

Gene expression

Reliable gene signatures for microarray classification: assessment of stability and performance

Chad A. Davis[†], Fabian Gerick[†], Volker Hintermair[†], Caroline C. Friedel, Katrin Fundel, Robert Küffner and Ralf Zimmer*

Institute of Informatics, Ludwig-Maximilians-Universität München, Amalienstrasse 17, 80333 Munich, Germany

Received on January 10, 2006; revised on June 21, 2006; accepted on July 18, 2006

Advance Access publication July 31, 2006

Associate Editor: Alvis Brazma

ABSTRACT

Motivation: Two important questions for the analysis of gene expression measurements from different sample classes are (1) how to classify samples and (2) how to identify meaningful gene signatures (ranked gene lists) exhibiting the differences between classes and sample subsets. Solutions to both questions have immediate biological and biomedical applications. To achieve optimal classification performance, a suitable combination of classifier and gene selection method needs to be specifically selected for a given dataset. The selected gene signatures can be unstable and the resulting classification accuracy unreliable, particularly when considering different subsets of samples. Both unstable gene signatures and overestimated classification accuracy can impair biological conclusions.

Methods: We address these two issues by repeatedly evaluating the classification performance of all *models*, i.e. pairwise combinations of various gene selection and classification methods, for random subsets of arrays (*sampling*). A model score is used to select the most appropriate model for the given dataset. Consensus gene signatures are constructed by extracting those genes frequently selected over many samplings. Sampling additionally permits measurement of the stability of the classification performance for each model, which serves as a measure of model reliability.

Results: We analyzed a large gene expression dataset with 78 measurements of four different cartilage sample classes. Classifiers trained on subsets of measurements frequently produce models with highly variable performance. Our approach provides reliable classification performance estimates via sampling. In addition to reliable classification performance, we determined stable consensus signatures (i.e. *gene lists*) for sample classes. Manual literature screening showed that these genes are highly relevant to our gene expression experiment with osteoarthritic cartilage. We compared our approach to others based on a publicly available dataset on breast cancer.

Availability: R package at <http://www.bio.ifi.lmu.de/~davis/edaprakt>

Contact: ralf.zimmer@bio.ifi.lmu.de

INTRODUCTION

Microarrays have become a standard means of investigating different states of biological systems on the basis of the expression of genes on a genome-wide level. A broad range of statistical analysis

methods have been developed to guide the biological interpretation of these data. Typical first questions on measurements with various sample classes, e.g. tissues or disease states, are (1) how can a particular sample be classified, i.e. assigned to a sample class with high reliability? and (2) what are the molecular features best representing the differences between classes, sample subsets or individual samples? The answers to these questions could help towards a first interpretation of the data, i.e. could improve the analysis of future measurements and their assignment to a specific sample class, which could suggest further work or experiments. Moreover, gene or feature lists could provide a starting point for understanding the processes responsible for the differences between states, possibly indicating and identifying well-known and new target genes.

A wide variety of machine learning methods have been proposed for classification tasks related to microarrays, including support vector machines (SVM), *k* nearest neighbors (*k*NN), decision trees (DT) and many others (Dudoit *et al.*, 2002). Compared to the number of samples, the number of features (i.e. genes or spots represented on the microarray) is large, which affects the performance of the classifiers. Therefore, a number of feature subset selection (FSS) methods have been developed for gene selection in microarray data (for a review see Guyon and Elisseeff, 2003). However, using an arbitrarily fixed combination of FSS method and classifier (a *model*) may sacrifice performance that could have been achieved with another model. The systematic pairwise combination of FSS methods with classifiers produces models with varying performance and, thus, a ranking of possible models.

Often, one is tempted to rely on a best classification method, with a certain performance, or to restrict the analysis to the top ranking genes/features. A method to resolve these issues by comprehensive evaluation and selection of models has been proposed by Statnikov *et al.* (2005). Another difficulty due to the relatively small number of samples is the instability of the performance of models that are simply built upon fixed gene signatures. This is due to the fact that gene signatures depend heavily on the actual dataset and are especially sensitive to noise (Guyon and Elisseeff, 2003). Repeated random sampling of arrays before feature selection can be used to assess the quality of FSS methods with regard to signature stability (Bi *et al.*, 2003). For a recent discussion of issues in published studies see also Michiels *et al.* (2005).

The focus of this paper is to combine these strategies by selecting models that exhibit not only good classification performance but also stable signatures.

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Our approach begins by selecting an appropriate model from a predefined library of commonly used models. Each model in the library is a combination of a feature selection method, a classifier and fixed parameters. Our model selection procedure then evaluates all models from this library for the given dataset. We repeat this evaluation over several random samplings of expression arrays to be analyzed. On one hand, this permits an estimation of the quality of the FSS methods via the stability of their gene signatures. On the other hand, it enables an estimation of the quality of the classifiers via the consistency of their classification performance. An additional cross validation has been performed to estimate the degree of overfitting that occurs within the random samplings during the model selection procedure to provide an estimation of the expected performance as well as an extra indication of the quality of the signature.

METHODS

In order to determine reliable features and stable classification accuracies, together with an estimate of the overall performance, we propose the following procedure (StabPerf), outlined in Figure 1. For given sets of gene expression measurements (arrays) for the respective classes, we use sampling to select gene signatures and classify arrays for all combinations of FSS and classification methods. Parameter combinations for individual models (i.e. the combination of feature subset selection, classifier and fixed *parameters*) need to be specified beforehand for the given dataset, i.e. an optimization in parameter space is not performed by our method. Nevertheless, sensible parameters need to be selected so that the feature selection methods do not produce empty sets. We define a new score to rank these models.

In the first step of StabPerf, a group of gene expression arrays is randomly split into a *training set* (6/7 of the arrays) and a *validation set* (1/7 of the arrays). This splitting is repeated several times (*sampling*), whereby the split in each iteration is randomized such that the classes are represented in each training set in the same proportions as they are in the complete dataset (*balanced sampling*).

Subsequently, for each training set an FSS method determines a list of relevant features (*gene signature*).

Then, for each gene signature from each FSS method, each classification method is trained on the respective training set, using only the features in the given signature. Accuracies for these combinations of FSS method and classifier are measured by predicting the outcome on the corresponding validation sets.

Next, we rank all possible models based on the stability of sampled gene signatures (signature stability) as well as the classification performance of the model. As multiple classification accuracies were calculated for each model, not only the total accuracy (fraction of correct predictions over all predicted samples) but also the median absolute deviation (MAD) of classification accuracies, over all sampling steps, is computed. We refer to the classification accuracy and its MAD as the *performance* of the classifier. The gene signature stability and the deviation of the classification accuracies are two important factors, neglected by many current approaches, both of which are used to select the best model (*model selection*). Finally, the chosen model is retrained on all arrays.

Classification performance of our model selection approach is estimated via stratified 10-fold cross validation (Hastie *et al.*, 2001) since the accuracy on the validation set was used during model selection and, thus, is biased.

Model selection

The optimal model, i.e. combination of an FSS method with a classification method, is chosen based on the stability of the signatures produced by the FSS method as well as the distribution of classification accuracies.

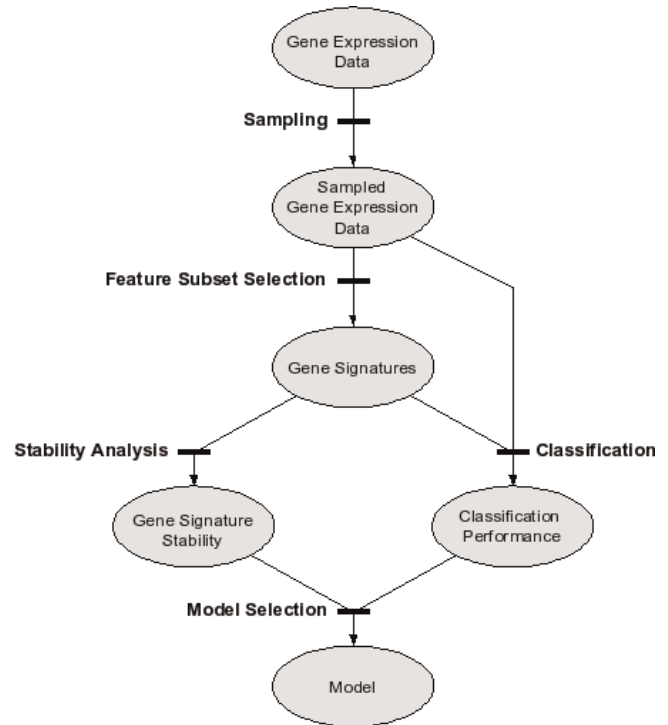


Fig. 1. Overview of the analysis pipeline StabPerf (STABLE model selection by optimizing reliable classification PERFORMANCE). Random sampling from the given gene expression dataset produces multiple training subsets. Gene signatures that separate the classes are derived by FSS methods for each training set and are used to train a classifier. The stability of the gene signatures and the classification performance on the validation set are used to select a best model, based on combinations of FSS and classification method.

Since classifiers can be sensitive to noise as well as over-fitting, we introduce an additional measure into our model selection score that considers the stability of an FSS method. We prefer FSS methods that produce stable gene signatures, because features which are frequently selected over many different training sets are expected to be the most biologically relevant ones with respect to the differences between sample classes. The model score can be parameterized to select models specific to the needs of particular users.

For a given FSS method S , let F be the list of all features, which have been selected in at least one of n sampling steps, i.e. for at least one training subset. Let $freq(f)$ be the number of sampling steps in which a feature $f \in F$ has been selected. The non-adjusted stability $Stab_{NA}$ of the FSS method is conceptually an inverse variance:

$$Stab_{NA}(S) = \frac{\sum_{f \in F} (freq(f)/n)}{|F|}. \quad (1)$$

We introduce the length adjusted stability $Stab$, as the non-adjusted stability does not yet account for the artificial increase in stability that occurs with increasingly long gene signatures. For example, consistently producing the signature containing every feature on the array would result in 100% stability. Therefore, non-adjusted stability is penalized by the median number of selected features μ , as a fraction of the total number of features per array $|features|$, weighted by a penalty factor α . Here, we use $\alpha = 10$, which, with approximately 7500 features per array, effectively limits signatures to at most 750 features.

$$Stab(S) = \max[0, Stab_{NA}(S) - \alpha * (\mu/|features|)]. \quad (2)$$

Second, we define the classification performance $Perf(M)$ for a given model $M = (S, C)$, i.e. combination of FSS method S and classification method C . Here, we award a model M for a high total accuracy $Acc(M)$ over all sampling steps and penalize a high median absolute deviation $MAD(M)$. This penalty is also adjustable, depending on the objectives of the analysis. We use $\beta = 0.5$:

$$Perf(M) = \max [0, Acc(M) - \beta * MAD(M)]. \quad (3)$$

The model score (in the range of [0,1]):

$$Modscore(M) = \gamma * Stab(S) + (1 - \gamma) * Perf(M), \quad (4)$$

is a weighted combination of stability and classification performance, where γ is also adjustable. We use $\gamma = 0.5$, so that both stability and classification performance contribute equally.

Consensus gene signatures

The complete procedure described here can be considered a FSS method itself, referred to as ConsGS. ConsGS uses the stability measure of an FSS method, calculated from all its gene signatures, and retains only those features that occur in more than a given fraction τ of all signatures. τ controls the sensitivity and selectivity of the method. Features selected by ConsGS are expected to be more biologically relevant, as it eliminates noise by selecting only those features that are consistently found to be significant. Furthermore, the feature frequencies over all gene signatures provide a relevance ranking for the selected features, which can be used as a starting point for evaluating candidate genes.

Feature subset selection

Various FSS methods for producing gene signatures are investigated. Here, we use the distinction made in Inza *et al.* (2004) between filter methods, which select genes based on statistical correlations between expression values and sample classes, and wrapper methods, which select genes based on their ability to separate sample classes, using a given classifier. The following filter methods were used in this study:

- **F-test (FT_t).** All features with an F -test statistic above a chosen threshold t are selected, e.g. FT₃₀ selects all features with a statistic $t \geq 30$.
- **Pearson correlation (PC).** All features are selected with an absolute Pearson correlation r between expression values and sample groups above a chosen threshold t , i.e. $|r| > t$. Alternatively, the top n genes with the highest absolute Pearson correlation coefficient (referred to as PC_(n)) are selected, e.g. PC₍₅₀₎ to select the top 50 genes.
- **P -value combined with fold change (PV _{f} , FC _{f}).** Genes are selected having both t -test P -values lower than a threshold t (default: $t = 0.001$) as well as log fold changes $|\log_2(\text{fold-change})|$ higher than a threshold f (e.g. $f = 2.5$), between any two sample groups. The fold change is calculated as the ratio of the midmeans (mean of values between the 25th and 75th percentiles) of the expression values for each pairwise combination of patient groups.

Additionally, over-representation analysis (ORA) is used for optimizing both biological as well as statistical relevance (Draghici *et al.*, 2003). ORA is widely used to analyze the distribution of GO (Ashburner *et al.*, 2000) terms within gene signatures. Moreover, approaches use GO annotations to improve the classification, e.g. Lottaz and Spang (2005). Our approach, however, uses GO/ORA only for post-processing gene lists of the aforementioned FSS methods, thereby incorporating additional biological knowledge.

The last two FSS methods are similar to wrapper methods in that they use trained classifiers to estimate the significance of genes with respect to sample class separation. They are different from wrapper methods, however, in that they do not explicitly make use of classification accuracy to determine gene significance. This is intended to avoid basing gene selection solely on maximizing classification accuracy.

- **Decision tree-based (DT_t).** A decision tree is trained and genes are selected starting at the root and moving down the tree, as long as they exceed a minimum significance threshold t ($t = 0.01$) (Breiman *et al.*, 1984).
- **SVM-based (SVM_t).** A linear SVM is trained, which assigns a normed weight to each gene. Genes having a normed weight above a threshold t ($t = 0.00001$) are selected.

Classification

We consider a number of classification methods that are commonly used in the literature:

- **NSC (Nearest Shrunken Centroid)** (Tibshirani *et al.*, 2002). A modification of the nearest centroid classifier, which assigns a sample to the class with the nearest centroid. The modification consists of ‘shrinking’ the centroids of each class toward the overall centroid of all classes.
- **k NN (5NN, with $k = 5$)** (Mitchell, 1997). A sample is assigned to a class based on a voting scheme among its k nearest neighboring samples.
- **SVM (Boser *et al.*, 1992).** Classes are separated by finding a maximal margin hyperplane between them, either in the original feature space, or in a higher dimensional space, depending on the kernel function used. We considered SVMs both with third order polynomial kernels (SVM-P) and radial basis function kernels (SVM-R). The R library used is based on LibSVM (Chang and Lin, 2001, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>).
- **DT (Breiman *et al.*, 1984).** Each internal node in the tree describes a test on a feature and has one child node for each possible value or range of values of the feature. A class label is associated with each leaf node and the classification of a sample is determined by following the path from the root to a leaf.

Data preparation

We applied our approach to a large and difficult osteoarthritis (OA) experiment with 78 single-channel cDNA expression arrays (Aigner *et al.*, 2006). Each array contains 7467 spots corresponding to 3468 genes, which represent the input features that are processed by the various FSS methods. The samples were divided into four classes according to the stages of degenerative cartilage progression, leading to a four-way classification problem. The arrays represented healthy (18 patients), early degenerative cartilage (20), peripheral OA (21) or central OA (19) states, whereby healthy and early patients have not been diagnosed with OA, and peripheral OA and central OA describe two variants of late OA. Scaled median absolute deviation (SMAD) (Dudoit and Yang, 2003) was applied as the most suitable between-array normalization method for this dataset (Fundel *et al.*, 2005).

Additionally, we evaluated a publicly available dataset on breast cancer (van't Veer *et al.*, 2002) with 78 two-channel arrays containing some 25 000 human genes.

RESULTS

Pairwise combinations of the aforementioned FSS and classification methods were examined for different numbers of sampling steps. Each of these models is evaluated in each sampling step. We found that the performance of models can be highly variable when the number of sampling steps is low (Figure 2). This shows that performance estimates are unreliable without repeated sampling. For example, the initial classification performance (with only one sampling step) of the model FT₃₀/5NN was over-estimated by ~20% points, compared to its performance after 250 sampling iterations. Performance underestimates can also be observed in Figure 2, where FT₅₀/NSC gains ~10% points, showing that

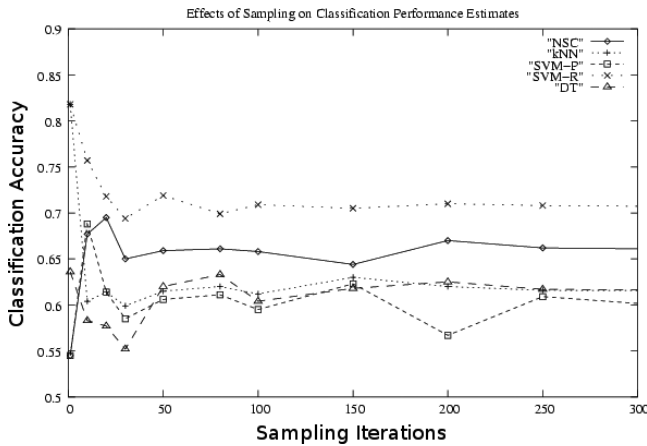


Fig. 2. Stability of classification accuracy. A large number of sampling steps leads to more stable and reliable accuracy estimations of the models. The five models (i.e. 11–15 in Table 1) shown here all use the FT₃₀ FSS method.

repeated sampling prevents both types of errors caused by variable classifier performance. Based on these observations, we set the number of sampling steps to 400, so that the performance of all classifiers becomes sufficiently stable. These 400 sampling steps produce 400 trained models for each FSS and classification method, whose accuracy distribution provides a better estimate of the reliability of the model by considering the median accuracy and the MAD of the accuracy (see Table 1).

Combining FSS and classification methods

The results of all pairwise combinations of the seven FSS and five classification methods examined here are summarized in Table 1. We see that no FSS method and classification method is consistently superior with respect to all criteria, rather the performance of the different models is highly variable. This is the principle behind the model scoring function. For example, when we compare models 27 (DT_{0,01}/5NN) and 34 (SVM_{0,00001}/SVM-R) we see that both achieve a total accuracy of 68.4%, however, model 34 is penalized for the higher MAD (18.0%) of its accuracy, while model 27 benefits from having a much higher gene signature stability (53.7%) than that of model 34 (22.4%). However, model 27 includes more genes (234) on average than model 34 (199) in order to achieve this signature stability. Nonetheless, stable gene signatures are more important than short gene signatures, causing model 27 to be rated superior to model 34 on this data. In another comparison, models 21 (PV_{0,01}FC_{2.0} + ORA/NSC) and 31 (SVM_{0,00001}/NSC) show very similar accuracies (73.4% and 72.9%, respectively) and identical MADs (12.4%), however model 21 achieves a higher gene signature stability (63.9% versus 22.4%) and accomplishes this with shorter signatures (139 genes versus 199 genes), leading to a higher rating of model 21 over model 31. Because the weightings of the different factors are configurable, the model ranking can be fine-tuned to the individual goals of the researcher.

When one considers the top model (model 24, highlighted), one sees that it does not have the highest total accuracy, the lowest MAD, the shortest gene signatures or even the highest signature stability. Nonetheless this model is ranked first because it represents the best overall performance for the given dataset, striking the best

Table 1. Model scoring

No	FSS and Classifier	Acc	MAD	Stab _{NA}	μ	Modscore
PC_{0.6}						
1	NSC	74.6%	13.5%			0.424
2	5NN	69.8%	14.8%			0.400
3	SVM-P	53.6%	13.5%	51.1%	250	0.323
4	SVM-R	72.0%	13.5%			0.414
5	DT	60.9%	13.5%			0.359
PC_[50]						
6	NSC	60.6%	14.8%			0.422
7	5NN	59.7%	14.8%			0.417
8	SVM-P	53.3%	13.5%	37.9%	50	0.389
9	SVM-R	62.5%	13.5%			0.435
10	DT	62.5%	13.5%			0.435
FT₃₀						
11	NSC	65.9%	12.4%			0.503
12	5NN	61.5%	13.5%			0.465
13	SVM-P	58.7%	12.4%	38.9%	99	0.446
14	SVM-R	70.6%	10.8%			0.544
15	DT	61.5%	13.5%			0.466
PV_{0,001}FC_{2.5}						
16	NSC	70.3%	13.5%			0.536
17	5NN	71.2%	13.5%			0.543
18	SVM-P	55.4%	8.1%	66.6%	165	0.428
19	SVM-R	70.2%	13.5%			0.535
20	DT	63.0%	9.4%			0.486
PV_{0,01}FC_{2.0}+ORA						
21	NSC	73.4%	12.4%			0.563
22	5NN	77.6%	14.8%			0.592
23	SVM-P	59.0%	9.9%	63.9%	139	0.453
24	SVM-R	77.2%	12.4%			0.594
25	DT	66.7%	12.4%			0.509
DT_{0,01}						
26	NSC	69.6%	12.4%			0.533
27	5NN	68.4%	12.4%			0.523
28	SVM-P	56.9%	13.5%	53.7%	234	0.428
29	SVM-R	70.4%	13.5%			0.537
30	DT	62.2%	13.5%			0.471
SVM_{0,00001}						
31	NSC	72.9%	12.4%			0.559
32	5NN	67.4%	18.0%			0.503
33	SVM-P	64.0%	15.7%	22.4%	199	0.481
34	SVM-R	68.4%	18.0%			0.512
35	DT	64.7%	12.4%			0.493

Results for 35 models (5 FSS methods combined with 7 classifiers) over 400 sampling steps for the OA dataset. No: Model number, Acc: total classification accuracy, MAD: median absolute deviation of accuracy, Stab_{NA}: stability of FSS method, μ : median length of gene signatures, Modscore: see Equation 4. The table compares the various criteria and the final model score for all the models evaluated on the OA dataset. The best values in each column are highlighted, as well as the best overall model according to our model score. Models are grouped by FSS method. Model 6 is based on the methods used in Michiels *et al.* (2005). Model 24 was selected by our StabPerf approach.

balance between performance and stability. If we consider an arbitrary fixed model, for example model 6, based on the methods used in Michiels *et al.* (2005), one sees that systematically evaluating all combinations of FSS and classification methods with StabPerf was able to uncover a model (i.e. model 24) with both higher total accuracy as well as better signature stability for more genes. The fixed threshold of 50 genes in PC_[50] excludes many

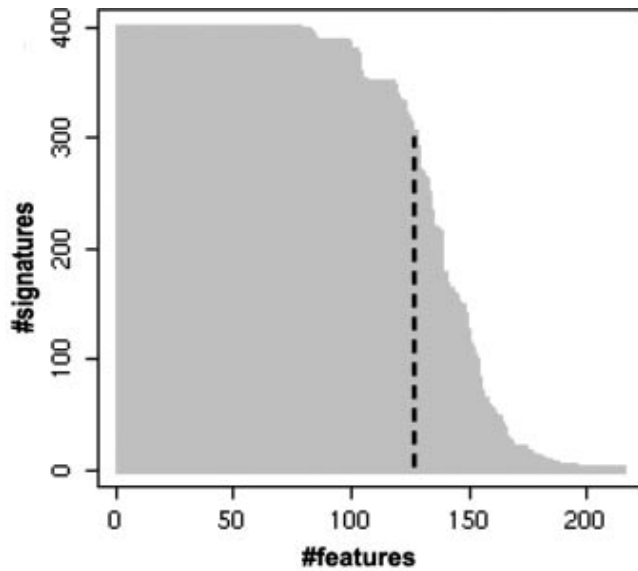


Fig. 3. Stability plot of FSS method $PV_{0.01}FC_{2.0}+ORA$. The plot shows how many features occur in how many signatures out of 400 samples. The stability of the FSS method is the area under the curve (63.9%). In total, 215 spots (96 genes) are selected for a signature by $PV_{0.01}FC_{2.0}+ORA$. 127 spots (33 genes) occur in more than 300 (75%) of the 400 signatures sampled (left of dashed line). Data on total and 75% stable features for the other FSS methods are collected in Table 2.

potentially significant genes, which explains its poor stability (37.9%) compared to, e.g. that of $PC_{0.6}$ (51.1%), which selects genes exceeding a minimum relevance criterion.

Consensus gene signatures

For a given FSS method, some features are more consistently represented across the 400 generated gene signatures than others. In Figure 3 the frequencies of occurrence of all features selected by the chosen FSS method $PV_{0.01}FC_{2.0}+ORA$ are shown (features occurring in none of the 400 signatures not shown). We see that there is a large number of features that are present in every signature and a fair number of additional features that are present in at least $\tau = 75\%$ (300) of the signatures.

We applied this procedure to all seven FSS methods. Table 2 shows to what extent the median signature length can be reduced by building consensus signatures. While our analysis is based on spots on the array (i.e. features), one can also map spots to the genes they represent. The frequency of occurrence of a gene in the 400 signatures is defined here as the average frequency of occurrence of its corresponding spots in the 400 signatures. A gene is retained in the consensus gene signature if it was selected in 75% of these signatures. The threshold τ of 75% may be adjusted, extending or shortening the consensus signatures (column Stable in Table 2), allowing control over the sensitivity and selectivity of the method.

A literature search was conducted to compare the 400 individual pooled signatures (215 spots \doteq 96 genes) to the consensus signature (127 spots \doteq 33 genes) of the top performing combination $PV_{0.01}FC_{2.0}+ORA/SVM-R$. Of the 33 genes retained, 32 (97%) have already been attributed to the context of OA in the literature (Table 3). Among the 96 genes occurring at least once in any of

Table 2. Consensus gene signatures for different FSS methods

FSS	Sum		Med.		Stable		Red.
	Genes	Spots	Genes	Spots	Genes	Spots	
$PC_{0.6}$	174	301	98	241	83	178	26.1%
$PC_{[50]}$	44	94	23	50	11	32	36.0%
FT_{30}	100	197	45	99	25	61	38.4%
PVFC	98	201	65	165	49	124	24.9%
PVFC+ORA	96	215	43	139	33	127	8.6%
$DT_{0.01}$	171	316	93	234	65	155	33.8%
$SVM_{0.00001}$	677	759	147	199	41	50	74.9%

FSS: FSS method, Sum: number of features that occur at least once in any of the 400 signatures produced by the given FSS method, Med.: median signature length, Stable: number of features occurring in over 300 (75%) of the 400 signatures from the given FSS method, Red.: Amount of reduction from Median spots to Stable spots, PVFC: $PV_{0.001}FC_{2.5}$, PVFC + ORA: $PV_{0.01}FC_{2.0}+ORA$. Spots refers to array spots, while Genes identifies how many genes these spots correspond to (see Section Consensus Gene Signatures). With respect to the median length, developing a consensus signature shortened gene signatures by 34.7% on average. The reducibility of a method is inversely related to its stability (Table 1). The 33 genes identified by $PV_{0.01}FC_{2.0}+ORA$ are shown in Table 3.

the 400 signatures, 66 (68.8%) were found in the context of OA. This is still high above average, as once would expect, because the $PV_{0.01}FC_{2.0}+ORA$ method tends to select functionally coherent lists of genes due to the matching between genes and controlled vocabularies. However, it shows that the frequency-based filtering provides an even more stringent list of OA-related genes.

If sampling is neglected in such a scenario, the ratio of relevant genes to selected genes will be lower. For example, with $PV_{0.01}FC_{2.0}+ORA$ 90.3% (SD: 2.7%) of the genes in a signature are found to be OA-related, on average. However, the consensus signature contained 97.0% OA-related genes, which is more than two SDs above the mean. As $PV_{0.01}FC_{2.0}+ORA$ incorporates biological knowledge, the consensus signature produces only a small, but significant, improvement. Less stable FSS methods are expected to benefit even more from these consensus signatures. Additionally, without sampling, it is not clear, how much of the observed classification performance is due to random correlations in the expression data. Moreover, the manual validation of the selected genes is more challenging, as the signatures are longer and contain fewer relevant genes.

The effectiveness of the procedure depends on the selectivity and sensitivity of individual FSS methods. An FSS method, such as $PC_{0.6}$ with a high sensitivity and low selectivity allows more effective determination of consensus signatures. Overly selective FSS methods, e.g. $PC_{[50]}$ on this dataset exclude many potentially stable genes, making it more difficult to produce a meaningful consensus signature. As seen in Table 2, only 11 genes are retained in the consensus gene signature for $PC_{[50]}$, compared to 83 genes from $PC_{0.6}$. This provides fewer stable candidate genes for further examination unless the stability threshold τ is lowered.

Overall classification performance on the OA dataset

The overall classification accuracy of StabPerf was estimated as 73.4% (SD: 11.0%) by 10-fold stratified cross validation for the difficult four-way classification task on the OA dataset. The accuracies can be based on different selected models for different folds. The

Table 3. Stable genes from FSS method $PV_{0.01}FC_{2.0}+ORA$

Frequency	Gene	Description
400	CLECSF1	C-type lectin (cartilage-derived)
400	COL11A1	Collagen, type XI, alpha 1
400	COL11A2	Collagen, type XI, alpha 2
400	COL1A2	Collagen, type I, alpha 2
400	COL2A1	Collagen, type II, alpha 1 (primary osteoarthritis)
400	COL3A1	Collagen, type III, alpha 1
400	COL4A1	Collagen, type IV, alpha 1
400	COL6A1	Collagen, type VI, alpha 1
400	EFEMP1	EGF-containing fibulin-like extracell. matrix prot.
400	LOC83690	CocoaCrisp
400	MT1X	Metallothionein 1X
400	MT2A	Metallothionein 2A
400	OGN	Osteoglycin (osteoinductive factor, variant 3)
400	SOD2	Superoxide dismutase 2, mitochondrial
400	SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)
400	TXNIP	Thioredoxin interacting protein
399	FN1	Fibronectin 1, variant 1
399	IGFBP3	Insulin-like growth factor binding protein 3
399	RPS2	Ribosomal protein S2
398	COL1A1	Collagen, type I, alpha 1
392	COL5A1	Collagen, type V, alpha 1
390	MT1G	Metallothionein 1G
388	COL9A2	Collagen, type IX, alpha 2
387	DUSP1	Dual specificity phosphatase 1
386	GPX3	Glutathione peroxidase 3 (plasma)
379	COL6A3	Collagen, type VI, alpha 3, variant 5
352	PRSS11	Protease, serine, 11 (IGF binding)
351	ANGPTL2	Angiopoietin-like 2
337	MT1E	Metallothionein 1E (functional)
331	OGN	Osteoglycin (osteoinductive factor, variant 1)
321	SFRP4	Secreted frizzled-related protein 4
316	MMP3	Matrix metalloproteinase 3
313	MMP2	Matrix metalloproteinase 2 (type IV collagenase)

Those 33 genes (from 127 array spots), ranked by frequency, occurring in at least 75% of the 400 signatures, that are retained in the consensus gene signature (see Table 2). For all these genes (except LOC83690) we found evidence in a literature search for being associated with OA.

OA disease classes normal and early as well as peripheral OA and central OA are known to be difficult to separate (Fundel *et al.*, 2005). If one simplifies this classification problem, by combining the two classes normal with early as well as peripheral OA with central OA, we observe an average accuracy of 97.5%.

Data on breast cancer

We applied our methods to a second experiment comprising 78 arrays on breast cancer (van't Veer *et al.*, 2002). The aim of this study was to examine if breast cancer patients could be determined that survive for at least five years free of metastases after treatment. This dataset has also been analyzed by Michiels *et al.* (2005) based on the same filter that has been used in the original publication. Here, the same protocol as in model 9 (compare Table 1) has been used differing only in the number of sampling steps (van't Veer *et al.*

use 500). Across all three studies (including our own) a two-class cross-validation accuracy of $\sim 60\%$ was estimated showing that the patient groups are difficult to separate from gene expression profiles alone. Furthermore, a *Modscore* of 0.083 shows that signatures generated from this dataset cannot be expected to yield a good classification performance or signature stability. We tested all 35 models that have been used for the OA data. The criteria for all FSS methods have been relaxed to accommodate for the difficult dataset ($PC_{0.3}$, FT_{20} , $PV_{0.1}FC_{1.5}$, $SVM_{0.0001}$) leading to accuracies of the different models between 50 and 67% (detailed data not shown). Here, the *Modscore* was anticorrelated with performance showing that either moderate accuracy or moderate feature stability could be achieved, but not both.

DISCUSSION

The method StabPerf proposed in this article addresses two objectives in classification tasks of microarray data. The first objective focuses on achieving high classification performance by evaluating FSS methods and classification methods in order to determine the most appropriate combination specifically for the given dataset. The second objective requires that FSS methods and the performance of the classifier are stable, i.e. resilient against variations between sets of expression arrays. We addressed these two objectives by evaluating all possible combinations of the examined FSS and classification methods and by subjecting each of these combinations to repeated random sampling.

Of course, such a strategy can only be meaningfully executed if a reasonably large dataset is being analyzed. In our case, we investigated an OA dataset with four classes and approximately twenty measurements per class. This is one of the largest OA datasets available and is large enough to allow accurate stability and reliability analyses with StabPerf. In addition, the dataset is difficult in that distinguishing the four classes is difficult both at the phenotypic and the gene expression level. In particular, this is true for the comparison between normal and early as well as between peripheral and central OA.

Neglecting the first objective, i.e. relying on a particular classification method, leads to suboptimal classification performance. This can, at the cost of increased computational effort, be avoided by evaluating all pairwise combinations of FSS and classification methods. Indeed, different models may be appropriate for different datasets and the StabPerf approach prevents one from being bound to an inappropriate model for the given dataset. On the contrary, StabPerf chooses the best model for any dataset.

If the second objective of the strategy is compromised, i.e. repeated random sampling is not used, one cannot account for the classification performance that results from learning random correlations. Compared to approaches without sampling, the classification accuracies obtained from our approach may appear inferior in some cases. However, StabPerf provides a more realistic estimation of performance and corrects for overly optimistic and overly pessimistic results. Indeed, it is preferable to estimate the expected performance on future datasets by using a distribution of accuracies, as provided by such a sampling approach.

We also do not follow the approach of standard FSS methods, which simply select genes that exhibit strong differential regulation between sample groups. Such signatures tend to be unstable and contain many genes that may not be relevant to the biological

question at hand. Additionally, the detailed manual validation of standard signatures is more time consuming compared to that of a consensus signature (Table 2). This consensus signature is built from the most stable genes identified by sampling a given FSS method. The length of this consensus signature is determined by the chosen frequency threshold, allowing control over the sensitivity and selectivity of the method. The number of features required for further studies can easily be adjusted at this point by adjusting the threshold. This should avoid shortcomings of arbitrarily limiting the features at the FSS step. For the OA dataset, we were able to show that the consensus signature genes are much more likely to be related to the disease than genes occurring less frequently.

It is important to avoid performance sacrifices on the one hand but also to avoid unrealistic performance estimates on the other. This also requires that sensible parameters need to be manually chosen specifically for the given dataset beforehand to avoid FSS methods returning empty sets of features. We are aware that this might introduce a bias into performance estimation but we expect this bias to be negligible as we do not perform systematic parameter optimization.

We showed that the proposed strategy is well suited to address both types of problems and that the achieved accuracy of 73.4% for the four-way and of 97.5% for the two-class classification problem, as estimated by 10-fold stratified cross validation, is indeed realistic for the given OA dataset. As in many cases biological classes are indeed difficult to separate, the computational cost associated with our approach is justified.

For comparison, we also analyzed a publicly available dataset on breast cancer (van't Veer *et al.*, 2002) that has also been analyzed by Michiels *et al.* (2005). They found that the two classes of patients were difficult to separate (accuracy of 60%) and that feature signatures were unstable in different samplings. In terms of model classification performance our results are consistent with Michiels *et al.* (2005). Signatures extracted from this data are not reliable, which was clearly reflected by our *Modscore*.

In general, we see that processes, such as gene selection that occur earlier in the analysis of microarray data have pronounced effects on classification performance and reliable gene signatures. Therefore, it is imperative that many FSS methods be evaluated, in terms of signature stability, and that these methods be parameterized to emphasize sensitivity over selectivity. As we showed in the case of the Pearson correlation, e.g. simply selecting the top 50 genes prevents one from being able to generate meaningful consensus signatures. Indeed, consensus signatures generated from stable FSS methods have been shown to provide concise and reliable gene lists, as confirmed by checking the disease relevance of the corresponding genes.

CONCLUSION

New methods for feature subset selection and classification of microarray data continue to appear in the literature. For a given dataset, our new model selection approach StabPerf finds the most appropriate combination of FSS and classification method out of a predefined model library. Our sampling approach for all models, as implemented in StabPerf, delivers reliable gene signatures and robust classification performance estimates for a given measurement to be analyzed by computing data similar to Table 1.

The systematic analysis performed by StabPerf involves computational costs, which depend on the number of sampling steps and the number of FSS and classification methods used. However, the

StabPerf procedure, which is available as a free R package, has been designed to exploit parallel processing facilities when in a LAM/MPI environment (Burns *et al.*, 1994). For the data and parameter settings to reproduce Table 1 ~5.5 h on 20 Intel Xeon CPUs were required (without cross validation). Thus, the involved computations need not necessarily delay the analysis process in realistic applications. The advantages with respect to stability and reliability in the analysis of valuable microarray measurements justify the additional computational effort by sampling, systematic model evaluation and cross validation. For a given microarray experiment the StabPerf procedure allows one to avoid typical mistakes in early analysis steps and to obtain realistic classification performance estimates as well as stable, and therefore more reliable, ranked relevant gene lists of customizable length for further processing.

ACKNOWLEDGEMENTS

We would like to thank Dr. Thomas Aigner (Osteoarticular and Arthritis Research, Institute of Pathology, University of Leipzig), Dr. Klaus Lindauer (Bioinformatics, Sanofi-Aventis, Frankfurt), Dr. Joachim Saas and Dr. Eckart Bartnik (Therapeutic Department Thrombosis and Angiogenesis, Sanofi-Aventis, Frankfurt) for helpful comments and discussions. This work has partially been funded by projects BEX (Sanofi-Aventis) and BFAM (bmbf).

Conflict of Interest: none declared.

REFERENCES

- Aigner, T. *et al.* (2006) Large-scale gene expression profiling major pathogenetic pathways of cartilage degeneration in osteoarthritis. *Arthritis and Rheum.* in press.
- Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–9.
- Bi, J., Bennett, K., Embrechts, M., Breneman, C. and Song, M. (2003) Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, **3**:1229–1243.
- Boser, B.E., Guyon, I. and Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh, ACM Press, NY, pp. 144–152.
- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth & Brooks, Monterey.
- Burns, G., Daoud, R. and Vaigl, J. (1994) LAM: An open cluster environment for MPI. In John W. Ross, *Proceedings of Supercomputing Symposium 94*, University of Toronto, pp. 379–386.
- Chang, C. and Lin, C. (2001) LibSVM: a library for support vector machines.
- Draghici, S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
- Dudoit, S. and Yang, J.Y.H. (2003) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In Parmigiani, G., Garrett, E.S., Irizarry, R.A. and Zeger, S.L. (eds), *The Analysis of Gene Expression Data: Methods and Software*, Springer, NY, pp 73–101.
- Dudoit, S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
- Fundel, K., Küffner, R., Aigner, T. and Zimmer, R. (2005) Data Processing Effects on the Interpretation of Microarray Gene Expression Experiments. In Torda, A., Kurtz, S. and Rarey, M. (eds), *German Conference on Bioinformatics (GCB) 2005, Hamburg, Lecture Notes in Informatics*, Gesellschaft für Informatik, Bonn, pp. 77–91.
- Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach. Learning Res.*, **3**, 1157–1182.
- Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*. Springer-Verlag, NY.
- Inza, I. *et al.* (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.*, **31**, 91–103.
- Lottaz, C. and Spang, R. (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics*, **21**, 1971–1978.
- Michiels, S. *et al.* (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Mitchell,T.M. (1997) *Machine Learning*. McGraw-Hill, NY.

Statnikov,A. *et al.* (2005) A comprehensive evaluation of multcategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–72.

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