BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 17 2010, pages 2136–2144
doi:10.1093/bioinformatics/btq345

Gene expression

Advance Access publication June 30, 2010

Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients

Marc Johannes[1,*], Jan C. Brase[1], Holger Fröhlich[2], Stephan Gade[1], Mathias Gehrmann[3], Maria Fälth[1], Holger Sültmann[1] and Tim Beißbarth[4,*]

[1]German Cancer Research Center, Cancer Genome Research, Im Neuenheimer Feld 280, 69120 Heidelberg, [2]Bonn-Aachen International Center for IT, Algorithmic Bioinformatics, Dahlmannstrasse 2, 53113 Bonn, [3]Siemens Healthcare Diagnostics Products GmbH, Diagnostics Research Germany, Nattermannallee 1, 50829 Cologne and [4]University Medicine Göttingen, Medical Statistics, 37099 Göttingen, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: One of the main goals of high-throughput gene-expression studies in cancer research is to identify prognostic gene signatures, which have the potential to predict the clinical outcome. It is common practice to investigate these questions using classification methods. However, standard methods merely rely on gene-expression data and assume the genes to be independent. Including pathway knowledge a priori into the classification process has recently been indicated as a promising way to increase classification accuracy as well as the interpretability and reproducibility of prognostic gene signatures.

Results: We propose a new method called Reweighted Recursive Feature Elimination. It is based on the hypothesis that a gene with a low fold-change should have an increased influence on the classifier if it is connected to differentially expressed genes. We used a modified version of Google's PageRank algorithm to alter the ranking criterion of the SVM-RFE algorithm. Evaluations of our method on an integrated breast cancer dataset comprising 788 samples showed an improvement of the area under the receiver operator characteristic curve as well as in the reproducibility and interpretability of selected genes.

Availability: The R code of the proposed algorithm is given in Supplementary Material.

Contact: m.johannes@DKFZ-heidelberg.de; tim.beissbarth@ams.med.uni-goettingen.de

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

Received on December 19, 2009; revised on May 21, 2010; accepted on June 26, 2010

1 INTRODUCTION

Clinical cancer research is driven by two basic tasks. One is to elucidate the functions of genes and pathways to get mechanistic insights in tumor development and progression and to identify novel drug targets. Another task is the identification of prognostic and predictive gene signatures that allow the prediction of the clinical outcome and response to treatment.

In breast cancer, clinical predictors for relapse events (e.g. lymph-node status and histological grade) often fail to classify tumors according to their clinical behavior. The molecular heterogeneity of the tumors leads to differing outcomes in patients with the same disease stage. Therefore, several studies have been performed to identify prognostic gene signatures based on high-throughput gene-expression profiling. Various gene signatures have been identified, which have the potential to improve the current risk stratification of breast cancer patients (Desmedt *et al.*, 2007; Liu *et al.*, 2007; van't Veer *et al.*, 2002; Wang *et al.*, 2005; Yu *et al.*, 2007).

Classification methods are used in this context to identify genes, which are able to discriminate between clinical relevant endpoints. One example is the support vector machine (SVM), which is a supervised learning algorithm that covers a broad range of applications (Vapnik, 1995; Vapnik and Cortes, 1995). Due to its ability to handle high-dimensional data, the SVM is a good tool for analyzing gene-expression data and its superior performance over other methods has been shown (Brown *et al.*, 2000; Furey *et al.*, 2000).

However, genes identified by standard methods, which exclusively rely on gene-expression data, often lack stability, i.e. applying the same method to different datasets results in gene lists that have only a small number of genes in common (Ein-Dor *et al.*, 2005). Recent studies have shown that standard methods can be improved in terms of classification accuracy as well as stability of gene lists by including a priori knowledge of pathways into the classification process (Bellazzi and Zupan, 2007; Binder and Schumacher, 2009; Chuang *et al.*, 2007; Lee *et al.*, 2008; Rapaport *et al.*, 2007; Su *et al.*, 2009; Yousef *et al.*, 2009; Zhu *et al.*, 2009). These methods benefit from pathway knowledge since genes are not treated as independent. Most of the methods mentioned above are based on the hypothesis that genes in close proximity which are connected to each other should have similar expression profiles. Therefore, these algorithms look for predictive subnetworks instead of single genes. However, this hypothesis might not always be valid since pathways can be completely shut down by only one regulating gene and one would not observe a systematic effect in the whole pathway. Hence, the accuracy of predictions might be further

*To whom correspondence should be addressed.

improved by algorithms that are adapted to use pathway knowledge and are able to choose single genes without demanding that these genes are within the same subnetwork.

In this article, we introduce a new algorithm called Reweighted Recursive Feature Elimination (RRFE). The task of including both, gene-expression data as well as pathway knowledge, is accomplished by using the recursive feature elimination (RFE) algorithm (Guyon *et al.*, 2002), which is based on SVMs in combination with GeneRank (Morrison *et al.*, 2005), a modified version of Google's PageRank algorithm. The PageRank algorithm is based on the idea that a web page should be highly ranked if other highly ranked pages contain links to it. Morrison *et al.* (2005) transferred the idea of PageRank to genes, and we included it into the classification.

RRFE was applied to four single datasets as well as an integrated dataset consisting of 788 breast cancer samples. Two distinct analyses were carried out. The purpose of the first analysis was to highlight differences in gene lists, when network knowledge is used or not. Therefore, the status of the human epidermal growth factor receptor 2 (ERBB2), which is a frequently amplified oncogene in breast cancer, was predicted. In the clinics, the ERBB2 status is routinely analyzed, since the ERBB2 receptor is linked to the prognosis and alternative treatment strategies exist for ERBB2 positive breast cancer patients. In a second analysis, one of the major challenges in current clinical cancer research (Ein-Dor *et al.*, 2006) was addressed, i.e. the prediction of relapse events in breast cancer.

2 SYSTEM AND METHODS

2.1 Support vector machines (SVM)

The SVM is a statistical learning method for building classification models. Briefly, assume as training data, a set of examples $(\mathbf{x}_1, y_1)\ldots(\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \mathbb{R}^P$ represent n input vectors and $y_i \in \{-1, 1\}$ are the corresponding class labels. Any hyperplane can be defined as $\{\mathbf{x} : f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle + b = 0\}$. The SVM searches for the pair $(\widehat{\mathbf{w}}, \hat{b})$ that maximizes the margin between both classes. A new sample with input vector $\mathbf{x}$ is then assigned to class $\text{sign}(\hat{f}(\mathbf{x})) = \text{sign}(\langle \widehat{\mathbf{w}}, \mathbf{x}\rangle + \hat{b})$. The maximum-margin hyperplane is found by the following optimization problem:

$$\underset{\mathbf{w},\xi}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{subject to} \quad y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 - \xi_i, \forall i$$

where ξ_i are slack variables that measure the degree of misclassification. Hence, bounding the sum $\sum_{i=1}^{n}\xi_i$ bounds the amount by which predictions are allowed to fall on the wrong side of the margin. Thus, C is a tuning parameter that controls the trade-off between the slack variable penalty and the size of the margin. The solution to (1) can be found by introducing Lagrangian variables. It can be shown that the weight vector $\mathbf{w}$ can be written as:

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \ , \tag{2}$$

where α_i are the corresponding Lagrange multipliers.

The interested reader is referred to Vapnik (1995), Burges (1998) and Schölkopf and Smola (2001) for a more detailed introduction. All SVMs were trained and tested by using the kernlab package (Karatzoglou *et al.*, 2004).

2.2 Recursive feature elimination (RFE)

In order to permit a feature selection when using SVMs, Guyon *et al.* (2002) introduced an algorithm called RFE.

Feature selection is performed in order to eliminate uninformative variables which, in turn, should lead to a better generalization performance, i.e. the SVM should achieve a better classification performance on previously unseen patterns. In addition to this, a reduced set of features may also give a better insight into the underlying model to be learned and a computational speedup. Other benefits might be cost reduction, in biological applications, for example, where only a smaller subset of genes has to be measured to detect a particular disease with the same accuracy as before.

For SVMs with a linear kernel, as used in this manuscript, RFE uses $\|w\|^2$ as a ranking criterion for the importance of a feature, that is, the feature with the smallest impact on the norm of $\mathbf{w}$ is removed. Thus, in order to get the rank of the j-th feature, the j-th element of the weight vector has to be calculated as

$$|w_j| = \left|\sum_{i=1}^{n}\alpha_i y_i x_{ij}\right| \tag{3}$$

where x_{ij} is the j-th element of the i-th feature vector. After calculating (3) for each of the P features they can be ordered according to their importance (high-value means more important). As a next step a specific amount of features from the bottom of the ordered list are discarded. This process of training the SVM, calculating (3) and removing of a specific amount of uninformative features is repeated until the set of surviving features is empty. For each SVM, trained on a certain number of features, a measure of performance has to be calculated in order to afterwards find the optimal number of features. This can, for example, be done by cross-validation or by using a theoretical concept such as the span estimate (Chapelle and Vapnik, 2000).

2.3 GeneRank

Morrison *et al.* (2005) recently introduced a modified version of Google's PageRank algorithm (Brin and Page, 1998), which they called GeneRank. The authors adapted PageRank to use biological data: based on a biological network and a vector of expression changes, GeneRank calculates a rank for each node in the network. GeneRank increases the rank of a gene with a low fold change if it is connected to genes with a higher change in expression. The algorithm proposed in this article uses a transformed version of the GeneRank result to alter the ranking criterion of RFE and thus incorporates pathway knowledge into the classification.

In order to get a ranking for each node in the network, GeneRank, for $\mathbf{r} \in \mathbb{R}^P$, solves the following linear system:

$$(\mathbf{I} - d\mathbf{G}\mathbf{D}^{-1})\mathbf{r} = (1-d)\mathbf{e}\mathbf{x}. \tag{4}$$

Here, $\mathbf{I}$ is the identity matrix. $d \in (0, 1)$ is a fixed parameter which is called the 'damping factor'. It controls the influence of the network structure and the fold change information on the ranking. A value of $d \to 0$ results in a ranking mostly affected by the fold change information whereas $d \to 1$ corresponds to a ranking that is more dependent on the network structure. $G \in \{0, 1\}^{P\times P}$ is the adjacency matrix of the underlying biological network. This biological network can, for example, be composed of gene–gene interaction or protein–protein interaction (PPI). Thus, a node in the network corresponds to a gene or a protein and the edges encode for an interaction between those. Hence, $g_{ij} = g_{ji} = 1$ if two nodes are connected and $g_{ij} = g_{ji} = 0$ otherwise. D is a diagonal matrix with entries

$$d_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ \deg_i & \text{otherwise} \end{cases}$$

where $\deg_i := \sum_{j=1}^{P} g_{ij}$ is the degree of node i in the network. $\mathbf{e}\mathbf{x} \in \mathbb{R}^P$ is a vector of absolute value of expression change between two biological groups.

After (4) has been solved, $\mathbf{r}$ can be used to rank the features.

2.4 Protein-Protein interaction (PPI) networks

For the evaluations of the algorithm presented in this article (see Section 3.1), information about PPIs was used as prior knowledge. However, throughout

the article the term pathway knowledge is used to refer to any kind of connectivity data, since the new algorithm is not limited to PPI data.

Information about PPI were obtained from the Human Protein Reference Database (HPRD; Prasad *et al.*, 2009). The flatfile of binary PPIs (Release 8, June 9, 2007) was obtained from their servers. After mapping the proteins to the genes present on the microarray used in the experiments (cf. Section 2.5), an interaction matrix G of dimension 7896×7896 was created with

$$g_{ij} = g_{ji} = \begin{cases} 1 & \text{if proteins } i \text{ and } j \text{ interact,} \\ 0 & \text{otherwise.} \end{cases}$$

The mapping resulted in a matrix with 59 924 non-zero elements, that represents 29 962 interactions since the matrix is symmetric.

2.5 Datasets

Four breast cancer Affymetrix HG-U133A microarray datasets, comprising a total of 788 microarrays, were downloaded from the NCBI GEO data repository. All datasets included breast cancer patients that were treated with surgery and did not receive systemic therapy. The datasets were recently analyzed and published by Schmidt *et al.* (2008). The first dataset, the Mainz cohort (Schmidt *et al.*, 2008) represents 153 lymph node-negative, relapse-free patients and 47 lymph node-negative patients that had a relapse (median relapse time 6.04 years; GEO accession number GSE11121). The second dataset, the Rotterdam cohort (Wang *et al.*, 2005) represents 286 lymph node-negative breast cancer samples (including 106 relapse events with median relapse time 7.17 years; GSE2034). Samples from the third and fourth datasets were selected by GSM numbers as previously described by Schmidt *et al.* (2008). Thus, after filtering, the data made available by Loi *et al.* (2007) consisted of 125 samples (GSE6532) including 49 relapse events (median relapse time 7.7 years) and the dataset of Desmedt *et al.* (2007) contained 177 patients (GSE7390), with 85 relapse events (median relapse time 7.42 years).

2.5.1 Data preprocessing The raw data were preprocessed using robust multichip average (RMA; Irizarry *et al.*, 2003). After combining the single datasets, quantile normalization (Bolstad *et al.*, 2003) was performed in order to remove inter-dataset effects. Mapping of the PPIs to the probe sets present on the HG-U133A microarray resulted in 13 671 features with prior knowledge from a total of 22 283 features present on the chip. All annotation data concerning the HG-U133A microarray was obtained from the R-package *hgu133a.db* (Carlson *et al.*, 2009) in the R-Bioconductor environment (Gentleman *et al.*, 2004).

2.5.2 Determination of the ERBB2 *status* *ERBB2* is a frequently amplified oncogene in breast cancer. Determination of ERBB2 status is important, since ERBB2-positive patients have a poor prognosis and targeted treatment strategies (i.e. monoclonal antibody against ERBB2—Trastuzumab) are available for ERBB2-positive breast cancer patients. The ERBB2 status is routinely detected by immunohistochemistry in the clinics. Since ERBB2 status was not available for all samples, ERBB2 receptor status was determined using Affymetrix probe set *216836_s_at* as previously described by Rody *et al.* (2009) and Brase *et al.* (2010). The classification of the ERBB2 status by microarrays has recently been shown to have a concordance of 96% when compared with immunohistochemistry data (Roepman *et al.*, 2009). The expression values of the *ERBB2* probe set showed a bimodal distribution over the 788 samples. By visual inspection, 11.45 was chosen as cutoff to assign the patients into ERBB2 positive and negative. Afterwards *ERBB2* was removed from the raw data which left in 22 281 features for classification.

2.6 Experimental setup and measure of improvement

Classification was performed for each of the four datasets as well as the combined set comprising all 788 samples. All comparisons were evaluated in a 10-fold cross-validation (CV; Hastie *et al.*, 2009; Kohavi, 1995) that was repeated five times with different split positions.

In 10-fold CV, the data is randomly split into 10 equally sized subsamples. In a next step, the classifier is trained on nine subsamples and subsequently used to predict the class-labels of the retained subsample. Thus, after repeating the cross-validation process 10 times, with each of the 10 subsamples used once as test set, one obtains a prediction for every sample in the dataset. Using these predictions, a measure of performance can be computed. In this article, the area under the receiver operator characteristic curve (AUC; Swets, 1988) was used. In order to calculate a standard deviation for the AUC, the CV was repeated five times, yielding five AUCs for each dataset. Due to this setup a total of 50 (5×10) classifiers were trained on each of the five datasets. The AUCs were calculated using the ROCR package (Sing *et al.*, 2005) for the R software (R Development Core Team, 2009).

To test whether or not the new RRFE method gained a significant increase in AUC compared with SVM-RFE the implementation of the Wilcoxon rank sum test in the R software was used. For each dataset, using five AUCs obtained by applying $5\times$ 10-fold CV with RRFE versus five AUCs obtained by applying $5 \times$ 10-fold CV with SVM-RFE (or any other method), the null hypothesis that the distributions differ by a location shift of 0 was tested against the (one-sided) alternative that the distribution of RRFE is greater compared with that of SVM-RFE. A P-value ≤ 0.05 was considered as significantly different.

3 RESULTS

3.1 Reweighted recursive feature elimination (RRFE)

First, a PPI matrix was created using data obtained from the HPRD (cf. Section 2.4). Afterwards, the gene-expression dataset was split up, 90% were used to train the algorithm and the remaining 10% were used after the training procedure to evaluate the performance of the classifier on unseen data (Fig. 1).

As a next step, the training data were used to calculate the fold-change between two biological groups, e.g. relapse and no relapse. The PPI matrix and the fold-change information were used to calculate a ranking of the features by using the GeneRank algorithm (cf. Section 2.3), which was previously implemented for the R software. However, it is important to mention that the PPI matrix contains information about proteins and the microarray consists of probe sets targeting genes. Thus, the probe set data had to be combined in order to get weights for genes. If a gene was targeted by more than one probe set, the fold-change for each probe set was calculated and then averaged. Whenever a gene was represented by more than one protein in the PPI matrix, all proteins derived by that gene were given the same fold-change. After applying GeneRank, the ranks obtained for a particular gene were assigned to the corresponding probe sets. For all experiments performed, the damping factor d was set to 0.5 as suggested in Morrison *et al.* (2005).

The next step was to train an SVM with a linear kernel on all features. The optimal choice for the parameter C in Equation (1) was found by using the *span estimate* (Chapelle and Vapnik, 2000), i.e. for each $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ an SVM was trained and the corresponding *span estimate* for each model was calculated. The *span estimate* is a theoretical concept which allows to calculate a bound on the *leave-one-out* (LOO) error. The LOO error is an approximately unbiased estimate of the generalization performance of the SVM. It has been shown in the literature (Chapelle *et al.*, 2002) that the span estimate provides a good and computationally efficient criterion for model and feature selection in SVMs.

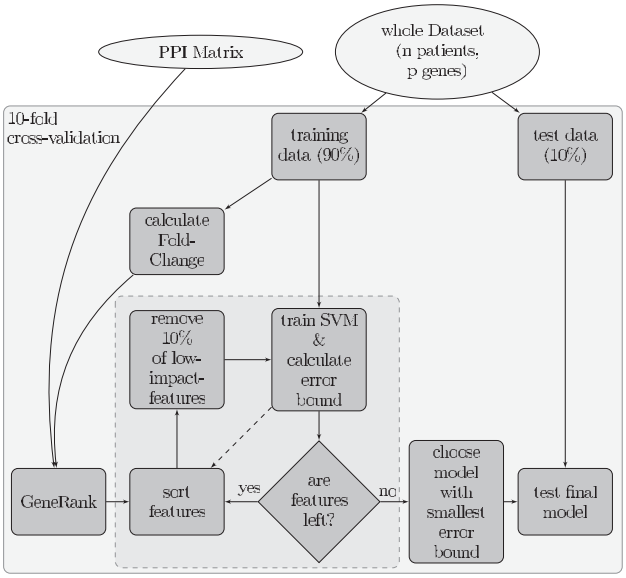

Fig. 1. General workflow of RRFE.

After training, the genes were ordered according to the RFE criterion (dashed line in Fig. 1) in combination with the GeneRank weight for this particular feature. The weights were combined by using a function ϕ, which was defined as:

$$\phi(w_i, r_i) = w_i \frac{1}{\text{rank}(r_i)} \tag{5}$$

where

$$\text{rank}(r_i) = |\{r_l | r_l \geq r_i\}| \tag{6}$$

In Equation (5), w_i is the RFE weight obtained from (3) and r_i is the result of GeneRank for the i-th feature. Instead of using the original ranks returned by the GeneRank algorithm, the results were converted in a rank-based fashion by using the rank function (6). This means that the feature with the highest GeneRank gets a weight of 1, the feature with the second highest rank gets the weight 2, the next 3, …, P. The transformation was done in order to avoid single genes having a weight which is extremely higher than that of all others, a situation that might occur if these genes have a degree (connections to other genes) that is much higher than the degree of the others. Due to the transformation of the GeneRank result and the fact that the RFE weight is multiplied with the reciprocal of it, the feature with the highest GeneRank weight keeps its RFE weight and all subsequent RFE weights are decreased. In contrast to the standard SVM-RFE algorithm, this ranking takes into account both the impact of a particular feature on the weight vector of the hyperplane and the connectivity of that feature in the underlying biological network. After the list has been ordered, 10% of the genes with the smallest impact were discarded. As in the original SVM-RFE algorithm, this procedure of eliminating genes was repeated until the set of surviving features was empty (dashed box in Fig. 1).

After finishing the feature elimination process, the SVM with the optimal number of features had to be selected. Since each SVM was tuned by using the *span estimate*, a bound on the LOO error was inherently available for each model. This bound was used to select the SVM with the best generalization performance. Subsequently, the selected SVM was used to predict the class labels of the remaining 10% of the samples, which were previously assigned to the test set.

Since 10% of the data were used for testing, the process of splitting the data in test and training set had to be repeated 10 times until each sample was exactly once a member in the test set (light gray solid box in Fig. 1). Thus, after 10 folds the class label of each sample was predicted. These predictions were then used to calculate the overall AUC of the method. However, to provide a better estimate of the variation of the AUC, the whole CV process (Fig. 1) was repeated five times.

3.2 RRFE performs stable feature selection and improves the interpretability of gene lists

Overexpression of the *ERBB2* oncogene is a well-known factor associated with poor prognosis in breast cancer, which is frequently based on the amplification of a genomic region on chromosome 17. *ERBB2* amplification on the gene level frequently leads to an overexpression on the RNA and protein level. Several genes adjacent to *ERBB2* are co-amplified leading to an additional upregulation on the RNA level.

To demonstrate the impact of RRFE on the genes, which are chosen for the classification, 788 breast cancer samples were assigned to two groups according to their ERBB2 receptor status (cf. Section 2.5.2). Of the total breast cancer patients, 102 were defined as ERBB2 receptor positive and 686 as ERBB2 negative. The *ERBB2*-specific probe sets were omitted, and all other genes that could be mapped onto a pathway were used to predict the *ERBB2* status of the patients. Since some of the genes are lying within the same amplicon, adjacent to the *ERBB2* gene, they are highly correlated with ERBB2 status (due to co-amplification). Hence, by choosing these genes, a classifier should be able to predict the status of the patients with high accuracy. Therefore, the focus of this experiment was not to compare the performance of the methods but rather to highlight changes in the genes which have been used by the methods to predict the ERBB2 status of the patients. Due to the way RRFE incorporates prior knowledge into the classification process one would expect differences in the selected gene sets.

RRFE was used to predict the ERBB2 status and subsequently compared with standard SVM-RFE. As expected, after cross-validation, an AUC close to 1 could be observed with both methods (data not shown). However, different genes were selected when the standard method was compared with the RRFE method (Table 1). More than 50% (6 out of 10) of the genes that were selected most often by the standard SVM-RFE are located on chromosome 17, adjacent to the *ERBB2* gene (Table 1, upper panel). Those genes are identified due to their correlation with the class labels, that is, the receptor status. However, due to the correlation the information gained from these features is redundant.

Three of these correlated genes found by RFE are contained in the gene list of RRFE as well, which is most probably due to their high fold-change (Table 1, lower panel). However, in addition several other genes were chosen as a result of including prior knowledge, i.e. the connectivity.

To functionally characterize genes that have been chosen by both methods as well as to test for enriched pathways within these gene lists the tool *gene2pathway* (Fröhlich *et al.*, 2008) has been used.

Table 1. Results of the *ERBB2* receptor status prediction

Gene symbol	Chromosome	Times chosen	Log fold-change	Connections
SVM-RFE				
GRB7	**17**	50	−2.105	14
NDUFA7	19	31	0.308	2
MED24	**17**	29	−1.479	4
LRRC59	**17**	24	−0.729	1
CRKRS	**17**	23	−1.825	5
PHB	**17**	22	−0.512	14
CD86	3	22	−0.099	4
MED1	**17**	21	−1.627	26
ACTG1	17	21	0.024	35
NR2F1	5	20	−0.538	13
RRFE				
GRB7	**17**	50	−2.105	14
EGFR	7	49	0.032	151
WFDC2	20	48	1.044	1
EWSR1	22	48	−0.008	97
TP53	17	48	0.240	242
PRKACA	19	48	0.022	131
LRRC59	**17**	48	−0.730	1
SMAD3	15	47	0.072	166
CRKRS	**17**	47	−1.825	5
PRKCA	17	47	0.038	162

Top 10 genes chosen by both methods after CV, i.e. a gene would have been chosen 50 times if it was considered as important by all classifiers. (upper panel) genes chosen by the SVM-RFE algorithm. (lower panel) genes considered as important by RRFE. Bold numbers indicate genes which lie adjacent to the ERBB2 gene on chromosome 17.

This tool retrieves the pathway membership of a gene from the KEGG database (Kanehisa and Goto, 2000) and afterwards uses Fisher's exact test to identify significantly enriched pathways within the genes which were identified by the classifiers. All *P*-values were corrected for multiple testing with the method by Benjamini and Yekutieli (2001). After applying *gene2pathway* to the top 100 genes which have been chosen most often by SVM-RFE no significantly overrepresented pathways could be found (data not shown). However, using the top 100 selected genes chosen by RRFE the pathway *ERBB signaling* ($P = 1.69 \times 10^{-11}$) showed up to be significantly enriched among others.

Besides changes in the gene lists, RRFE chooses genes more consistently compared with SVM-RFE. The classification performance was evaluated in a five times repeated 10-fold CV, therefore 50 classifiers were trained (Section 2.6). Thus, if a gene was considered as important by every classifier, it would have been chosen 50 times. SVM-RFE chooses *GRB7* 50 times (Table 1), but the second gene in the list, *NDUFA7*, was only chosen 31 times. The 10th gene, *NR2F1*, was chosen by <50% of the SVM-RFE classifiers. RRFE chooses *GRB7* 50 times as well. However, the next gene is still chosen by 49 out of 50 classifiers. *PRKCA*, the 10th gene in the RRFE list, survived the elimination process in 94% of the classifiers.

3.3 RRFE improves the prediction of relapse events in breast cancer

RRFE was applied to predict the event of a relapse in four single datasets as well as in a combined dataset comprising 788 breast cancer samples. For each dataset, two analyses were performed.

Table 2. AUC obtained by predicting relapse events on five datasets

	SVM-RFE	RRFE	*P*-value	SVM-RFE	RRFE	*P*-value
Combined	0.649	0.671	0.028	0.657	0.688	0.008
GSE11121	0.614	0.667	0.016	0.588	0.657	0.016
GSE2034	0.659	0.708	0.008	0.673	0.727	0.028
GSE7390	0.528	0.516	0.345	0.519	0.536	0.016
GSE6532	0.503	0.609	0.004	0.518	0.585	0.004

Columns 1 and 2 show the difference in AUC when prior knowledge is used (RRFE) or not (SVM-RFE). Column 3 shows the *P*-values obtained from testing whether there is a significant difference between the AUCs. Columns 4 and 5 show the AUC for both methods when all genes were used for classification. The last column shows the *P*-values obtained by carrying out the same test as above.

First, only genes for which prior knowledge was available, i.e. which were present in the PPI matrix, were used. Although this is the most straightforward way to include prior knowledge it might not be the best, because many potentially informative features have to be discarded. Therefore, a second analysis was performed with the aim to use as much information as possible, that is, all features present on the microarray. Since it is not possible to obtain a GeneRank weight for genes that are not present in the PPI matrix, the smallest weight returned by GeneRank was assigned to these genes. Due to this setup, SVM-RFE was compared with RRFE 10 times (five datasets, two analysis on each).

In 9 out of these 10 comparisons, RRFE was able to significantly increase the AUC (Table 2 and Fig. 2). No improvement could be observed in the dataset of Loi *et al.* (2007). However, the performance was similar to that of the standard SVM-RFE algorithm.

Pathway overrepresentation analysis (Fröhlich *et al.*, 2008) was carried out on the top 100 genes selected by each method on the combined dataset (788 samples; only genes that could be mapped onto a pathway). Within the top 100 selected genes of SVM-RFE no significantly enriched pathways could be found. Analysis of the top 100 genes selected by RRFE revealed *Cell Growth and Death* ($P = 2.656 \times 10^{-12}$), *Cancers* ($P = 1.880 \times 10^{-11}$) and *Cell cycle* ($P = 6.607 \times 10^{-08}$) as significantly overrepresented among others.

The stability of the gene lists was investigated as well. Of the top 100 genes selected by RRFE, 87 were chosen by each of the 50 classifiers. The remaining 13 genes were selected 49 times. With SVM-RFE no gene was selected by all classifiers, the top gene was chosen 46 times. SVM-RFE seems not to perform a stable feature selection since the 100th feature was considered as important by only 36% of the classifiers.

3.4 Comparison to other classification methods

For comparing the performance of RRFE to other state-of-the-art methods, the combined dataset with 788 samples was used. The methods were used to predict if the patients had a relapse event or not. RRFE was compared with both standard methods that use solely expression data as well as methods that use expression data in combination with pathway knowledge.

3.4.1 Standard classification methods RRFE was first compared to the nearest shrunken centroid method (Tibshirani *et al.*, 2002). After five CVs this method reached an average AUC of 0.590. In a second analysis, generalized additive models were fitted

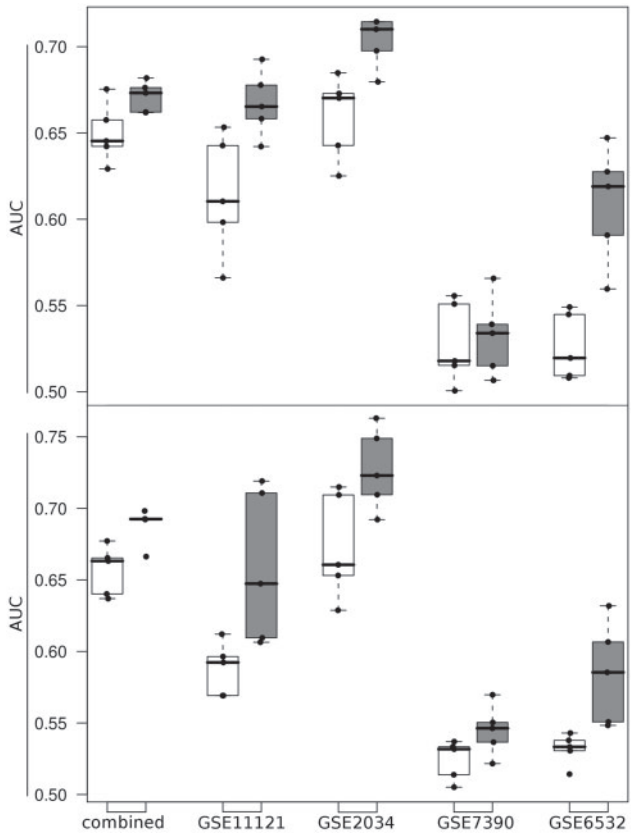

Fig. 2. Boxplots of AUCs obtained by CV. The *x*-axis denotes the five datasets (combined = integrated dataset; GSE = GEO accession number for the single datasets). The upper part shows results of classification on 13 671 features for which pathway knowledge was available. In the lower part, all 22 283 features were used. The white boxplots show the result of SVM-RFE and the gray ones show the result of RRFE.

by likelihood-based boosting (Tutz and Binder, 2006) using the GAMBoost package of Binder and Schumacher (2009). The average AUC of GAMBoost was 0.559.

3.4.2 Classification methods that use prior knowledge In order to have fair comparisons all methods mentioned in this subsection were adopted to use PPI data as basis of pathway knowledge.

First, RRFE was compared with the network-based SVM introduced by Zhu *et al.* (2009). They proposed a network-based penalty which is based on the sum of the F_∞-norms (Zou and Yuan, 2008) applied to the neighboring gene pairs. With the hinge loss of the SVM punished by their network-based penalty, the pathway knowledge is directly included into the optimization process of the SVM. The performance was evaluated by CV and after five repeats an average AUC of 0.607 was observed (Table 3). However, it is worth noting that we followed the analysis of Zhu *et al.* (2009) and focused only on genes with the highest variation (i.e. standard deviation of the $\log_2$ expression level ≥ 0.8 across 788 samples).

In order to evaluate the performance of algorithms distinct from SVMs, we have chosen a boosting algorithm for the second comparison. Binder and Schumacher (2009) recently proposed a method called PathBoost which is based on likelihood-based boosting (Tutz and Binder, 2006). PathBoost is similar to GAMBoost mentioned above. However, PathBoost uses the additional knowledge about pathways to adapt the penalty of connected genes in the course of boosting steps. The main parameter of the method is the number of boosting steps which was optimized by an inner CV. After five CV runs, the average AUC of PathBoost was 0.564 (Table 3).

Table 3. Method comparison on the combined dataset

Method	AUC (95% CI)
RRFE	0.671 (0.662–0.681)
Network-based SVM	0.607 (0.574–0.645)
PathBoost	0.564 (0.547–0.601)
PAM	0.590 (0.569–0.623)
GAMBoost	0.559 (0.534–0.588)

Shown are the average AUC and the 95% confidence interval (CI) after CV.

4 DISCUSSION

Many high-throughput gene-expression profiling studies have been performed with the aim to identify putative new biomarkers for risk stratification of breast cancer patients (Desmedt *et al.*, 2007; Liu *et al.*, 2007; van't Veer *et al.*, 2002; Wang *et al.*, 2005; Yu *et al.*, 2007). However, a major drawback of the gene lists identified in these studies is that the overlap is poor (Ein-Dor *et al.*, 2005; Sontrop *et al.*, 2009).

Classification methods are often used to computationally find survival-related gene signatures. However, genes which are known to be involved in breast cancer are not necessarily identified by current methods because they only detect genes that exhibit a large change in expression (Chuang *et al.*, 2007), which might be only downstream effectors of the key players. For example, minute changes in transcription factor expression may lead to considerable expression changes of many genes regulated by the transcription factor. The key players, which are correlated to cancer development and not exhibit high fold changes, are known to be highly connected to those genes that significantly change their expression level (Chuang *et al.*, 2007). Hence, the performance of classification algorithms might be improved if they are adapted to use the gene-expression data coming from microarray experiments and exploit the information that is stored in biological networks, that is, the connectivity of the genes.

In this article, we introduced RRFE, which includes pathway knowledge into the classification process by modifying the feature-ranking criterion of RFE (Guyon *et al.*, 2002). Motivated by the findings mentioned above, we used GeneRank (Morrison *et al.*, 2005), a modified version of the PageRank algorithm (Brin and Page, 1998), to alter the ranking criterion of RFE and thus incorporate the prior knowledge. The PageRank algorithm, used by Google, is based on the hypothesis that a web page should be highly ranked in a search result if other highly ranked pages contain links to it. This hypothesis is similar to what is observed in gene networks. The assumption of our approach is that a gene with a low fold change should have an increased influence on the classifier if it is connected to differentially expressed genes. The combination of GeneRank and RFE gives

highly connected genes the chance to influence the classifier and in turn help deciphering the underlying biological process. Thus, RRFE accounts for the fact that many functionally relevant genes might not be detectable with current techniques and hence decrease the amount of unexploited information in the data.

The graph structure which is needed by GeneRank was built from binary PPI data coming from the HPRD. This database is known to contain manually curated information for a high number of proteins. Thus, there was a high overlap between proteins in the database and corresponding genes on the microarrays used in our experiments. Therefore, by using HPRD, we were able to keep a high number of features for the classification. However, it is worth mentioning that RRFE showed a similar performance when it was evaluated using network data coming from KEGG (Kanehisa and Goto, 2000) and the ConsensusPathDB (Kamburov *et al.*, 2009) (data not shown). Nevertheless, RRFE can easily be adapted to use other sources of prior knowledge.

In order to judge the performance of RRFE on a dataset with a large number of cases, we integrated four single datasets into one big dataset as recently described by Schmidt *et al.* (2008). However, combining datasets is not a trivial task and one has to take care of several issues: only probe sets which are present on every microarray should be considered, the patients should have received a similar or no treatment and the chips should be normalized together in order to get rid of within-array differences. After careful consideration, we thus decided to use the datasets described by Schmidt *et al.* (2008) for all experiments performed in this article. The reason for choosing these particular datasets was that all samples were measured on the same microarray platform and the patients did not receive any systemic treatment. However, to be sure that our method does not benefit from the combination of the four datasets, we also evaluated it on the single ones.

In the first analysis, the influence of pathway information on the composition of gene lists was assessed. Therefore, classification of the ERBB2 status in 788 samples was chosen. Breast cancer patients are classified due to their ERBB2 receptor status in the clinics. ERBB2-positive patients can be treated with a specific therapeutical antibody directed against the ERBB2 receptor (Emens, 2005). *ERBB2* classification is a good example to demonstrate the gene selection, since ERBB2 receptor overexpression is based on an amplification of a region on chromosome 17. Therefore, adjacent genes within the *ERBB2* amplicon are simultaneously overexpressed and detected by common classification methods due to their high expression fold change. This is a well-known example for a gene cluster showing a high correlation in a certain molecular breast cancer subtype (ERBB2 positive). However, a large proportion of the upregulated genes (based on co-amplification) are apparently not associated with the intrinsic biology and the adverse clinical outcome of the ERBB2 breast cancer subtype. Therefore, the idea behind this analysis was to show how the gene lists differ when prior knowledge (PPI data) is used or not.

Our analyses showed that the standard SVM-RFE selects many of the overexpressed genes, which are located on chromosome 17 adjacent to *ERBB2*. However, ERBB2-positive breast cancer is a distinct subtype, which is correlated with adverse risk factors (Slamon *et al.*, 1987; Sorlie *et al.*, 2001). Therefore, classification methods should also detect genes, which are correlated with the intrinsic biology rather than being co-amplified. The analyses demonstrated that different gene sets were selected by the RRFE classification method and the pathway overrepresentation showed that RRFE is indeed able to detect those genes that are correlated with the intrinsic biology (see below).

Another result that could be observed in this analysis is that the gene list of RRFE is more stable compared with SVM-RFE. In addition, the genes chosen by RRFE are more closely related to cancer: a pathway overrepresentation analysis revealed *ErbB signaling* among some others as significantly enriched in the top 100 genes chosen by RRFE. No overrepresented pathways could be observed with the standard algorithm.

Besides consistent feature selection, a classifier should be able to classify patients with a high sensitivity as well as with a high specificity. To investigate the performance of RRFE, it was used to predict if a breast cancer patient will suffer from a relapse event after surgery. The classification was performed in two different settings. In the first instance, classification was performed on all genes which could be mapped onto a pathway. The second test was done using all measured genes. Both analyses were performed on all datasets, which resulted in 10 comparisons. When the results of RRFE were compared with SVM-RFE, a significant increase in AUC could be observed in 9 out of 10 cases. Again, a pathway overrepresentation analysis was performed for the top 100 genes selected most often by each of the methods. The gene list of RRFE revealed several cancer-related pathways as significantly enriched. No enriched pathways could be found in the gene list produced by SVM-RFE. Thus, the pathway knowledge improves classification accuracy and enables the classifier to identify cancer relevant genes. In addition, the pathway data causes the classifier to be more robust against noise in the data, since it significantly increases the AUC even on the combined dataset with 788 samples where the experiments have been performed by different investigators in different labs on different days.

To further investigate the performance of RRFE, it was compared with both standard classification methods and classification methods that take a priori into account the knowledge of pathways. In order to obtain comparable results, the methods that use pathway knowledge were adopted to use the same knowledge-base as RRFE, i.e. the HPRD. In all comparisons RRFE reached the highest AUC.

We think, that it is important to mention that biological knowledge is biased and under permanent change, which is especially true for pathway data. Therefore, when using pathway knowledge one has to bear in mind that this information suffers from an annotation bias, and thus the density of connections in certain regions of the network might be higher due to the fact that these parts of the network are better understood. Therefore, it is desirable to be able to control the influence of the pathway knowledge on the ranking of the genes, which can be done by adjusting the damping factor. The value of the damping factor should depend on the reliability of the underlying pathway data.

We believe that with increasing quality and amount of pathway data, the classification performance of methods should increase as well.

5 CONCLUSION

We presented a new classification method called RRFE that a priori takes pathway knowledge into account. This is achieved by using a modified version of the PageRank algorithm in combination with RFE. In our experiments, we used PPI data as prior knowledge.

However, the algorithm can also be used with other types of pathway data. We showed that gene lists become more consistent when pathway knowledge is used. In addition, compared with SVM-RFE the AUC could be significantly improved in 9 out of 10 experiments. RRFE also outperformed two other classification methods that include a priori the knowledge of pathways as well as two standard classification methods that are commonly used in the field of microarray analysis.

In conclusion, RRFE increases both, the classification performance in clinically relevant questions as well as the amount of information that is extracted from the data.

ACKNOWLEDGEMENTS

We thank Christian Bender and Ruprecht Kuner for help and discussions and Dirk Ledwinka for IT support. The authors are responsible for the contents of this publication.

Funding: German Federal Ministry of Education and Science (BMBF) in the framework of the program for medical genome research (NGFN, IG-Prostate Cancer, FKZ: 01GS0890); BMBF project BreastSys in the platform Medical Systems Biology; the Clinical Research Group 179 through the German Research Foundation (DFG).

Conflict of Interest: none declared.

REFERENCES

Bellazzi,R. and Zupan,B. (2007) Towards knowledge-based gene expression data mining. *J. Biomed. Inform*., **40**, 787–802.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Binder,H. and Schumacher,M. (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, **10**, 18.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Brase,J.C. *et al.* (2010) ERBB2 and TOP2A in breast cancer: a comprehensive analysis of gene amplification, RNA levels, and protein expression and their influence on prognosis and prediction. *Clin. Cancer Res.*, **8**, 36.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, **30**, 107–117.

Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Burges,C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Carlson,M. *et al.* (2009) *hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)*. R package version 2.2.12. Available at http://bioconductor.org/packages/2.4/data/annotation/html/hgu133a.db.html (last accessed date July 9, 2010).

Chapelle,O. and Vapnik,V. (2000) Bounds on error expectation for support vector machines. *Neural Comput.*, **12**, 2013–2036.

Chapelle,O. *et al.* (2002) Choosing multiple parameters for support vector machines. *Mach. Learn.*, **46**, 131–159.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 10.

Desmedt,C. *et al.;* TRANSBIG Consortium (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

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. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Emens,L.A. (2005) Trastuzumab: targeted therapy for the management of HER-2/neu-overexpressing metastatic breast cancer. *Am. J. Ther.*, **12**, 243–253.

Fröhlich,H. *et al.* (2008) Predicting pathway membership via domain signatures. *Bioinformatics*, **24**, 2137–2142.

Furey,T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Gentleman,R.C. *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

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

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction*. Springer, New York.

Irizarry,R. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kamburov,A. *et al.* (2009) Consensuspathdb–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623–D628.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Karatzoglou,A. *et al.* (2004) kernlab – an S4 package for kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 1137–1143.

Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.

Liu,R. *et al.* (2007) The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N. Engl. J. Med.*, **356**, 217–226.

Loi,S. *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.*, **25**, 1239–1246.

Morrison,J.L. *et al.* (2005) Generank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.

Prasad,T.S.K. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res*., **37**, D767–D772.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Rody,A. *et al.* (2009) Gene expression of topoisomerase II alpha (TOP2A) by microarray analysis is highly prognostic in estrogen receptor (ER) positive breast cancer. *Breast Cancer Res. Treat.*, **113**, 457–466.

Roepman,P. *et al.* (2009) Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer. *Clin. Cancer Res.*, **15**, 7003–7011.

Schmidt,M. *et al.* (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res*., **68**, 5405–5413.

Schölkopf,B. and Smola,A.J. (2001) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Slamon,D.J. *et al.* (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, **235**, 177–182.

Sontrop,H. *et al.* (2009) A comprehensive sensitivity analysis of microarray breast cancer classification under feature variability. *BMC Bioinformatics*, **10**, 389.

Sorlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Su,J. *et al.* (2009) Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*, **4**, e8161.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

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

Tutz,G. and Binder,H. (2006) Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**, 961–971.

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

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik,V. and Cortes,C. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

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

Yousef,M. *et al.* (2009) Classification and biomarker identification using gene network modules and support vector machines. *BMC Bioinformatics*, **10**, 337.

Yu,J.X. *et al.* (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, **7**, 182.

Zhu,Y. *et al.* (2009) Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, **10** (Suppl. 1), S21.

Zou,H. and Yuan,M. (2008) The F_∞-norm support vector machine. *Stat. Sin.*, **18**, 379–398.