BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 1 2007, pages 30–37
doi:10.1093/bioinformatics/btl543

Gene expression

Improved breast cancer prognosis through the combination of clinical and genetic markers

Yijun Sun[1,3,*], Steve Goodison[2], Jian Li[3], Li Liu[1] and William Farmerie[1]

[1]Interdisciplinary Center for Biotechnology Research, [2]Department of Surgery and [3]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA

Received on July 3, 2006; revised on October 11, 2006; accepted on October 15, 2006
Advance Access publication November 26, 2006
Associate Editor: David Rocke

ABSTRACT

Motivation: Accurate prognosis of breast cancer can spare a significant number of breast cancer patients from receiving unnecessary adjuvant systemic treatment and its related expensive medical costs. Recent studies have demonstrated the potential value of gene expression signatures in assessing the risk of post-surgical disease recurrence. However, these studies all attempt to develop genetic marker-based prognostic systems to replace the existing clinical criteria, while ignoring the rich information contained in established clinical markers. Given the complexity of breast cancer prognosis, a more practical strategy would be to utilize both clinical and genetic marker information that may be complementary.

Methods: A computational study is performed on publicly available microarray data, which has spawned a 70-gene prognostic signature. The recently proposed I-RELIEF algorithm is used to identify a hybrid signature through the combination of both genetic and clinical markers. A rigorous experimental protocol is used to estimate the prognostic performance of the hybrid signature and other prognostic approaches. Survival data analyses is performed to compare different prognostic approaches.

Results: The hybrid signature performs significantly better than other methods, including the 70-gene signature, clinical makers alone and the St. Gallen consensus criterion. At the 90% sensitivity level, the hybrid signature achieves 67% specificity, as compared to 47% for the 70-gene signature and 48% for the clinical makers. The odds ratio of the hybrid signature for developing distant metastases within five years between the patients with a good prognosis signature and the patients with a bad prognosis is 21.0 (95% CI: 6.5–68.3), far higher than either genetic or clinical markers alone.

Availability: The breast cancer dataset is available at www.nature.com and Matlab codes are available upon request.

Contact: sun@dsp.ufl.edu

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

1 INTRODUCTION

Breast cancer is the second most common cause of deaths from cancer among women in the United States. In 2006, it is estimated that about 212 000 new cases of invasive breast cancer will be diagnosed, along with 58 000 new cases of non-invasive breast cancer and 40 000 women are expected to die from this disease (Data from American Cancer Society, 2006). The major clinical problem of breast cancer is the recurrence of therapeutically resistant disseminated disease. In many patients, microscopic or clinically evident metastases have already occurred by the time the primary tumor is diagnosed. Chemotherapy or hormonal therapy reduces the risk of distant metastases by one-third. However, it is estimated that about 70% patients receiving treatment would have survived without it. Therefore, being able to predict disease outcomes more accurately would help physicians make informed decisions regarding the potential necessity of adjuvant treatment, and may lead to the development of individually tailored treatments to maximize the efficacy of treatment. Consequently, this would ultimately contribute to a decrease in overall breast cancer mortality, a reduction in overall heath care cost and an improvement in patients' quality of life.

Despite significant advances in the treatment of primary cancer, the ability to predict the metastatic behavior of tumors remains one of the greatest clinical challenges in oncology. Two commonly used treatment guidelines are the St. Gallen (Goldhirsch *et al*., 2003) and NIH (Eifel *et al*., 2000) consensus criteria that determine whether a patient is at a high risk of tumor recurrence and/or distant metastases based on a panel of clinical markers, such as age of patient, tumor size, the number of involved lymph nodes at the time of surgery and the aggressiveness of the cancer based on histopathological parameters. These criteria are less than precise in predicting therapy failure, with only 10% specificity at the 90% sensitivity level[1]. A more accurate prognostic criterion is urgently needed to avoid over- or under-treatment in newly diagnosed patients.

It has been recently established that related cellular phenotypes are generally reflected in the related patterns of cellular transcripts, implying the possibility of classifying cellular states by monitoring gene expression profiles (Golub *et al*., 1999). Identifying a gene signature using microarray data for breast cancer prognosis has been a central goal in some recent large-scale exploratory studies. In van't Veer *et al*., 2002, a 70-gene signature (also known as the Amsterdam signature) was derived from a cohort of 78 breast cancer patients, the prognostic value of which was further validated in

*To whom correspondence should be addressed.

[1]The specificity is defined as the rate of correctly predicting the lack of need of the adjuvant systemic therapies when the therapies are indeed not necessary, and the sensitivity is the rate of administering the adjuvant systemic therapies when indeed these therapies are effective.

a larger dataset (van De Vijver *et al*., 2002). More recently, a 76-gene signature was identified and successfully used to predict distant metastases of lymph node-negative primary breast cancer (Wang *et al*., 2005). These studies have shown that gene profiling can achieve a much higher specificity than the current clinical systems (50% versus 10%) at the same sensitivity level. These results are considered groundbreaking in breast cancer prognosis. A prospective and randomized study involving ∼800 breast cancer patients, referred to as MINDACT (Microarray In Node negative Disease may Avoid ChemoTherapy), is currently being conducted in Europe in order to evaluate the prognostic value of the 70-gene signature (Loi *et al*., 2006).

The predictive values of these gene signatures are usually demonstrated through comparison with the conventional St. Gallen and NIH consensus criteria. Though the results favor gene signatures, the comparison is somewhat unfair since both St. Gallen and NIH consensus criteria perform risk assessment by following the rules derived heuristically from clinical experiences rather than carefully optimized rules[2]. Edén *et al*. showed experimentally that the clinical markers, when used as the features in a well trained neural network (NNW), performed similarly to a gene based prognostic system (Edén *et al*., 2004), which is in sharp contrast with the conclusions drawn in the existing studies (van't Veer *et al*., 2002; van De Vijver *et al*., 2002; Weigelt *et al*., 2005). Moreover, most of the existing studies attempt to use a genetic marker based prognostic system to replace the existing clinical rules, rather than incorporating the valuable clinical information. Given the complexity of breast cancer prognosis, a more practical strategy, as suggested by Brenton *et al*., 2005, is to utilize both clinical and genetic markers that may contain complementary information. This may lead to a more economical and accurate prognostic system. In this paper, we conduct a computational study to demonstrate the feasibility of this strategy.

The key challenge to deriving a hybrid prognostic signature from both genetic and clinical markers is feature selection. One characteristic of microarray data, different from most of the classification problems we encounter, is the extremely large feature dimensionality compared to the small sample size. The curse of dimensionality (Duda *et al*., 2000; Trunk, 1979) becomes a serious problem. Here, we use our recently developed I-RELIEF algorithm to select a small feature subset such that the performance of a learning algorithm is optimized. I-RELIEF employs a feature weighting strategy that assigns each feature a real-valued number, instead of a binary one, to indicate its relevance to a learning problem. The feature weighting strategy enables the employment of well established optimization techniques, and thus allows for efficient algorithmic implementation that is critical for microarray data analysis. We use a rigorous experimental protocol to estimate the classification parameters and the prognostic performance of the new hybrid signature and other prognostic approaches, including the 70-gene signature, the clinical markers alone, and the conventional St. Gallen criterion. Survival data analyses are performed to compare the different prognostic approaches. Our results clearly demonstrate the superiority of the hybrid signature over a prognostic system that uses only genetic or clinical markers.

[2]The St. Gallen consensus criterion: tumor ≥2 cm, estrogen receptor negative grade 2–3, patient <35 years old (any one of these criteria met suggests high distant metastases risk.); the NIH consensus is similar to that of St. Gallen, but with tumor >1 cm.

2 MATERIALS AND METHODS

2.1 Dataset

A computational study is performed on van't Veer's data (van't Veer *et al*., 2002). This dataset contains expression profile information derived from samples collected from 97 lymph node-negative breast cancer patients 55 years old or younger, and associated clinical information including age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER) and progesterone receptor (PR) status. Among the 97 patients, 46 developed distant metastases within 5 years and 51 remained metastases free for at least 5 years. The isolation of RNA from cancerous tissues, labeling of complementary RNA (cRNA), the competing hybridization of labeled cRNA with a reference pool of cRNA from all tumors to arrays containing 24 481 gene probes, quantization and normalization of fluorescence intensities of scanned images are detailed described in the previous publication (van't Veer *et al*., 2002). The task is to build a computational model to accurately predict the risk of distant recurrence of breast cancer (using a 5-year post-surgery period as the defining point commonly used in the literature). Except for a simple re-scaling of each feature value to be between 0 and 1, no other preprocessing is performed. The re-scaling is performed by using the formula:

$$\hat{\mathbf{x}}_n^{(i)} = \frac{\mathbf{x}_n^{(i)} - \min_m \mathbf{x}_m^{(i)}}{\max_m \mathbf{x}_m^{(i)} - \min_m \mathbf{x}_m^{(i)}},$$

where $\mathbf{x}_n^{(i)}$ is the ith feature in the nth sample. In the following, we drop the hat in $\hat{\mathbf{x}}_n^{(i)}$ for notational brevity.

2.2. Feature selection

Feature selection plays a critical role in the success of a learning algorithm in problems involving a significant number of irrelevant features. Here, we use the term feature to refer to both genetic and clinical markers. Microarray profiling is a powerful technique that allows researchers to examine the expression levels of tens of thousands of genes in a cell or a tissue simultaneously. However, it also poses a serious challenge to the existing machine-learning algorithms. With relatively small sample size, a learning algorithm can easily overfit training data, resulting in a zero training error but a very poor generalization performance on unseen data. A commonly used practice to correct for overfitting is to select a small feature subset such that the performance of a learning algorithm is optimized. Compared to the classifier design, feature selection still, to date, lacks a rigorous theoretical treatment. Most existing feature selection algorithms rely on heuristic combinatorial search and thus cannot provide any guarantee of optimality. This is largely due to the difficulty in defining an objective function that can be easily optimized by some well-established optimization techniques. In the presence of thousands of irrelevant genes, even heuristic searches become computationally unfeasible. For this reason, in microarray data analysis, nearly all of the gene selection algorithms resort to filter type methods that evaluate genes individually, e.g. *t*-test and Fisher score (Dudoit *et al*., 2002; Golub *et al*., 1999). The limitations of filter methods for feature selection are summarized as follows:

(1) Filter methods are unable to remove redundant features. For example, if a gene is top ranked, its co-regulated genes will also have high ranking scores. It is a well-established fact in machine learning that redundant features may not improve but rather deteriorate classification performance (Kohavi and John, 1997). This fact is largely ignored in many microarray data analyses. From the clinical perspective, the examination of the expression levels of redundant genes will not improve clinical decisions but increase medical examination costs needlessly.

(2) Filter methods evaluate the goodness of features individually, while neglecting the possible correlation information among them (Li *et al*., 2004; Dudoit *et al*., 2002). Some features may receive low ranking scores when evaluated separately, but can provide critical information when combined with other features. One possible solution to this problem is to use wrapper type methods that use a classifier to evaluate the goodness of selected feature subsets (Kohavi and John, 1997). However, with tens of thousands of features, it is computationally unfeasible to perform the combinatorial searching required in a wrapper method.

We have recently developed a new feature selection algorithm, referred to as I-RELIEF (Sun and Li, 2006) to alleviate the aforementioned drawbacks of filter methods and the computational issue of wrapper methods. I-RELIEF is one of the first feature selection algorithms that utilize the performance of a nonlinear classifier when searching for informative features, and yet can be implemented efficiently through optimization and numerical analysis techniques, instead of combinatorial searching. Below we present a brief review of I-RELIEF. Let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ denote a training dataset, where $\mathbf{x}_n$ is the nth data sample and $y_n \in \{\pm 1\}$ is the corresponding class label, i.e. metastasis or no metastasis. The ith component of $\mathbf{x}_n$ records the expression level of the ith gene in the nth sample. We define a margin for the sample $\mathbf{x}_n$ as $\rho_n = d(\mathbf{x}_n - \mathrm{NM}(\mathbf{x}_n)) - d(\mathbf{x}_n - \mathrm{NH}(\mathbf{x}_n))$, where $\mathrm{NM}(\mathbf{x}_n)$ and $\mathrm{NH}(\mathbf{x}_n)$ are the nearest miss and nearest hit of $\mathbf{x}_n$, which can be regarded as two functions that given an input $\mathbf{x}_n$ return the nearest neighbors of $\mathbf{x}_n$ from the opposite and same classes, respectively, and $d(\cdot)$ is a distance function defined as $d(\mathbf{x}) = \sum_i |\mathrm{x}_i|$. Note that $\rho_n > 0$ if only if $\mathbf{x}_n$ is correctly classified by a one-nearest-neighbor classifier. One natural idea is to scale each feature such that the averaged margin in a weighted feature space is maximized:

$$\begin{aligned} &\max_{\mathbf{w}} \sum_{n=1}^{N} \rho_n(\mathbf{w}) \\ &= \max_{\mathbf{w}} \sum_{n=1}^{N} \sum_{i=1}^{I} w_i(|\mathbf{x}_n^{(i)} - \mathrm{NM}^{(i)}(\mathbf{x}_n)| - |\mathbf{x}_n^{(i)} - \mathrm{NH}^{(i)}(\mathbf{x}_n)|) \\ &\quad \text{s.t. } \|\mathbf{w}\|_2^2 = 1, \mathbf{w} \geq 0, \end{aligned} \tag{1}$$

where $\rho_n(\mathbf{w})$ is the margin of $\mathbf{x}_n$ computed with respect to $\mathbf{w}$. The constraint $\|\mathbf{w}\|_2^2 = 1$ prevents the maximization from increasing without bound, and $\mathbf{w} \geq 0$ ensures that the w-weighted distance is a metric. We have proven that the optimization scheme in Equation (1) can be solved with a closed-form solution, and is equivalent to the well-known RELIEF algorithm (Kira and Rendell, 1992; Sun and Li, 2006). Note that the use of the block distance in the margin definition is consistent with the original formulation of RELIEF; other distance functions can also be used. For example, in Gilad-Bachrach *et al*., 2004, Euclidean distance is used in defining a margin, which, however, leads to a difficult nonconvex optimization problem. Due to the feedback of the performance of a nonlinear classifier when searching for useful features, RELIEF usually performs better than filter methods. One major drawback of RELIEF, however, is that the nearest-neighbors are defined in the original feature space, which is highly unlikely to be the ones in the weighted space. In the presence of many irrelevant features, which is the case in microarray data analysis, the performance of RELIEF can degrade significantly. I-RELIEF provides an analytic solution to mitigate the problem of RELIEF.

We first define two sets $\mathcal{M}_n = \{i : 1 \leq i \leq N, y_i \neq y_n\}$ and $\mathcal{H}_n = \{i : 1 \leq i \leq N,\ y_i = y_n, i \neq n\}$, associated with each sample $\mathbf{x}_n$. Suppose that we have known, for each sample $\mathbf{x}_n$, its nearest hit and miss, the indices of which are recorded in the set $\mathcal{S}_n = \{(s_{n1}, s_{n2})\}$, where $s_{n1} \in \mathcal{M}_n$ and $s_{n2} \in \mathcal{H}_n$. Then the objective function we want to optimize can be formulated as

$$C(\mathbf{w}) = \sum_{n=1}^{N} (\|\mathbf{x}_n - \mathbf{x}_{s_{n1}}\|_{\mathbf{w}} - \|\mathbf{x}_n - \mathbf{x}_{s_{n2}}\|_{\mathbf{w}}), \tag{2}$$

where $\|\mathbf{x}\|_{\mathbf{w}} = \sum_i w_i |x_i|$. Equation (2) can be easily optimized by using RELIEF. However, we do not know the set $\mathcal{S} = \{\mathcal{S}_n\}_{n=1}^N$. By following the principle of the Expectation Maximization algorithm, we regard the elements of $\{\mathcal{S}_n\}_{n=1}^N$ as hidden random variables, and derive the probability distributions of the unobserved data. We first make a guess on the weight vector $\mathbf{w}$. The probability of the ith data point being the nearest miss of $\mathbf{x}_n$ if $i \in \mathcal{M}_n$, or being the nearest hit of $\mathbf{x}_n$ if $i \in \mathcal{H}_n$, can be naturally defined as

$$P_m(i \,|\, \mathbf{x}_n, \mathbf{w}) = \frac{f(\|\mathbf{x}_n - \mathbf{x}_i\|_{\mathbf{w}})}{\sum_{j \in \mathcal{M}_n} f(\|\mathbf{x}_n - \mathbf{x}_j\|_{\mathbf{w}})},$$

and

$$P_h(i \,|\, \mathbf{x}_n, \mathbf{w}) = \frac{f(\|\mathbf{x}_n - \mathbf{x}_i\|_{\mathbf{w}})}{\sum_{j \in \mathcal{H}_n} f(\mathbf{x}_n - \mathbf{x}_j\|_{\mathbf{w}})},$$

respectively, where $f(\cdot)$ is a kernel function. One commonly used kernel function is $f(d) = \exp(-d/\sigma)$, where σ is a user defined parameter. In the experiment, we set $\sigma = 2$ based on our empirical experience. (In the Supplementary material, we show that the choice of the tuning parameter is not critical, and the algorithm performs similarly for a large range of sigma values.) For notational brevity, we define $\alpha_{i,n} = P_m(i \,|\, \mathbf{x}_n, \mathbf{w}^{(t)})$, $\beta_{i,n} = P_h(i \,|\, \mathbf{x}_n, \mathbf{w}^{(t)})$, $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_2 = 1, \mathbf{w} \geq 0\}$, $\mathbf{m}_{n,i} = |\mathbf{x}_n - \mathbf{x}_i|$ if $i \in \mathcal{M}_n$, and $\mathbf{h}_{n,i} = |\mathbf{x}_n - \mathbf{x}_i|$ if $i \in \mathcal{H}_n$. I-RELIEF can be summarized as follows:

Step-1: After the tth iteration, the Q function is calculated as:

$$\begin{aligned} Q(\mathbf{w} \,|\, \mathbf{w}^{(t)}) &= \mathrm{E}_{\{\mathcal{S}\}}[C(\mathbf{w})] \\ &= \sum_{n=1}^{N} \left(\sum_{i \in \mathcal{M}_n} \alpha_{i,n} \|\mathbf{x}_n - \mathbf{x}_i\|_{\mathbf{w}} - \sum_{i \in \mathcal{H}_n} \beta_{i,n} \|\mathbf{x}_n - \mathbf{x}_i\|_{\mathbf{w}} \right) \\ &= \sum_{n=1}^{N} \left(\sum_j w_j \sum_{i \in \mathcal{M}_n} \alpha_{i,n} m_{n,i}^j - \sum_j w_j \sum_{i \in \mathcal{H}_n} \beta_{i,n} h_{n,i}^j \right) \\ &= \mathbf{w}^T \sum_{n=1}^{N} (\bar{\mathbf{m}}_n - \bar{\mathbf{h}}_n) = \mathbf{w}^T \mathbf{v}, \end{aligned} \tag{3}$$

where $\bar{\mathbf{m}}_n = \sum_{i \in \mathcal{M}_n} \alpha_{i,n} \mathbf{m}_{n,i}$ and $\bar{\mathbf{h}}_n = \sum_{i \in \mathcal{H}_n} \beta_{i,n} \mathbf{h}_{n,i}$.

Step-2: The re-estimation of $\mathbf{w}$ in the $(t+1)$ th iteration is:

$$\mathbf{w}^{(t+1)} = \arg\max_{\mathbf{w} \in \mathcal{W}} Q(\mathbf{w} \,|\, \mathbf{w}^{(t)}) = (\mathbf{v})^+ / \|(\mathbf{v})^+\|_2, \tag{4}$$

where $(\nu_i)^+ = \max(\nu_i, 0)$. The above two steps iterate alternatingly until convergence, i.e. $\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| < \theta$, where θ is a small positive number. In Sun and Li, 2006, we have mathematically proven that I-RELIEF converges to a unique solution regardless of the initial weights if the kernel function is properly selected. The convergence is usually achieved within a few iterations.

I-RELIEF combines the merits of both filter and wrapper methods. Note that the objective function optimized by I-RELIEF approximates the leave-one-out accuracy of a nearest-neighbor classifier. Therefore, I-RELIEF can be regarded as a wrapper method, and thereby it naturally addresses the issues of feature correlation and the removal of redundant features. Moreover, I-RELIEF can be solved analytically, and thus avoids any heuristic combinatorial search. The effectiveness of the algorithm has been demonstrated through large-scale experiments on simulated data and six microarray datasets (Sun and Li, 2006). In the Supplementary material, a simulation study of I-RELIEF on a toy example is presented for illustration purpose.

3 EXPERIMENTS

3.1 Experimental setup

In a computational study using microarray data with small sample sizes, special care must be taken in experimental protocols to avoid possible overfitting of a computational model to training data. One particular problem in many microarray data analyses is an

incomplete cross-validation method that uses the same dataset for both training and testing, resulting in over-optimistic performances not reproducible in other independent validation studies (Simon *et al*., 2003; Simon, 2005; Brenton *et al*., 2005). To avoid this problem, we adopt a rigorous experimental protocol proposed in Wessels *et al*., 2005 with the leave-one-out cross validation (LOOCV) method. In each iteration, one sample is held out for testing and the remaining samples are used for training. The experimental protocol consists of two loops: inner and outer loops. In the inner loop, LOOCV is performed to estimate the optimal classification parameters based on the training data provided by the outer loop. In the outer loop, the held-out sample is classified by using the best parameters from the inner loop. The experiment is repeated until each sample has been used for testing.

The classification parameters that need to be specified in the inner loop include the kernel width of I-RELIEF, the structural parameters of a classifier (e.g. the regularization parameter in SVM and the number of the hidden nodes in NNW), and the number of the features used in a classifier, which leads to a multi-dimensional parameter searching. To make the experiment computationally feasible, we adopt some heuristic simplifications. Linear discriminant analysis (LDA) is used to estimate classification performances. One major advantage of LDA, compared to other classifiers, such as SVM and NNW is that LDA has no structural parameters. We then predefine the kernel width $\sigma = 2$, and estimate the number of features through LOOCV in the inner loop. The use of LDA is further justified by other research work. Simon pointed that there may not be sufficient information in most microarray datasets to support nonlinear classifiers (Simon, 2005). In the analyses performed by van't Veer *et al*., the 70-gene signature was derived from the same dataset, and the samples were classified using a correlation based classifier. It can be shown that the correlation based classifier is a special case of LDA, with the within-class scatter matrix being replaced by an identity matrix **I**. In Edén *et al*., 2004 where a NNW classifier was constructed, it was found through cross-validation that a NNW without hidden layers performed the best, which is actually a linear classifier. We comment that a comprehensive parameter searching may lead to a more accurate prediction performance but with a much higher computational complexity.

We demonstrate the predictive values of the hybrid prognostic signature derived from the genetic and clinical markers by comparing its performance with those of the clinical markers that are used as the features in a well trained LDA classifier, St. Gallen criterion and the 70-gene signature[3]. The performances of the 70-gene signature and the clinical markers are estimated through LOOCV. Hence, the held-out testing sample is not involved in the identification of a gene signature. It should be noted that the signature identified in each iteration is very likely to be different from the one reported in van't Veer *et al*., 2002. However, the LOOCV error provides us with an unbiased estimation on how the gene signature so-produced performs on unseen data (Simon, 2005) (c.f. Section 3.2).

[3]We follow the experimental procedure outlined in Veer *et al*., 2002 that first identified the top 70 genes and then assessed its predictive value by using a correlation based classifier. The detailed description is presented in the Supplement.

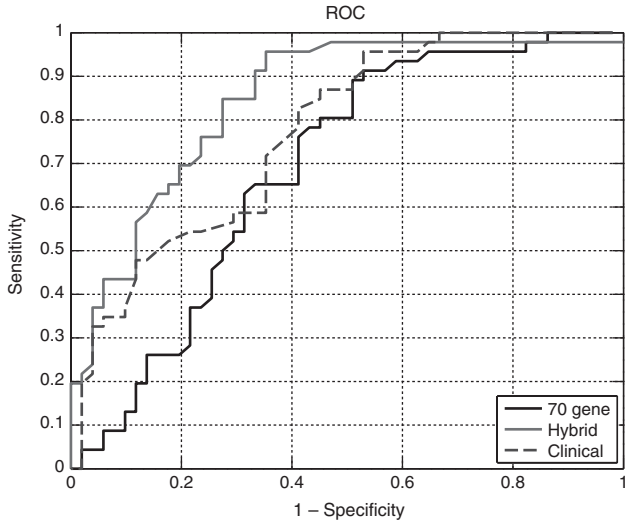

Fig. 1. ROC curves of three methods. A colour version of this figure is available as supplementary data.

Table 1. Prognostic Results (90% sensitivity)

Methods	Specificity	Odd ratio (95% CI)	Hazard ratio (HR) HR (95% CI)	*P*-value
70-gene	24/51 = 47%	9.3 (2.9–30.0)	6.0 (2.0–17.0)	<0.001
St. Gallen	6/51 = 12%	N/A	N/A	N/A
Clinical	25/51 = 48%	9.3 (2.9–30.0)	6.2 (2.2–17.6)	<0.001
Hybrid	34/51 = 67%	21.0 (6.5–68.3)	11.1 (3.9–31.5)	<0.001

3.2 Results

With a small sample size, some performance measurements, such as odds ratios are heavily influenced by the choice of a decision threshold. A receiver operating characteristic (ROC) curve obtained by varying a decision threshold can give us a direct view on how a classifier performs at the different sensitivity and specificity levels. In Figure 1, we plot the ROC curves of three classifiers based on the hybrid signature, the 70-gene signature and the clinical markers. We observe that the hybrid signature significantly outperforms both the 70-gene signature and the clinical markers, whereas the latter two approaches perform similarly. By following the study of van't Veer and colleagues (van't Veer *et al*., 2002), a threshold is set for each classifier such that the sensitivity of each classifier is equal to 90%. The corresponding specificities are computed and reported in Table 1. For comparison, the specificities of the St. Gallen criterion are also reported. Both the 70-gene signature and the clinical markers significantly outperform the St. Gallen criterion, as reported in the literature, and the hybrid signature improves the specificities of the 70-gene signature and the clinical markers by 20%. We point out that our estimation of the specificity of the 70-gene signature is worse than that reported in van't Veer *et al*., 2002 and Weigelt *et al*., 2005 (47% versus 73%), but is consistent with that in the follow-up validation study of the 70-gene signature on a larger dataset (van De Vijver *et al*., 2002) (53%). This is because 76 samples in van't Veer's dataset that were used for performance estimation were also involved in the identification

of the gene signature, which led to a biased estimate of the prediction performance of the signature. Our result suggests that LOOCV can effectively correct for this bias.

We compute the odds ratio (OR) of four approaches for developing distant metastases within five years between the patients with a good prognostic signature and the patients with a bad prognosis. The results are reported in Table 1. We observe that the 70-gene signature has the same OR (9.3, 95% confidence interval (CI): 2.9−30.0) as the clinical markers. This result is consistent with the findings reported in Edén *et al*., 2004. We also note that the hybrid signature gives a much higher OR (21.0, 95% CI: 6.5−68.3) than either genetic or clinical markers.

To further demonstrate the predictive value of the hybrid signature in assessing the risk of developing distant metastases in breast cancer patients, survival data analyses of four approaches are also performed[4]. The Kaplan–Meier curve of the hybrid signature, plotted in Figure 2, shows a significant difference in the probability of remaining free of distant metastases in patients with a good signature and the patients with a bad prognostic signature (P-value <0.001). The Mantel–Cox estimation of hazard ratio of distant metastases within five years for the hybrid signature is 11.1 (95% CI: 3.9−31.5, P-value <0.001), which is superior to either genetic or clinical markers alone.

More experimental results can be found in the Supplementary material.

3.3 Hybrid signature

With a small sample size, each iteration in LOOCV may generate a different prognostic signature since training data are different (Simon, 2005). In our study, we find that the majority of the iterations identify the same hybrid signature that consists of only three gene markers and two clinical markers (Supplementary Table 1). Note that the hybrid signature is markedly shorter than the 70-gene signature.

The two clinical markers in the hybrid signature are tumor grade and angio-invasion. Histological grading of tumors has been shown in numerous studies to provide useful prognostic information in breast cancer (Elston *et al*., 1991). The grade represents a morphological assessment of the degree of differentiation of the tumor as evaluated by the percentage of tubule formation, the degree of nuclear pleomorphism and the presence of mitoses. Grade 1 tumors have a low risk of metastases; grade 2 tumors have an intermediate risk of metastases and grade 3 tumors have a high risk of metastases. Patients with grade 1 tumors have a significantly better survival rate than those with grade 2 or 3 tumors (Elston *et al*., 1991). An essential step in the metastatic cascade is (lympho)vascular invasion, or the penetration of tumor cells into lymph and/or blood vessels in and around the primary tumor. Accordingly, the observation of 3 or more tumor cell emboli in tumor-associated vessels has been correlated with the presence of LN metastases and with poor prognosis in patients with breast cancer (de Mascarel *et al*., 1998; Pinder *et al*., 1994).

[4]It is not clear whether at 5 years post-surgery, patients had died from distant metastasis or that the clinicians were unable to continue follow-up for other reasons. Some researchers (Edén *et al*., 2004) treated the patients who survived more than 5 years as if they lost follow-up, while in our experiment, we consider them as ''dead''. Therefore, the results of the 10-year prognosis are not reliable.

The three genetic markers in the hybrid signature include AL080059, CEGP1 and PRAME, of which CEGP1 and AL080059 are also listed in the 70-gene signature. The CEPG1 gene (also known as SCUBE2, EGF2-like 2 and ASCL3), is located on human chromosome 11p15 and has homology to the achaete-scute complex (ASC) of genes in the basic helix–loop–helix (bHLH) family of transcription factors. The exact biological role for CEGP1 (SCUBE2) is still unknown, but the gene encodes a secreted and cell-surface protein containing EGF and CUB domains (Yang *et al*., 2002). The epidermal growth factor motif is found in many extracellular proteins that play an important role during development, and the CUB domain is found in several proteins implicated in the regulation of extracellular process, such as cell–cell communication and adhesion (Grimmond *et al*., 2000). Expression of SCUBE2 has been detected in vascular endothelium and may play important roles in development, inflammation and perhaps carcinogenesis (Yang *et al*., 2002). The expression of SCUBE2 was recently reported to be associated with ER status in a recent SAGE-based study of breast cancer specimens (Abba *et al*., 2005). The AL080059 label refers to a sequence obtained from a human cDNA clone, but subsequent analysis has revealed significant homology with the TSPY-like 5 (TSPYL5) gene, and with other human proteins, including NAPs, factors which play a role in DNA replication (Schnieders *et al*., 1996). It is thought that NAPs act as histone chaperones, shuttling histones from their site of synthesis in the cytoplasm to the nucleus. Histone proteins are involved in regulating chromatin structure and accessibility and therefore can impact gene expression (Rodriguez *et al*., 1997), thus, a role in tumor cell phenotype can be proposed. Although both AL080059 and CEGP1 were found to be significantly over-expressed in our studies of a breast tumor metastases model (Goodison *et al*., 2005), neither the AL080059 nor CEGP1 genes have been evaluated independently in human cancers. Conversely, the expression of the preferentially expressed antigen in melanoma (PRAME) gene has been linked to human disease, including cancer. PRAME is classed as a cancer-testis antigen (CTA), a group of tumor-associated antigens that represent possible target proteins for immuno-therapeutic approaches. Their expression is encountered in a variety of malignancies but is negligible in healthy tissues, with male germinal cells being the exception (Juretic *et al*., 2003). PRAME was first discovered in a patient with melanoma (Ikeda *et al*., 1997), and has since been found to be expressed in a large variety of cancer cells including squamous cell lung carcinoma, medulloblastoma, neuroblastoma, renal cell carcinoma and acute leukemia (Matsushita *et al*., 2003). Our study raises the possibility that therapies targeted to PRAME may be beneficial in breast cancers also.

4 DISCUSSION

We present some discussion on the optimality and uniqueness of prognostic signatures. Due to these issues, among others, the appropriateness of the existing gene signatures being ready for clinical trials is currently under hot debate (Brenton *et al*., 2005; Weigelt *et al*., 2005; Loi *et al*., 2006). Since many potential readers of this paper are from the oncology community, we start the discussion with a relatively simple machine-learning example. This example was first used by Trunk (Trunk, 1979) to demonstrate the existence of the curse of dimensionality. We find that

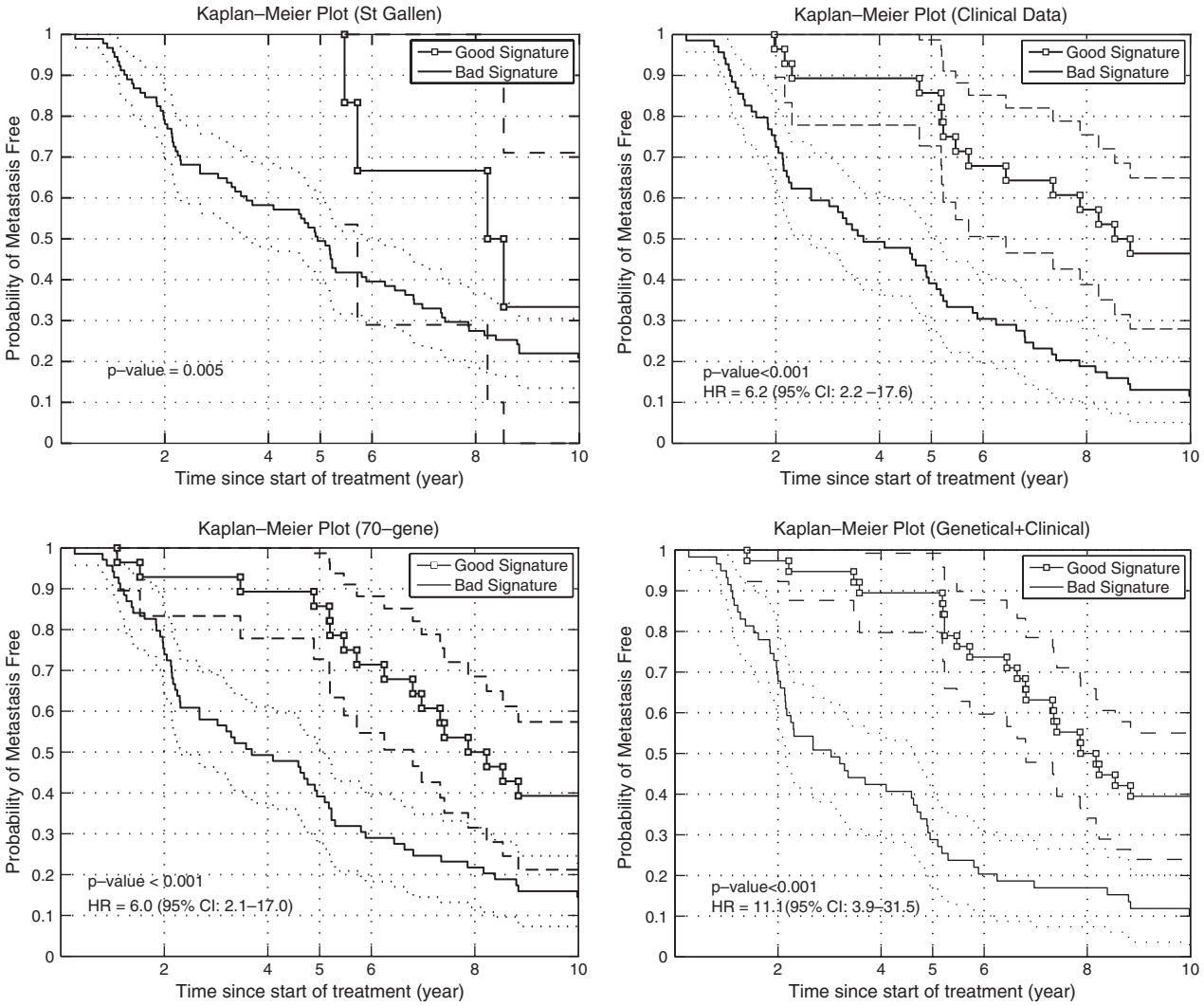

Fig. 2. Kaplan–meier estimation of the probabilities of remaining distant metastases free in patients with a good or bad prognostic signature, determined by the St. Gallen criterion, clinical markers, 70-gene signature and hybrid signature. The *P*-values is computed by the use of log-rank test.

Trunk's experiment, when applied to the research of breast cancer prognosis, reveals to us much beyond the curse of the dimensionality. Consider the following binary classification problem. The *a priori* probabilities $P(C_1) = P(C_2) = 1/2$, and the class conditional probabilities are Gaussian, given by $p(\mathbf{x}\,|\,C_1) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$ and $p(\mathbf{x}\,|\,C_2) \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I})$, respectively, where $\boldsymbol{\mu}$ is the mean vector, the ith components of which is $(1/i)^{1/2}$, and $\mathbf{I}$ is the identity matrix. The task is to construct a classifier based on a given training dataset. All of the *a priori* knowledge is given except for the mean vector $\boldsymbol{\mu}$, which is estimated from training data. The classification accuracies, as a function of the number of features used in a constructed classifier for four different training sample sizes (i.e. 20, 50, 100 and 200) averaged from 100 runs, are plotted in Figure 3. From the figure, we arrive at the following observations:

(1) For a given sample size, the inclusion of additional features beyond a certain point leads to a higher error. It can be shown that with a finite sample size, the classification error converges to one-half when the number of features goes to infinite (Trunk, 1979). Note that in Trunk's data model, each feature contains a certain amount of discriminant information. This observation, when applied to breast cancer prognosis, implies that with a limited number of training samples, some features, though having some predictive values in breast cancer prognosis when evaluated individually, do not necessarily improve the predictive performance of a computational model when used together with other features, and sometimes may even deteriorate performance. This highlights the need for performing feature selection.

(2) With the increase of sample sizes, the numbers of the features corresponding to the optimum performance are also increased. For example, with 20 samples, the classification accuracy peaks around 30 features, whereas for 200 samples, the peak occurs around 200 features (Fig. 3). This observation, when applied to the research of breast cancer prognosis,

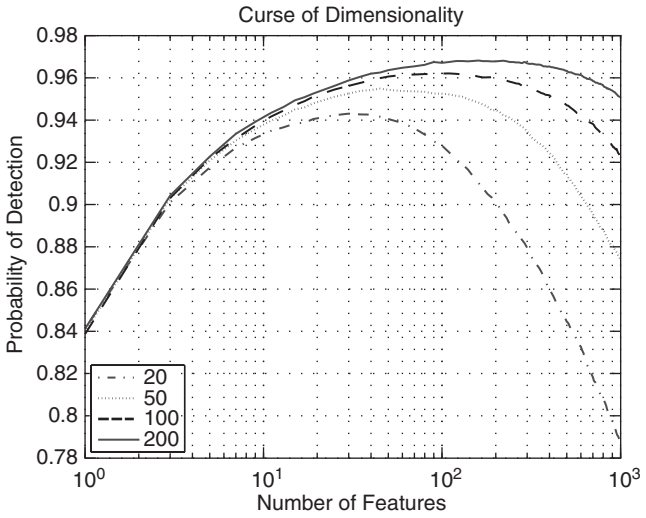

Fig. 3. Trunk's experiment showing the curse of dimensionality. A colour version of this figure is available as supplementary data.

indicates that a prognostic signature derived from a small training dataset is likely to be lengthened when a larger training dataset is used.

(3) There exists a range in the feature dimensionality where a classifier achieves the close-to-optimum performance. Moreover, with the inclusion of more samples, the range becomes even larger. (Note that the *x*-axis of Fig. 3 is log-scaled.) It means that, for a given training dataset, there may exist several different signatures with a similar predictive value. This observation, together with other factors (e.g. the existence of co-regulated genes and the use of different microarray platforms and data processing algorithms. Interested readers may refer to Dalton *et al*., 2006 for more detailed discussion.), may provide an explanation as to why the gene signatures identified in some recent independent studies are different.

For clinical applications, what we are really interested in is not whether there exists several different signatures having a similar predictive value, but whether these signatures have achieved the optimum, or close-to-optimum performance. After decades of research on breast cancer prognosis, many prognostic markers have been reported in the literature, including clinical markers and gene signatures. Many of them are single-marker prognostic and predictive studies. A critical question remains unanswered to date: *what is the best we can perform in breast cancer prognosis given all clinical and genetic information using advanced computational algorithms?* Without further optimization, an expensive clinical validation trial of a prognostic signature may merely repeat the already established predictive values of the signature, and yet, cannot prove its optimality for clinical applications. In this paper, we have presented a computational study clearly demonstrating the feasibility of utilizing both clinical and genetic information simultaneously for more accurate breast cancer prognosis. We believe that this is an advantageous direction to pursue in future breast cancer prognosis studies.

Our experiment is based on van't Veer's data, which was obtained from only 97 tissue samples. We demonstrate through Trunk's experiment that identifying prognostic signatures for breast cancer prognosis is necessarily an ongoing and dynamic process, in which, with the inclusion of more patient samples, a prognostic signature will be continuously lengthened and refined, whereby the performance of a prognostic signature will be improved accordingly and finally stabilized.

5 RELATED WORK

From the machine-learning perspective, it is a straightforward idea to integrate all available information for a classification task. Some efforts have been made in this direction for breast cancer prognosis but with little success. Ritz (Ritz, 2003) combined both genetic and clinical information in a NNW for breast cancer prognosis but found that the combination did not improve the performance. Dettling *et al*., 2004 applied penalized logistic regression analysis to predict cancer prognosis for the same dataset. They found that none of the clinical variables entered the model and concluded that the clinical data did not contain any useful independent information for prediction, given the gene expression profile. In Gevaert *et al*., 2006, a Bayesian network was developed to perform breast cancer prognosis. The results showed that although a Bayesian network that used both genetic and clinical information can lead to a simpler classifier with fewer genes, which is consistent with our finding, the Bayesian network performed similarly to the 70-gene signature. We emphasize that these negative results do not necessarily mean that the clinical data contains no additional information to the genetic data; it only tells us that with their models the applied combination strategy did not work. This highlights the difficulty of designing a successful combination strategy.

6 CONCLUSION

In this paper, we applied a new mathematical model to predict the likelihood of disease recurrence and metastases in breast cancer. Our preliminary study has shown that a hybrid signature can provide significantly improved prognostic specificity over the existing gene signatures and the current clinical systems by about 20% and 60%, respectively. We have also presented an informative discussion on the issue of the curse of the dimensionality in the context of breast cancer prognosis. We believe that researchers, particularly from the oncology community, should benefit from the discussion.

To fully address the question of what is the best we can perform in breast cancer prognosis given all available information, as posed in Section 4, larger-scale computational studies involving more patient data and which compare different learning algorithms are required and are under way in our laboratory.

ACKNOWLEDGEMENTS

Conflict of Interest: none declared.

REFERENCES

Abba,M. *et al.* (2005) Gene expression signature of estrogen receptor α status in breast cancer. *BMC Genomics*, **6**, 74–81.

Brenton,J. *et al.* (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J. Clin. Oncol.*, **23**, 7350–7360.

Dalton,W. *et al.* (2006) Cancer biomarkers–an invitation to the table. *Science*, **312**, 1165–1168.

de Mascarel,I. *et al*. (1998) Obvious peritumorous emboli: an elusive prognostic factor reappraised: multivariate analysis of 1320 node-negative breast cancers. *Eur. J. Cancer*, **34**, 58–65.

Dettling,M. *et al*. (2004) Finding predictive gene groups from microarray data. *J. Multivariate Anal.*, **1**, 106–131.

Duda,R. *et al*. (2000) *Pattern Classification*, J. Wiley, NY.

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.

Edén,P. *et al*. (2004) 'Good old' clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur. J. Cancer*, **40**, 1837–1841.

Eifel,P. *et al*. (2000) National Institutes of Health consensus development conference statement: adjuvant therapy for breast cancer. *J. Natl. Cancer Inst.*, **93**, 979–989.

Elston,C. *et al*. (1991) Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, **19**, 403–410.

Gevaert,O. *et al*. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, **22**, 184–190.

Gilad-Bachrach,R. *et al*. (2004) Margin based feature selection—theory and algorithms. In *Proceedings of 21st International Conference Machine Learning,* pp. 43–50.

Goldhirsch,A. *et al*. (2003) Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. *J. Clin. Oncol.*, **21**, 3357–3365.

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

Goodison,S. *et al*. (2005) The RhoGAP protein DLC-1 functions as a metastasis suppressor in breast cancer cells. *Cancer Res.*, **65**, 6042–6053.

Grimmond,S. *et al*. (2000) Cloning, mapping, and expression analysis of gene encoding a novel Mammalian EGF-related protein (SCUBE1). *Genomics*, **70**, 74–81.

Ikeda,H. *et al*. (1997) Characterization of an antigen that is recognized on a melanoma showing partial HLA loss by CTL expressing an NK inhibitory receptor. *Immunity*, **6**, 199–208.

Juretic,A. *et al*. (2003) Cancer/testis tumour-associated antigens: immunohistochemical detection with monoclonal antibodies. *Lancet Oncol.*, **4**, 104–109.

Kira,K. and Rendell,L. (1992) A practical approach to feature selection. In *Proceedings of 9th International Conference Machine Learning,* pp. 249–256.

Kohavi,R. and John,G. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Li,T. *et al*. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Loi,S. *et al*. (2006) Molecular forecasting of breast cancer: time to move forward with clinical testing. *J. Clin. Oncol.*, **24**, 721–722.

Matsushita,M. *et al*. (2003) Preferentially expressed antigen of melanoma (PRAME) in the development of diagnostic and therapeutic methods for hematological malignancies. *Leuk. Lymphoma*, **44**, 439–444.

Pinder,S. *et al*. (1994) Pathological prognostic factors in breast cancer. III. Vascular invasion: relationship with recurrence and survival in a large study with a long-term follow-up. *Histopathology*, **24**, 41–47.

Ritz,C. (2003) Comparing prognostic markers for metastases in breast cancer using artificial neural networks. Master thesis, Lund University, Sweden.

Rodriguez,P. *et al*. (1997) Functional characterization of human nucleosome assembly protein-2 (NAP1L4) suggests a role as a histone chaperone. *Genomics*, **44**, 253–265.

Schnieders,F. *et al*. (1996) Testis-specific protein, Y-encoded (TSPY) expression in testicular tissues. *Hum. Mol. Genet.*, **5**, 1801–1807.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.*, **95**, 14–18.

Simon,R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.*, **23**, 7332–7341.

Sun,Y. and Li,J. (2006) Iterative RELIEF for feature weighting. In *Proceedings of 23rd International Conference Machine Learning,* pp. 913–920.

Trunk,G. (1979) A problem of dimensionality: a simple example. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1**, 306–307.

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

van De Vijver,M. *et al*. (2002) A gene-expression signature as a predict of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

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

Weigelt,B. *et al*. (2005) Breast cancer metastasis: markers and models. *Nat. Rev. Cancer*, **5**, 591–602.

Wessels,L. *et al*. (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755–3762.

Yang,R. *et al*. (2002) Identification of a novel family of cell-surface proteins expressed in human vascular endothelium. *J. Biol. Chem.*, **227**, 46364–46373.