subset is held out in turn for testing purposes (test set), where the other four subsets are used for gene selection and training (training set). In each fold, support vector machine recursive feature elimination (SVM-RFE) (Guyon *et al.*, 2002) is used to rank the training set genes in decreasing order by their discriminative expression pattern. An SVM classifier with a linear kernel is employed to test the success rate of various numbers of genes from the top of the list ranging from 5 to 100 (in increments of 5) on the held out set. The number of genes ultimately selected for classification, N, is the number that maximizes the average 5-fold cross validation success rate.

The second and final stage in construction the predictive gene set is to rerank all genes by applying SVM-RFE on the full working set and select the N-top genes (the method is explained in detail in the Supplementary information).

2.5 Estimating the classification performance

Since the majority of the samples in our data sets are labeled as lung cancer versus a small number of normal lungs, the classification success rate is measured by the weighted average of true positives and true negatives

$$\frac{1}{2}\times\left(\frac{TP}{TP+FP}+\frac{TN}{TN+FN}\right)$$

where TP, FP, TN and FN denote true positives, false positives, true negative and false negative, respectively.

2.6 Constructing a repeatability-based gene list and gene core-set

Based on several predictive gene sets generated by different data samplings, we construct the RGL [sampling schemes are often used to increase certainty in the gene ranking e.g (Pepe *et al.*, 2003)]. The genes in the RGL are ranked according to their repeatability, i.e. their frequency in the different generated predictive gene sets, such that genes which are most frequent are at the top of the list. Whenever a gene is represented by multiple probe sets, we kept for further analysis only one probe set which exhibits its maximal repeatability frequency. We should note that RGLs which are based on a standard bootstrapping scheme to construct the predictive gene sets are similar to RGLs produced by the resampling method used in our analysis (see section 2.3 in the Methods). The Spearman correlation between the RGL lists produced by the two resampling methods is 0.89 and 0.86 for the Michigan and Harvard RGLs, respectively.

2.7 Ranking the joint core genes

The joint core genes are associated with two scores of repeatability frequency originating from the two RGLs obtained from the independent data sets. To rank the joint core genes (which appear in both gene core-sets) we first sort the repeatability scores obtained from the two independent lists leading to one unified list in which each gene appears twice. The ranking of each gene in the joint core is based on averaging the two positions of the gene in the unified sorted list.

3 RESULTS

3.1 Unstable ranked gene lists in a tumor versus normal binary classification task

It has been previously shown that the instability problem occurs in complex bioinformatics challenges such as finding prognostic gene signatures (Ein-Dor *et al.*, 2005; Michiels *et al.*, 2005). Ranked gene lists produced in these studies were unstable and depended strongly on the subgroups of patients on which they were generated. We show that the instability problem is also observed in simpler questions like classification of tumor versus normal tissues. Figure 1 demonstrates the instability of the ranked gene lists constructed from repeatedly applying SVM-RFE to the gene expression profiles of different subgroups of patients drawn at random from the Michigan and Harvard data sets separately. Evidently, genes which are ranked high using one subgroup of patients may be ranked low in another (as evident in both data sets).

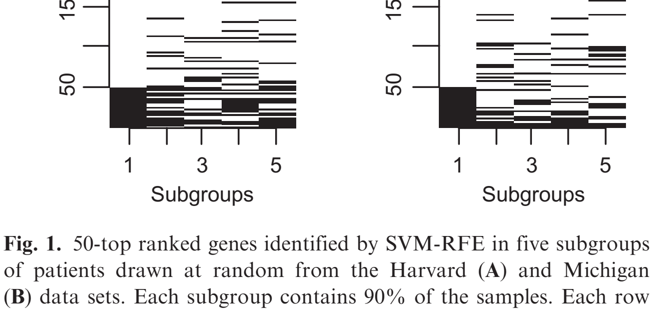

Fig. 1. 50-top ranked genes identified by SVM-RFE in five subgroups of patients drawn at random from the Harvard (**A**) and Michigan (**B**) data sets. Each subgroup contains 90% of the samples. Each row represents a gene and each column represents a different subgroup of patients. The genes are ordered by the leftmost column and the top 50 genes are marked by a line.

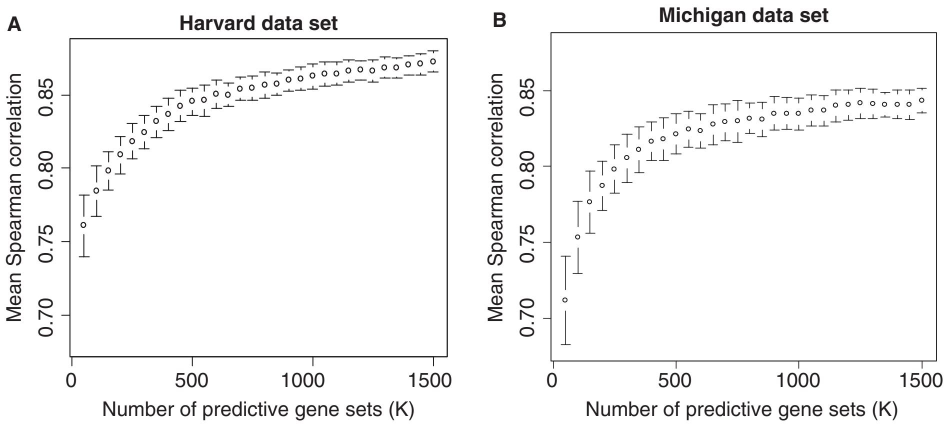

Fig. 2. Assessing the consistency of RGLs in the Harvard (**A**) and Michigan (**B**) data sets. The *x*-axis represents the number of predictive gene sets (K) and the *y*-axis represents the mean Spearman correlation between two RGLs produced over 100 samplings (see Methods section). SDs for each number of predictive gene sets are marked.

3.2. Constructing a consistent repeatability-based gene list

Our first challenge is to produce a consistent gene ranking method. Since our ranking procedure uses random samplings of the data (see Methods section) and hence is not deterministic, it is necessary to determine the number of predictive gene sets, *K*, sufficient for obtaining a consistent RGL. To this end, we repeat the gene ranking procedure twice, each time using *K* predictive gene sets, producing two different RGLs.

RGL consistency is evaluated by calculating the Spearman correlation between these two resulting RGLs. A high Spearman correlation obviously testifies to high consistency levels. This consistency test is performed for a varying *K* values, ranging from 50 to 1500 in intervals of 50. The resulting mean Spearman correlation increases with the number *K* of predictive gene sets used (Fig. 2). Throughout this work we use $K = 1000$, which evidently yields a consistent ranking. With $K = 1000$ the Harvard data set exhibits a mean Spearman correlation coefficient of 0.86 with a SD of 0.008, while the Michigan data set manifest a mean of 0.84 with a SD of 0.01.

Furthermore, the predictive gene sets which construct the RGLs reach high classification success rates. Mean success rates are 90% and 98.6% for the Harvard and Michigan data sets, respectively, testifying to the utility of the RGLs. The mean number of genes participating in a predictive gene set is 27.8 and 15.8 for the Harvard and Michigan data sets, respectively with SDs of 24.3 and 17.5.

Investigating the two RGLs, we observe that ~90% of the genes in both data sets do not participate in any of the predictive gene sets. Out of 4579 genes included in the two data sets, 547 genes comprise the gene core-set of the Harvard data set and 411 genes comprise the gene core-set of the Michigan data set. The distribution of the repeatability frequency in the core-sets of both data sets is presented in Supplementary Figure S2 (Supplementary information).

3.3 Repeatability-based gene lists are stable

A stable ranked gene list is unsusceptible to random partitioning of the data. The stability of RGLs produced for the Harvard and Michigan data sets is examined in Figure 3.

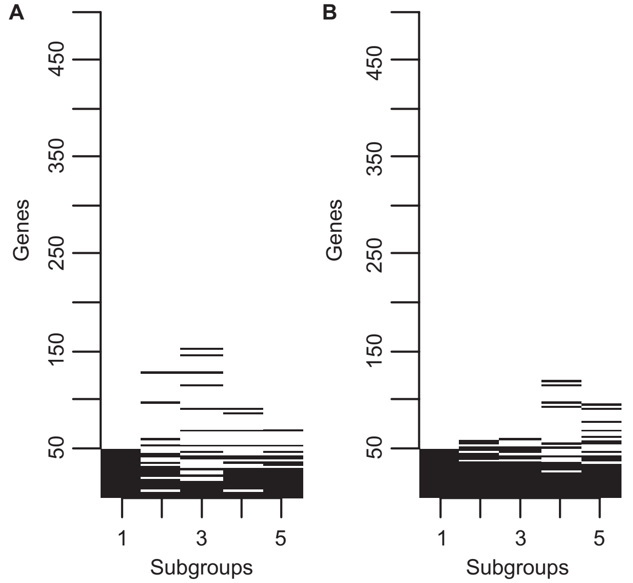

Fig. 3. 50-top ranked genes in five different RGLs produced by five random subgroups of patients drawn from the Harvard (**A**) and Michigan (**B**) data sets. Each subgroup contains 90% of the data. Figure layout is similar to Figure 1.

In contrast to the large variation in membership of the top 50 genes based on gene rankings by SVM-RFE (Fig. 1), the top 50 genes in the RGLs are reproducible. The mean overlap between the 50 top ranked genes of the different RGLs is 37 and 40.6 for the Harvard and Michigan data sets, respectively with SDs of 2.86 and 3.23, while the mean overlap between the 50 top ranked genes when using SVM-RFE as a ranking method (Fig. 1) is 24.1 and 26.8 for the Harvard and Michigan data sets, respectively with SDs of 8.34 and 9.54. These results suggest that indeed RGLs are stable, robust lists (it may be noted, however, that in the improbable case of identical data partitions, our method obviously leads to a less stable ranking than SVM-RFE, as the latter is deterministic). A comparison

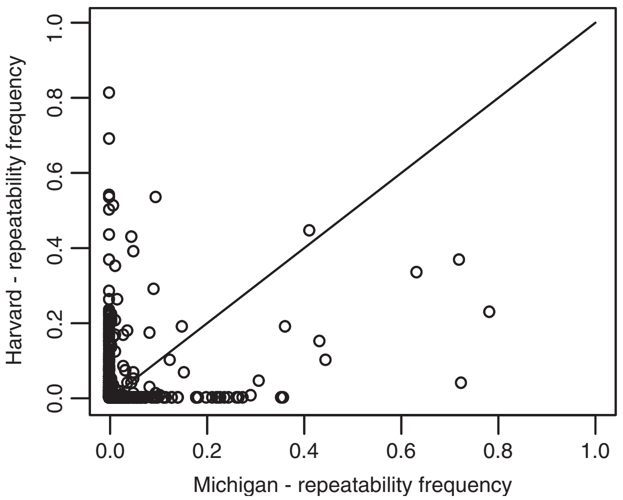

Fig. 4. Comparison of gene repeatability frequency between the Michigan and Harvard data sets. Each point represents a gene and its repeatability frequency in the Michigan data set (*x*-axis) versus its repeatability frequency in the Harvard data set (*y*-axis). The diagonal marks the position of genes which have equal repeatability frequencies in both data sets. A gene's repeatability frequency is given as the fraction out of the maximal 1000 repeats possible.

between the RGLs and the rankings produced by SVM-RFE is provided in the Supplemetary information.

3.4 Comparing gene rankings between data sets

Since the RGL ranking in each data set separately (Harvard versus Michigan) is rather stable one could expect that genes that are highly discriminative in one data set would be also highly discriminative in the second data set. Interestingly, this is not the case: the diagonal in Figure 4 marks the position of genes whose repeatability frequency is equivalent in both data sets. Evidently, only few points are located around the diagonal, whereas most points exhibit significant dissimilarity in their repeatability frequencies over the two data sets. Six out of the 10 top ranked genes in Harvard core-set do not appear in Michigan core-set, suggesting that these genes are 'data set specific' and may not be truly reflective of the underlying disease process. The top ranked genes in Michigan core-set are quite highly ranked in the Harvard core-set (eight out of the 10 top ranked genes in Michigan core-set appear in Harvard core-set). These genes are reproducible across the studies, testifying to their reliability.

The dissimilarity between data sets is also demonstrated by the low Spearman correlation of 0.173 between the RGLs of the Harvard and Michigan data sets.

3.5 Joint core magnitude

Since our goal is to examine whether relevant genes can be more effectively discovered by jointly analyzing two independent data sets, we focus on the joint core genes (obtained as described in the Methods section). The magnitude of the joint core of Michigan and Harvard data sets is 118 genes and is statistically significant ($P < 0.0025$ as none of the permutation runs reached the true joint core magnitude, see Supplementary information). The magnitude of the joint core remains significant across a variety of repeatability frequency thresholds used to determine the genes in the core-sets (Supplementary information Figure S3).

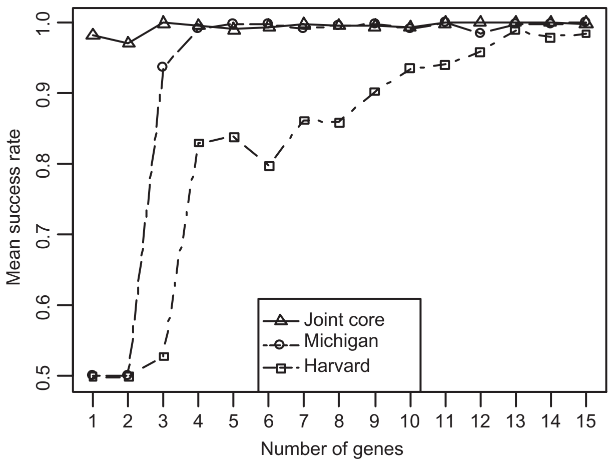

Fig. 5. The mean success rate of the top ranked genes of joint core (triangles), Michigan core-set (circles) and Harvard core-set (squares) on the Stanford data set. The top ranked genes include only the genes that appear in the Stanford data set. The *x*-axis represents the number of genes utilized by the classifier. For each number of selected genes the procedure is carried out 100 times, on different samplings of the Stanford data into training and test sets. The *y*-axis represents the mean success rate.

3.6 The joint core is transferable

A question remains, is the joint core more informative than the two independent core-sets? A pertaining test would investigate the transferability of these cores; that is, do they carry predictive information as for a new unseen data set, preferably even from a different technology? To this end, we test the classification performance of the different cores on the Stanford data set, an independent cross-platform microarray data of lung cancer. To evaluate the classification performance obtained with genes from the three cores (Harvard and Michigan core-sets and the joint core) on the Stanford data set, an SVM classifier is utilized in a standard train and test procedure. This procedure is repeated for an increasing number of genes selected from the top of the three ranked cores. This enables us to compare the classification performance of the ranked cores for the same number of genes each time.

The results show that the joint core outperforms the two independent core-sets, obtaining a high level of classification already with a very small number of highly ranked genes (<4). As observed in Figure 5, the first gene on the top ranked joint core (RAGE), achieves a high success rate of 98% on its own. The Michigan data set matches the joint core performance with four genes only where the Harvard data set requires the top 13 genes to match the joint core performance. As observed in Table 2, listing the genes in the different cores, the majority of genes in the top of the Harvard set are not in the joint core

Table 2. The 10-top ranked genes of the joint core, Michigan core-set and Harvard core-set

	Joint core	Michigan core-set	Harvard core-set
1	RAGE	TNXB	**SMAD6**
2	TNA	CA4	**GRK5**
3	FABP4	RAGE	**HYAL2**
4	TNXB	FABP4	TEK
5	*COX7A1*	FGR	**CD34**
6	PHLDA2	PHLDA2	*S100A3*
7	FGR	TNA	**FKBP1A**
8	TEK	*COX7A1*	TNA
9	TACSTD1	**CEACAM5**	**TLK1**
10	MAP4	**CASP1**	EMP2

Genes that do not appear in the joint-core are marked in bold. Genes that do not appear in the Stanford data set are marked in italics.

while in the Michigan set this is not the case. Interestingly, a marked increase in the Harvard set's success rate is reached (83%) by adding the fourth gene (TEK), which is the first in the Harvard list to appear in the joint-core.

3.7 Biological significance of the joint core genes

We turn to examine the biological function of the 118 genes composing the joint core, concentrating on their role in cancer. In a prominent review by Hanahan *et al.* (Hanahan and Weinberg, 2000), tumorigenesis is presented as a multistep process which manifests several essential alterations in cell physiology; these constitute the 'hallmarks of cancer'.

In the joint core, several representatives of each of these required alterations are found, as reported briefly below (their rank in the joint core is indicated in parentheses):

3.7.1 Self-Sufficiency in growth signals In cancer cells many oncogenes activate normal growth signaling pathways which yield uncontrolled proliferation (Hanahan and Weinberg, 2000). A putative representative of this class is ErbB3 (rank 72), belonging to the epidermal growth factor (EGF) receptor subfamily, which was shown to constitute a growth stimulatory loop particularly for non-small cell lung cancer (NSCLC) (Fong *et al.*, 2003).

3.7.2 Insensitivity to antigrowth signals Reduced expression of TGFβ receptor type III (TGFBR3, rank 36), an antiproliferative signal, is known to be associated with resistance to TGFβ and may play a role in tumorigenesis (Copland *et al.*, 2003).

3.7.3 Evading apoptosis The identified joint core is comprised of several genes related to apoptosis as indicated by their Gene Ontology class and KEGG pathway (Dennis *et al.*, 2003): PHLDA2 (rank 6), SPP1 (rank 21), ZBTB16 (rank 32), DNASE1L3 (rank 38), CSF2RB (rank 60), PML (rank 80), IGFBP3 (rank 81), TNFRSF25 (rank 82).

3.7.4 Sustained angiogenesis Several genes in the joint core are related to angiogenesis: TEK (TIE-2, rank 8), MDK (rank 15), EDNRB (rank 23), PECAM1 (CD31, rank 24) and ANG1 (rank 35), CDH5 (rank 65) (Ahmed *et al.*, 2000; Choudhuri *et al.*, 1997; Liao *et al.*, 2000; Takahama *et al.*, 1999). Interestingly, there may be a clinical potential in targeting these genes' pathways by producing anti-angiogenic agents. For example, it has been shown that blocking the TIE-2/ANG1 pathway inhibits, to a certain extent, tumor angiogenesis (Takahama *et al.*, 1999).

3.7.5 Tissue invasion and metastasis RAGE, the top ranked gene in the joint core list, was shown to be involved in motility and invasive behavior of cells. Furthermore, inhibition of RAGE-amphoterin signaling suppressed tumor growth and metastases in mice (Taguchi *et al.*, 2000). S100A4 (rank 94) is thought to mediate motility and invasiveness of cancer cells. It is a marker for poor patient prognosis in number of cancers (Li and Bresnick, 2006). Three other members of the S100 family were found to be in the joint core genes: S100A3 (rank 18), S100G (rank 30), S100A8 (rank 52). This fact may suggest an association between this family and lung cancer. Other genes in the joint core which are related to tissue invasion and metastases include: CAV1 (rank 13), SPP1 (rank 21) and SPINT2 (rank58) (Ho *et al.*, 2002; Rangaswami *et al.*, 2006; Suzuki *et al.*, 2003).

Obviously, though it is encouraging to find quite a few of these potentially involved genes in the joint core, their role in the actual pathogenesis of adenocarcinoma tumors remains to be explored. We additionally compared the joint core genes found by our analysis to the results of Jiang *et al.* (2004), a study with similar goals and applied to the same data sets (Harvard and Michigan). Jiang *et al.*, aiming to identify marker genes which are capable of differentiating adenocarcinoma from normal lung, discovered 13 and 10 marker genes by applying two different models to the data (five genes common to both models). Out of the union of 18 genes they identified 10 appear in our joint core, and out of the five genes common to both models four of them appear in the joint core (these results are evidently highly significant, $P<10^{-6}$). This high overlap reinforces the probability that genes in the joint core may be more reliable to the biology and diagnosis of adenocarcinoma tumors. However, one eminent complicating factor is that some of these genes (and in the joint core in general) may be overexpressed and some underexpressed (compared with their expression in normal, healthy tissue), and in parallel, their putative repressing/activating role in the molecular pathways which they belong to is in many cases still unknown.

4 DISCUSSION

A key component of gene-expression analysis is the identification of genes that play a pivotal role in the biological processes underlying the microarray experiment. With the increasing availability of microarray data sets there is a growing need for integrative computational methods that evaluate multiple independent microarray data sets. Meta-analysis methods are applied to reduce study-specific biases, aiming to yield results which offer improved reliability and validity. Here we propose a predictor-based meta-analysis approach that generates a robust predictive gene set. The method has its roots in ensemble learning methods frequently used in prediction and classification, where the underlying base learning algorithm is run multiple times, and a vote is taken on the resulting hypotheses. As confirmed experimentally in numerous cases by now, ensemble methods can efficiently reduce both the bias and the variance of learning algorithms and improve their overall accuracy (Dietterich, 2002). Using this method the genes are first ranked on different data sets, independently, according to their classification power, and then they are combined into a consolidated gene set, the joint core genes. Doing so, we address two main challenges; (1) *The instability problem:* When dividing a given data set into training and test sets, different divisions produce different ranked gene lists which subsequently give rise to different predictive gene sets. We show that this phenomenon is not restricted to complex computational challenges such as finding a prognostic gene signatures (Ein-Dor *et al*., 2005; Michiels *et al*., 2005), but is also observed in less challenging questions like binary classification of tumor versus normal tissues. Assuming that genes which are more essential for classification will appear more consistently in different predictive gene sets, we construct a ranked gene list termed RGL. The RGL demonstrates high stability, with an average overlap of approximately 39 genes between the top 50 genes of two different RGLs, generated from independent data divisions. (2) *Transferability*: How well do features learned in the context of one data set perform on a second, unseen, data set? Our results show successful transferability of the joint core genes to the unseen Stanford data set, in which the top three genes of the ranked joint core yield a classifier with an accuracy of 99.8%.

Applying the suggested gene ranking method to two prominent lung cancer data sets, the Michigan and Harvard data sets, results in a low Spearman correlation ($r = 0.173$) between the two RGLs, although each list by itself is stable. Moreover, genes exhibiting high classification power on one of the data sets (and thus were ranked at the top of the RGL) were ranked at the bottom of the corresponding RGL of the second data set. Observing that the two independent RGLs produced by our meta-analysis method are stable but exhibit a significant dissimilarity, leads us to attribute this dissimilarity to factors like biological differences among samples of different studies, differences in platform generation and differences in protocols, rather than to inner instability.

The joint core constructed by the meta-analysis approach focuses on genes which appear in the core-sets of both data sets, and hence are likely to be central to the phenomenon studied. The first gene in the ranked joint core, RAGE, exhibits a very high classification performance by itself. RAGE was shown to be strongly down-regulated in NSCLC patients compared to their paired normal lung tissues, not only on the transcriptional level (as revealed by this study) but also on a protein level (Schraml *et al*., 1997). These results may suggest RAGE as a potential marker for diagnosis of lung cancer.

Studying the transferability to the Stanford data set confirms that genes which are highly ranked only in one data set but are not part of the joint core are biased to their data set and thus exhibit low transferability. The joint core indeed shows improved transferability, demonstrating high classification even with a very small number of genes from the top of the ranked joint core. Although the joint core has better classification capability than the two separate cores, the joint core does not show a significant similarity to the Stanford core set. This may be due to the variation in platforms from which the data sets were produced (discussed in the Supplementary information).

The analysis method demonstrated in this study increases the reliability of identifying powerful predictive genes sets. The putative list of predictive genes identified may hold promise as therapeutic targets and diagnostic markers. Applying the method to other data sets and expanding the method beyond two data sets may enhance our biological understanding of previous microarray studies, with no extra experimental work.

ACKNOWLEDGEMENTS

We thank Isaac Meilijson, Gideon Dror, Yoav Benjamini, Danny Yekutieli and Yoram Oron for their valuable comments and suggestions. A. Kaufman is supported by the Yeshaya Horowitz Association through the Center of Complexity Science. E. Ruppin's research is generously supported by the Tauber Fund, the Yeshaya Horowitz Association through the Center of Complexity Science, the Israeli Science Fund (ISF) and the Israeli Ministry of Science and Technology (MOST).

Conflict of Interest: none declared.

REFERENCES

Ahmed,S.I. *et al*. (2000) Studies on the expression of endothelin, its receptor subtypes, and converting enzymes in lung cancer and in human bronchial epithelium. *Am. J. Respir. Cell Mol. Biol*., **22**, 422–431.

Beer,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med*., **8**, 816–824.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, **98**, 13790–13795.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.

Choudhuri,R. *et al*. (1997) An angiogenic role for the neurokines midkine and pleiotrophin in tumorigenesis. *Cancer Res*., **57**, 1814–1819.

Copland,J.A. *et al*. (2003) Genomic profiling identifies alterations in TGFbeta signaling through loss of TGFbeta receptor expression in human renal cell carcinogenesis and progression. *Oncogene*, **22**, 8053–8062.

Dennis,G. *et al*. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biology*, **4**, P3.

Dietterich,T. (2002) Ensemble learning. In *The Handbook of Brain Theory and Neural Networks*. MIT press, Cambridge, MA, pp. 405–408.

Dopazo,J. *et al*. (2001) Methods and approaches in the analysis of gene expression data. *Journal of Immunological Methods*, **250**, 93–112.

Ein-Dor,L. *et al*. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Ein-Dor,L. *et al*. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *PNAS*, **103**, 5923–5928.

Fong,K.M. *et al*. (2003) Lung cancer * 9: molecular biology of lung cancer: clinical implications. *Thorax*, **58**, 892–900.

Garber,M.E. *et al*. (2001) Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, **98**, 13784–13789.

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

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Ho,C.-C. *et al*. (2002) Up-regulated caveolin-1 accentuates the metastasis capability of lung adenocarcinoma by inducing filopodia formation. *Am. J. Pathol*., **161**, 1647–1656.

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

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

Li,Z.-H. and Bresnick,A.R. (2006) The S100A4 metastasis factor regulates cellular motility via a direct interaction with myosin-IIA. *Cancer Res*., **66**, 5173–5180.

Liao,F. *et al*. (2000) Monoclonal antibody to vascular endothelial-cadherin is a potent inhibitor of angiogenesis, tumor growth, and metastasis. *Cancer Res*., **60**, 6805–6810.

Lossos,I.S. *et al*. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med*., **350**, 1828–1837.

Michiels,S. *et al*. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Pepe,M.S. *et al*. (2003) Selecting differentially expressed genes from microarray experiments. *Biometrics*, **59**, 133–142.

Rangaswami,H. *et al*. (2006) Osteopontin: role in cell signaling and cancer progression. *Trends in Cell Biology*, **16**, 79–87.

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. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet*, **37**, S31–S37.

Schraml,P. *et al*. (1997) Differential messenger RNA and protein expression of the receptor for advanced glycosylated end products in normal lung and non-small cell lung carcinoma. *Cancer Res*., **57**, 3669–3671.

Shipp,M.A. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med*., **8**, 68–74.

Somorjai,R.L. *et al*. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**, 1484–1491.

Suzuki,M. *et al*. (2003) Bikunin target genes in ovarian cancer cells identified by microarray analysis. *J. Biol. Chem*., **278**, 14640–14646.

Taguchi,A. *et al*. (2000) Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases. *Nature*, **405**, 354–360.

Takahama,M. *et al*. (1999) Enhanced expression of Tie2, its ligand angiopoietin-1, vascular endothelial growth factor, and CD31 in human non-small cell lung carcinomas. *Clin. Cancer Res*., **5**, 2506–2510.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wang,Y. *et al*. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Warnat,P. *et al*. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265.