Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 587405, 7 pages
doi:10.1155/2009/587405

Research Article

Simpler Evaluation of Predictions and Signature Stability for Gene Expression Data

Yvonne E. Pittelkow[1] and Susan R. Wilson[1,2]

[1] *Centre for Bioinformation Science, MSI, The Australian National University, Canberra, ACT 0200, Australia*
[2] *School of Mathematics & Statistics, Faculty of Science, Prince of Wales Clinical School, Faculty of Medicine, University of NSW, Sydney 2052, Australia*

Correspondence should be addressed to Yvonne E. Pittelkow, yvonne.pittelkow@anu.edu.au

Received 3 March 2009; Revised 31 July 2009; Accepted 3 November 2009

Recommended by Satoru Miyano

Scientific advances are raising expectations that patient-tailored treatment will soon be available. The development of resulting clinical approaches needs to be based on well-designed experimental and observational procedures that provide data to which proper biostatistical analyses are applied. Gene expression microarray and related technology are rapidly evolving. It is providing extremely large gene expression profiles containing many thousands of measurements. Choosing a subset from these gene expression measurements to include in a gene expression signature is one of the many challenges needing to be met. Choice of this signature depends on many factors, including the selection of patients in the training set. So the reliability and reproducibility of the resultant prognostic gene signature needs to be evaluated, in such a way as to be relevant to the clinical setting. A relatively straightforward approach is based on cross validation, with separate selection of genes at each iteration to avoid selection bias. Within this approach we developed two different methods, one based on forward selection, the other on genes that were statistically significant in all training blocks of data. We demonstrate our approach to gene signature evaluation with a well-known breast cancer data set.

1. Introduction

The era of personalised medicine is being widely heralded, with claims including "all human illness can be studied by microarray analysis, and the ultimate goal [] is to develop effective treatments or cures for every human disease by 2050" [1]. The first step is the selection of predictors from the many (often tens of) thousands of measurements that can be made using modern microarray technology.

As is well known amongst statisticians, genes identified from a microarray analysis as predictors of outcome, and so included in a molecular signature, depend greatly on the selection of patients in the training set. Hence it has been advocated that validation by repeated random sampling (RRS) should be used [2]. RRS is advocated elsewhere, for example for selecting genes for classification [3]. The RRS approach can quickly become extremely computer intensive. Alternatively, the far less computer intensive method of 10-fold cross-validation has been shown empirically, for two microarray data sets, to perform very well when compared with repeated random sampling [4].

K-fold cross-validation is a resampling method that is widely used in practice for estimating performance of a prediction rule ("signature") when there is insufficient data to split the data into the ideal 3 parts: training, validation, and testing. For microarray experiments where the number, M, of measurements (say gene-expression values on an Affymetrix chip) is extremely large compared with the number, s, of samples (chips), it is starting to become clear that some traditional data statistical practices may not always be appropriate.

There is considerable confusion about the application of cross-validation and many investigators have not understood the importance of signature selection at each of the K separate steps [4]. If the signal-to-noise ratio of the underlying function is relatively strong then the signature may not vary

very much. However, if it is not strong, and this is very likely when the number of measurements is very large compared with the number of samples, then the signature may vary appreciably; see [5, 6]. Further, it has been noted that "the most striking finding when comparing the significant lists [from different studies] is the virtually complete lack of agreement in the included genes … the present lack of coherence [] warrants further examination" [7].

The reason for the differences could be attributed to a number of factors, including methodology (model, algorithm, gene selection, preprocessing steps such as normalisation, transformation, etc.), data (different samples, heterogeneous samples, poor measurements, poor experimental designs leading to the presence of confounders, etc.), and the underlying biology (different genes from the same pathway, considered later). The fragility of a prediction model has been demonstrated [8] by showing that published results could not be replicated using the same data, when training and tests sets were interchanged and the same algorithms applied.

Here we consider a relatively large study [9] that was used to predict distant metastasis of lymph-node-negative primary breast cancer based on determination of a single signature. The data are available in series GSE2034 at the NCBI Genebank GEO. These data consist of gene expression values from 286 Affymetrix U133a chips each with gene expression values for over 22 000 "genes". Of these patients, 209 were oestrogen receptor-positive (ER^+) and 77 patients were oestrogen receptor-negative (ER^-). Briefly, the samples were a subset of frozen tumour samples from patients with lymph-node negative breast cancer which had been submitted for steroid-hormone receptor measurement from an intake of 25 hospitals; further details in [9]. Metastasis status was determined from follow-up examinations or confirmed following patient report. The (statistical) sample selection was nonrandom and the data show considerable variation in many known breast cancer prognostic factors, such as therapy type, age, menopausal status, tumour grade, and stage for example. Data on these factors were not available on the web.

In the following, the ER^+ patient data are chosen for illustrative purposes, noting that this selection also controls some of the overall patient heterogeneity. Use of 10-fold cross-validation is explored to determine those genes that are included in molecular signatures (lists). In other words, signature stability is evaluated, as is the variability of prediction measures.

2. Methods

First, gene expression values were log transformed (base 2) after zero values were set to 0.1.

Two methods, both of which correct for selection bias [4], were developed and applied to these breast cancer data. Before detailing the methods we outline the basic steps of K-fold cross-validation:

(a) randomly divide the data into K separate and approximately equal parts called validation sets;

(b) leave out one of the validation sets, then (b_1) perform the analysis using the combined data of the remaining K-1 parts (referred to as the "training set") and, usually, (b_2) validate the analysis (that is often making prediction/s) on the "validation set";

(c) repeat (b) K times, so that following the (usual) b_2 step each one of the (separate) K validation sets is used once (e.g., for prediction/s).

For method (i), within each training set at step (b_1), we first did an initial screening by applying a univariate Cox proportional hazards regression model (Cox PHM) to select those genes in each of the training sets that were statistically significant for metastasis-free survival. A (partial) likelihood ratio test, comparing the partial likelihood estimated for the gene to the null model, was used. Genes were retained when the likelihood ratio test was statistically significant at level α_1. Next, restricting attention to just these retained genes, we used a Cox PHM with forward selection to select amongst the genes available following the initial gene selection. The selection was terminated when the statistical significance for inclusion into the Cox PHM, using a log likelihood ratio test, was greater than level α_2. For step (b_2), the resulting model was used to predict relapse-free survival at 5 years on the validation data set. To evaluate predictive performance, a binary outcome was defined [2], namely whether the predicted probability of a patient was relapse-free, was greater than, or less than, 0.5. Then, three measures of predictive performance (also called accuracy of classification) were estimated, namely the true positive fraction (TPF), also known as sensitivity, the false positive fraction (FPF) and the proportion of correctly predicted samples. Specificity is 1-FPF. In a clinical setting, sensitivity and specificity are more relevant than the proportion correctly predicted (i.e., 1 – classification error) that is often the measure reported in the bioinformatics literature. It is straightforward to show that the proportion of correctly predicted samples is a weighted average of sensitivity and specificity. Note that the proportion correctly predicted can be misleading if the outcome is rare or very common.

Method (ii) followed method (i) except that instead of forward selection we selected those genes which were significant at level α_1 in all training sets. Then all these genes were used in the signature with a multivariate Cox PHM being fitted at each iteration. The validation at step (b_2) in the second iteration was as described for method (i). This approach has been used successfully on data from proteomic profiling to distinguish malignant pancreatic cancer from benign disease [10].

There is no general rule for the value of K, and $K = 1$, 5, 10, and $\sqrt{s}$ (where s is the sample size) have all been suggested; the trade-off is between bias and variance [5]. Following Ambroise & McLachlan [4], we chose $K = 10$. This choice seems to balance the trade-off between bias with higher values of K and variability as observed for low K values, particularly $K = 1$ (the so-called leave-one-one cross validation). Note that any two 10-fold cross-validation training sets have approximately 80% of the total sample in common.

Table 1: Annotation analysis of the signatures 1 to 10; the table entries show the number of genes found in each signature to include the function shown in the first column, with a blank indicating zero.

Function	1	2	3	4	5	6	7	8	9	10
Cell cycle		2	2	1		3	1	3		
Cell Proliferation	1					2		1		1
DNA repair				1		1	1	1		
Immune response		1		1		2	3	1		1
Cell Growth	1			2			1			
Transcription	3	3	1	2	2	3	2	1	1	2
Cell-cell signal			1					1	1	
Development		2		1	2				1	
ATP binding	2	2	3		2	3	3	4	2	1
Nucleotide binding	3	4	3	1	1	4	2	4	2	
DNA binding				2	2		1	1	1	
Cell adhesion	1		1		2	2		3	1	1
Golgi stack	1				1			1	1	
Kinase activity	2	3	3	1	2	1	1	1	1	1
Transferase activity	2	2	3		3	1	1	2	1	1

Here the cross-validation was applied in stratified form, so that approximately equal numbers of distant (i.e., beyond 5 years) metastases patients were included in the validation sets. We found that our substantive conclusions changed little if this restriction was relaxed.

The Cox proportional hazards model was chosen for modelling metastasis free survival, t, because of censoring in the data. The Cox PHM models the conditional hazard as the product of a baseline hazard, $\lambda_0(t)$, and an exponential form of a linear function of covariates, $\mathbf{z}$, here $\log_2$ gene expression values, as follows:

$$\lambda(t \mid z) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{z}). \quad (1)$$

The assumption of proportional hazards was evaluated and found adequate for a small number of significant genes.

Known gene function was examined using a text search on annotation available from the Affymetrix web site. The R software [11] with the package "survival" was used for analysis.

3. Results

3.1. Method (i). The initial screening of genes at step (b_1), using a univariate Cox regression analyses, with $\alpha_1 \leq .001$ identified from 210 to 342 genes in the 10 training sets. Altogether across the 10 training sets, 675 unique genes were found.

The 10 signatures built using a Cox PHM with forward selection in each training set and with $\alpha_2 \leq .001$, consisted of 11 to 18 genes. Genes were generally uncorrelated within each signature.

The 10 signatures have very few genes in common; that is, the signatures are very unstable. For example, one gene occurred in 6 signatures (the most common), 4 genes occurred in 4 signatures and 8 genes in 3 signatures. Seventy genes occurred only once. Also very few of the 60 genes in Wang et al.'s [9] signature for the subset of ER^+ patients occurred in the 10 signatures. Three of the signatures included no genes in common with the Wang signature and the signature with the largest number of genes in common had only four genes in common.

The discrepancies between the ten signatures might be thought to be occurring because correlated genes in the same pathway/s are being selected. Table 1 shows the number of genes found in each signature with different functional classes. These classes are based on those in Wang et al. (Table 4) [9, 12]. Functional classes that were found in none or only one signature are not shown. This analysis is not meant to be definitive but to illustrate the lack of consistency in apparent function between the signatures. We note that the variation demonstrated here may reflect the complexity of signalling processes in possibly many pathways, and/or the presence of tumour subtypes [13] and/or the heterogeneity in the sample.

Table 2 shows signature performance indices estimated on the training sets. As is well known now, when the performance is estimated on the same set as used to train the signature, the performance is over optimistic (i.e., biased [4]). Table 2 shows that each signature predicts distant metastasis very well for the data on which they were trained, even though the signatures were composed of a relatively small number of genes (average 15). However, Table 3 shows that the performance is considerably worse when the signature is used to predict metastasis-free status on the validation set. The test error, estimated as the proportion *misclassified*, using distant metastasis within 5 years as the defining mark, varies between 0.19 (= 1 − 0.810) and 0.55 (= 1 − 0.450). The average unbiased estimate of error rate is 32.6%, while the biased estimate of error rate is only 11.5%. The variance of the unbiased test error is 122.7.

Table 2: Biased estimators of prediction performance for the ten signatures estimated on the training sets.

Signature	Prop. of true positives	Prop. of false positives	Prop. correctly predicted
1	0.943	0.213	0.891
2	0.943	0.206	0.892
3	0.926	0.230	0.874
4	0.934	0.230	0.879
5	0.959	0.213	0.901
6	0.942	0.197	0.896
7	0.942	0.262	0.874
8	0.967	0.164	0.923
9	0.950	0.279	0.874
10	0.942	0.344	0.846
Average	0.945	0.234	0.885

Table 3: Unbiased estimators of prediction performance for the ten signatures estimated on the validation sets.

Signature	Prop. of true positives	Prop. false positives	Prop. correctly predicted
1	0.750	0.714	0.579
2	1.000	0.800	0.778
3	0.769	0.714	0.600
4	0.786	0.714	0.619
5	0.929	0.429	0.810
6	0.857	0.571	0.714
7	0.929	0.714	0.714
8	0.538	0.714	0.450
9	0.857	0.571	0.714
10	0.857	0.429	0.762
Average	0.827	0.637	0.674

3.2. Method (ii). We demonstrate the second method using 2 values of α_1 for the initial screening. With $\alpha_1 \leq .001$, there were 59 genes that were statistically significant in all training sets for metastasis-free survival, using a univariate Cox PHM. To assess performance of this method we estimated the Cox PHM using these 59 genes and predicted relapse-free survival at 5 years in each corresponding validation set. The true positive fractions (TPF) and false positive fractions (FPF) are plotted on the left in Figure 1. The location of the average pair (TPF, FPF) = (0.793 0.581) is shown. The average TPF is slightly lower than for method (i) but so also is the FPF which is desirable. Given the variability of the estimates, these averages can be considered approximately the same. The lower range of the FPF is less here than for method (i), whereas the comparison of the FPF values in the figure and with corresponding column in Table 3 indicates an improvement over method (i). The average error rate from the validation sets is 33.3%, which is similar to method (i), but with variance 74.4 which is much lower.

When $\alpha_1 \leq .0001$ was used for the initial gene selection, 14 genes were identified for inclusion in the signature. The pairs (TPF, FPF) estimated on the validation set for this signature are plotted on the right of Figure 1. The averages of TPF and FPF are 0.896 and 0.548, respectively with associated variances 0.009 and 0.042. The average estimate of error rate has decreased to 25.5%, with variance 50.0.

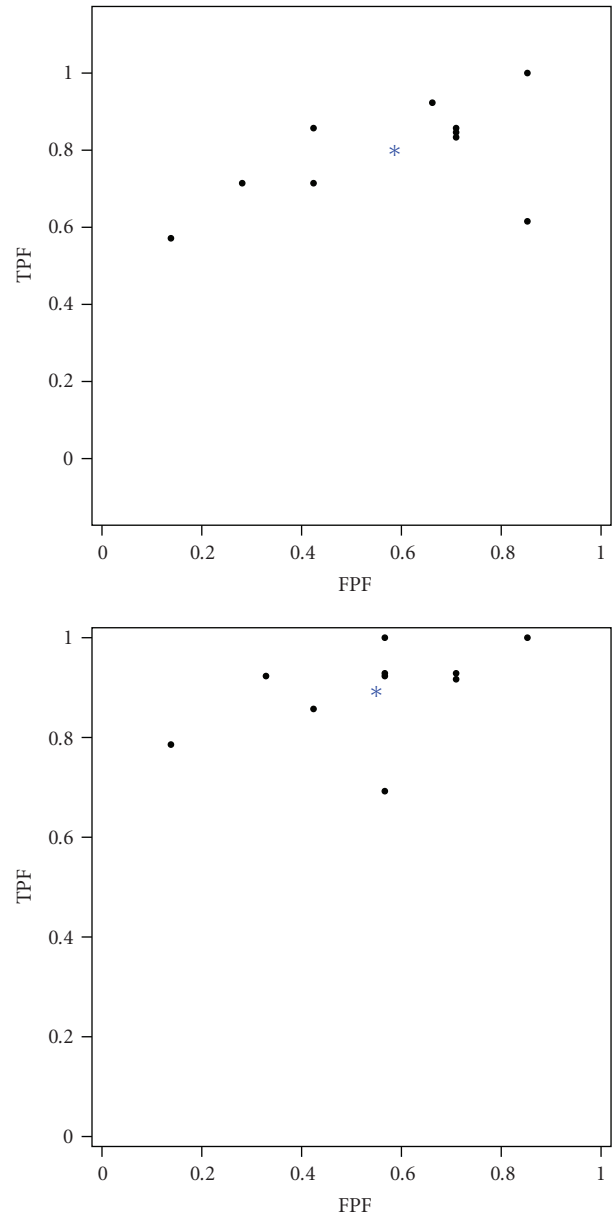

Figure 1: Performance assessment for method (ii). The pairs (TPF, FPF) estimated on the validation set are plotted in each graph. The assessment for the 59 gene signature ($\alpha_1 \leq .001$) is shown on the upper and that for the 14 gene signature ($\alpha_1 \leq .0001$) is shown on the lower.

4. Discussion

Cross-validation is shown to be a useful tool to assess performance and stability of molecular signatures used for prediction from microarray data. Such an evaluation is

necessary since the number of samples is small relative to the number of measurements. Alternatives to cross-validation for estimating predictive performance include the bootstrap and repeated random sampling [3, 4] for example; but for clarity in demonstrating instability of predictors, and because it is much less computer intensive, we concentrated here on cross-validation.

Within the cross-validation framework for developing gene signatures, there are many choices, including which initial screening approach to use and which method to build the signature. Each choice may lead to different genes being included in a signature. The choices made in the initial gene selection step here are somewhat arbitrary. Our initial choice, to select genes with significant ($\alpha_1 \leq 0.001$) association with breast cancer survival, was a strategy to reduce the number of genes so that the potential number was of the same order of magnitude as the number of samples. This initial gene selection step is not required for K-fold cross-validation in applications where the number of variables is closer to sample size. This screening of genes has interesting connections to the recent, independently derived, proposal of iterative sure independence screening (ISIS) for the linear model in ultrahigh dimensional feature space. ISIS was motivated by the need to deal with problems of large dimensionality for which accuracy of estimation and computational cost are two concerns. For the linear model, Fan and Lv [14] proposed reducing the dimensionality from high to a moderate scale that is below the sample size, by iteratively implementing a variable screening procedure and fitting the proposed model.

The presence of censored data indicates the use of a survival model for building the signatures. Since the Cox PHM has been found to be reasonably robust under misspecification [15], is well known in medical fields, and was used by Wang et al. [9], it was chosen for developing the signatures. This is preferable to the approach used by Ein-Dor et al. [16] that was based on genes correlated with survival, as "correlation" is not statistically appropriate in this context.

In method (i), to train the signature on the training data we used a stepwise procedure. It is well known that such methods find only a fraction of the models that fit the data well and can have the undesirable effect of over fitting the data. However, efficient screening of all models is not computationally possible given the number of genes (variables). Again, the use of a significance level of $\alpha_1 \leq 0.01$ is somewhat arbitrary. In further analyses (not reported) increasing α increased the number of genes and the lack of agreement between the signatures. The use of cross-validation helps to ascertain the extent of any over fitting but there is no known solution yet to finding the "optimal" model in this setting. It is good practice to use a number of approaches and to be satisfied only when substantive conclusions are unaltered by the algorithm, and basic assumptions are satisfied in so far as they can be tested. The importance of implementing the same gene selection rule/s for each of the training sets, in order to obtain an unbiased estimate of the prediction error, is stressed. Here there are two gene selection steps, and both need to be implemented at each training step. This is necessary because neither gene selection method is independent of the prediction method.

Method (ii) is novel, with its incorporation of selection based on variables that are common to *all* blocks using a cross-validation type of approach. This would appear to overcome the highly unstable list of genes identified as predictors of prognosis that is found using either repeated random sampling [3] or method (i), but needs to be further evaluated on other data sets. Note that method (ii) is comparatively simple, and computationally fast. Current research concerns evaluation of the performance and stability of signatures of both our variants of cross-validation with the SIS-style approaches to screening before fitting of the model/s. Since the data used for the selection of genes are not fully independent of the data used for validation, some bias may exist in estimates of performance [17] and alternative methods, including the use of 2-external cross-validation will be investigated.

There are many ways in which cross-validation can be developed so as to correct for selection bias [4], and we have used two. Another approach that has many similarities to our methods is to first consider a range of the numbers of genes to be used in the signature, say d_0 to d_1. Then *within* each of the K training blocks, perform K-1-fold cross-validation, and evaluate the signature for *each* value of d, selecting the value that gives the best classification rate for that block. The value of d selected may vary between the K training blocks; see [18] for further details. With the additional layer of cross-validation, as well as consideration of a range of values of d within each training block, this is more computer-intensive than our methods, although the associated variability may well be less. Comparative performance of these approaches is a future research project.

It is important to distinguish between the use of statistical models for prediction and for explanation [19]. In much of the literature this distinction is blurred. Here we have been concerned with choosing signatures for prediction. Some investigators attempt to use a signature to build pathway/s for explanatory purposes but such explanations are dubious when the signatures can be very unstable.

Sample design is an important, but often overlooked, consideration in microarray studies. It has been identified as a key issue [20], and we emphasize that poor sample design cannot be overcome with increased sample size. Ideally the sample should be representative of the population on which the signature is to be used. Again resampling methods can help to highlight problems. Note that most of the theoretical work underpinning justification for methodologies assumes that the samples are representative. There are many unknowns about how non-homogeneity in the sample affects the performance of the predictors.

In this data set there were confounders, including age, menopausal status, T stage, Grade, and PR status, that may be important predictors, including their possible interaction/s with particular genes, have not been considered here as they are not publicly available. Although Wang et al. [9] showed that these and other confounding factors were not statistically significant after adjusting for their gene signature

in the training set, their inclusion may be useful in predicting "new" samples. Their usefulness will arguably depend on how well these-factors are "accounted for" by the genes in the signature. The "best" signatures will in the end, most probably, be developed by a combination of biological expertise and computational algorithms.

Here we have demonstrated also the usefulness of using false positive and false negative fractions for assessing the performance of signatures. Particularly for medical applications, these measures of performance are more appropriate for assessing relevance of test performance than the overall classification (or misclassification) rate.

A promising approach has been to apply a protein-network-based method [21] that, instead of identifying markers as individual genes, considers them as subnetworks as extracted from protein interaction databases. Unfortunately the results suffer from selection bias. Two data sets were used [9, 22], although one of the data sets [22] included both node negative and node positive patients; just node negative patients could have been selected for inclusion in the analyses. It is important to correct for selection bias when estimating classification accuracy from use of the subnetwork markers. As outlined above, to remove selection bias, the subnetwork features should be identified at *each* step of the cross-validation, not "identified using all microarray samples before classification" [21, page 4].

5. Conclusion

We demonstrate the usefulness of the 10-fold cross-validation framework for assessing performance of signatures on a well-known Breast Cancer data set. Within this approach we developed two different methods, one based on forward selection, the other on genes that were statistically significant in all training blocks of data.

This paper demonstrates a lack of agreement between signatures estimated using a forward selection method on randomly selected subsets of the same data. Our novel method overcomes the instability of the forward selection method. It is computationally fast and is shown to have lower test error variance and FPF.

The 10-fold cross-validation framework is a very useful and less computer intensive method than some other re-sampling methods. It allows estimation of unbiased misclassification error as well as an assessment of the signature's sensitivity and specificity. Such evaluations are of utmost importance especially when the number of samples is small relative to the number of measurements.

The use of the pair of measurements, true positive fraction (sensitivity) and false positive fraction (1-specificity), provides a more clinically relevant evaluation of overall performance than the single measure given by the correctly predicted fraction (or its complement).

Although the hope of finding a single diagnostic tool is laudable, it is still too early to be making any strong claims. Given the complexity of human breast tumours, we agree that it is unlikely that "robust and internationally agreed signature gene lists will be accumulated in the near future" [7].

Finally, we have as yet to fully understand the enormous quantities of data that the rapidly evolving microarray technologies are giving us. Moreover, "although the use of gene-expression profiles in clinical practice is very appealing, we should be very cautious . . ." [23].

Acknowledgments

Part of this research was supported by the Australian Research Council (ARC) Grant DP0343727 and by the ARC Centre of Excellence for Mathematics and Statistics of Complex Systems (MASCOS). The authors thank the reviewers for their helpful comments.

References

[1] M. Schena, *Microarray Analysis*, Wiley-Liss, New York, NY, USA, 2003.

[2] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[3] S. G. Baker and B. S. Kramer, "Identifying genes that contribute most to good classification in microarrays," *BMC Bioinformatics*, vol. 7, article 407, pp. 1–7, 2006.

[4] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, NY, USA, 2001.

[6] R. Simon, "Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data," *British Journal of Cancer*, vol. 89, no. 9, pp. 1599–1604, 2003.

[7] T.-K. Jenssen and E. Hovig, "Gene-expression profiling in breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 634–635, 2005.

[8] R. Tibshirani, "Immune signatures in follicular lymphoma," *The New England Journal of Medicine*, vol. 352, no. 14, pp. 1496–1497, 2005.

[9] Y. Wang, J. G. M. Klijn, Y. Zhang, et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[10] C. J. Scarlett, R. C. Smith, A. Saxby, et al., "Proteomic classification of pancreatic adenocarcinoma tissue using protein chip technology," *Gastroenterology*, vol. 130, no. 6, pp. 1670–1678, 2006.

[11] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2005.

[12] Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo.

[13] J. Downward, "Cancer biology: signatures guide drug choice," *Nature*, vol. 439, no. 7074, pp. 274–275, 2006.

[14] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society B*, vol. 70, no. 5, pp. 849–911, 2008.

[15] J. M. Neuhaus, "Misspecification," in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Eds., vol. 5, pp. 3305–3307, John Wiley & Sons, Chichester, UK, 2004.

[16] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[17] I. A. Wood, P. M. Visscher, and K. L. Mengersen, "Classification based upon gene expression data: bias and precision of error rates," *Bioinformatics*, vol. 23, no. 11, pp. 1363–1370, 2007.

[18] G. J. McLachlan, J. Chevelu, and J. Zhu, "Correcting for selection bias via cross-validation in the classification of microarray data," in *Beyond Parametrics in Interdisciplinary Research: A Festschrift to P.K. Sen*, N. Balakrishnan, E. Pena, and M. J. Silvapulle, Eds., IMS Lecture Notes-Monograph Series, pp. 383–395, Institute of Mathematical Statistics, Hayward, Calif, USA, 2007.

[19] B. Ripley, "Computer intensive methods," in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton, Eds., vol. 2, pp. 1071–1075, John Wiley & Sons, Chichester, UK, 2004.

[20] E. Biganzoli, N. Lama, F. Ambrogi, L. Antolini, and P. Boracchi, "Prediction of cancer outcome with microarrays," *The Lancet*, vol. 365, no. 9472, pp. 1683–1686, 2005.

[21] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, pp. 1–10, 2007.

[22] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[23] A. Abdullah-Sayani, J. M. Bueno-de-Mesquita, and M. J. van de Vijver, "Technology insight: tuning into the genetic orchestra using microarrays—limitations of DNA microarrays in clinical practice," *Nature Clinical Practice Oncology*, vol. 3, no. 9, pp. 501–516, 2006.