BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 10 2005, pages 2200–2209
doi:10.1093/bioinformatics/bti370

Genome analysis

Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data

J. S. Yu[1,2], S. Ongarello[2,3], R. Fiedler[2], X. W. Chen[4], G. Toffolo[3], C. Cobelli[3] and Z. Trajanoski[5,*]

[1]School of Electronics Engineering and Computer Science, Peking University, China, [2]Institute for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria, [3]Department of Information Engineering, University of Padova, Italy, [4]Electrical Engineering and Computer Science Department, Information and Telecommunication Center, University of Kansas, USA and [5]Institute for Genomics and Bioinformatics, Christian-Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria

Received on September 14, 2004; revised and accepted on March 1, 2005
Advance Access publication March 22, 2005

ABSTRACT

Motivation: High-throughput and high-resolution mass spectrometry instruments are increasingly used for disease classification and therapeutic guidance. However, the analysis of immense amount of data poses considerable challenges. We have therefore developed a novel method for dimensionality reduction and tested on a published ovarian high-resolution SELDI-TOF dataset.

Results: We have developed a four-step strategy for data preprocessing based on: (1) binning, (2) Kolmogorov–Smirnov test, (3) restriction of coefficient of variation and (4) wavelet analysis. Subsequently, support vector machines were used for classification. The developed method achieves an average sensitivity of 97.38% (sd = 0.0125) and an average specificity of 93.30% (sd = 0.0174) in 1000 independent k-fold cross-validations, where $k = 2, \ldots, 10$.

Availability: The software is available for academic and non-commercial institutions.

Contact: zlatko.trajanoski@tugraz.at

1 INTRODUCTION

The novel biotechnology of high-throughput and high-resolution MALDI-TOF (matrix-assisted laser desorption and ionization time-of-flight) mass spectrometry (MS) makes it promising to explore the low-molecular-weight (LMW) region of the blood proteome for the diagnosis of significant patterns for various diseases (Lilien *et al*., 2003; Liotta *et al*., 2003; Petricoin and Liotta, 2003, 2004; Wulfkuhle *et al*., 2003). Molecular and statistical approaches to identifying ovarian cancer in the early stage are urgently needed, and much work has already been done (Anderson *et al*., 2003; Bao-Ling *et al*., 2002; Petricoin *et al*., 2002a,b; Qu *et al*., 2002; Vlahou *et al*., 2003; Wu *et al*., 2003; Yu *et al*., 2004). In this work we considered the SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) low-resolution and high-resolution raw MS data provided by National Cancer Institute (NCI),[1] relative to a study conducted to discriminate ovarian cancer from normal tissue. The published high-resolution data achieved with extensive quality control and assurance (QC/QA) analysis allow superior classification patterns when compared to those obtained with low-resolution instrumentation (Conrads *et al*., 2004). Recently NCI has published on its website the ovarian high-resolution QqTOF SELDI data mining results of the concordant m/z regions found by some particular classifications, with both sensitivity and specificity of almost 100%. Even now, universal robust methods for identifying ovarian cancer from MS data are still in development.

One of the best challenges is to keep the discriminatory features between two classes of interest while reducing the intolerable dimensionality (Duda *et al*., 2001). Petricoin *et al*. (2002a) used genetic algorithms and self-organizing clustering analysis to extract the discriminatory proteomic pattern from the low-resolution training set, achieving sensitivity and specificity of 100% simultaneously in some particular testing trials. Since the result of genetic algorithm converges to a local optimal solution, distinct random initializations of this iterative searching algorithm may lead to distinct solutions. This brings some problems when trying to identify significant biomarkers. Still, it is believed that there should be some interesting relationships between the extracted discriminatory patterns (Zhu *et al*., 2003). From the ovarian high-resolution MS data, Conrads *et al*. (2004) showed that the most frequent m/z ratios extracted by a similar method are 845.0895, 8602.237 and 8709.548. Vlahou *et al*. (2003) tested the method of classification and regression tree (CART) on the ovarian cancer discrimination from benign diseases and healthy controls, which resulted in a cross-validation accuracy of 81.5%.

In this work we have developed a novel method for dimensionality reduction and tested on a published ovarian high-resolution SELDI-TOF data set. We will show that the accuracy can be improved to 85–90% on the binned MS data. Several statistical methods for the classification of ovarian cancer based on MS spectra have been compared in Wu *et al*. (2003), and the method of random forest was demonstrated to outperform other methods like linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbor (k-NN), bagging (Bauer and Kohavi, 1999) and boosting classification trees.

*To whom correspondence should be addressed.

[1]See http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp for details.

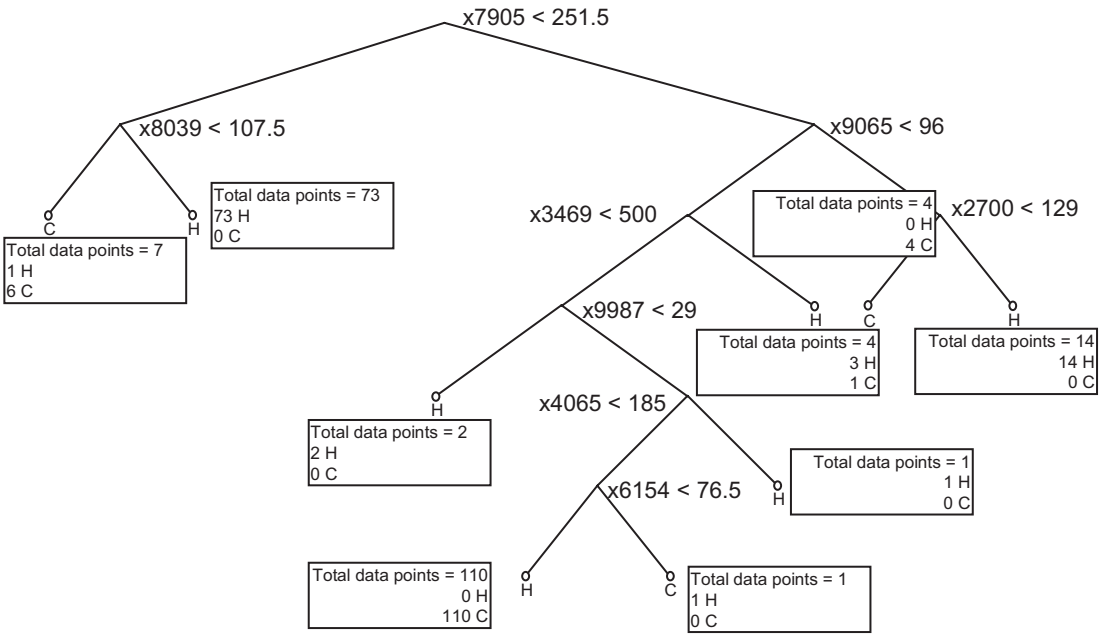

Fig. 1. Classification tree on the binned MS ovarian data, where x_i denotes the intensity at the i-th binned m/z ratio.

2 SYSTEMS AND METHODS

We make the following assumptions:

(1) If the intensity distributions of control and cancer are distinct at a specific m/z, this m/z ratio is supposed to be a useful feature to the classification.

(2) For the control and cancer data respectively, the m/z at which the random variable of intensity has a small coefficient of variation (CV) is representative.

The MS dataset can be written as $S = \{(x_i, y_i)|x_i \in \mathbb{R}^m, y_i = \pm 1, i = 1, 2, \ldots, n\}$, where $\boldsymbol{x}_i$ is an intensity vector according to a sorted sequence of m/z ratios and y_i is the class label of $\boldsymbol{x}_i$ (-1 for the healthy, $+1$ for cancer). A binary optimal classifier is a function $f{:}\mathbb{R}^m \to \pm 1$ such that $f(x_i) = y_i$ for both a training subset and testing subset of S. When the feature space is high-dimensional, feature selection becomes crucial as the first step towards pattern recognition. For the raw ovarian high-resolution SELDI-TOF dataset composed of 95 control samples and 121 cancer samples, the dimension of the original feature space is over 370 000.

To improve the performance of identifying ovarian cancer, we make use of a four-step data preprocessing procedure: (1) binning, (2) Kolmogorov–Smirnov (KS)-test based feature selection, (3) restriction of coefficient of variation and (4) discrete wavelet transformation. All the procedures do not depend on the particular classifier that will be used later, since they act just on the numerical characteristics of the MS data. Therefore, various classifiers could be trained and tested on the preprocessed data.

3 BINNING OF RAW MS DATA

In the first step, binning of raw MS data is performed (Fig. 1). Since the length of the observed m/z sequence varies in the raw MS data, we prepared the data as follows:

(1) Align the study sets according to the sorted union of m/z ratios into an intensity frame with missing data.

(2) To reduce the dimensionality, we binned the frame, at a given bin length $l > 0$, into a matrix A of m-by-n, where $n = 216$ (121 ovarian cancer samples and 95 control samples) and m is determined by l (Table 1). Each bin is an interval of the form $[b, b + l]$, where $b, l \in \mathbb{R}^+$. For the binned data, without ambiguity, the m/z ratio stands for the left boundary of an interval. After binning with $b \in \mathbb{N}, l = 1$, the dimension is reduced from 373 401 to 11 301.

Table 1. Mass spectrometry data matrix of control and cancer[a]

m/z	-1	$\cdots$	-1	1	$\cdots$	1
r_1	$x_{1,1}$	$\cdots$	$x_{1,k}$	$x_{1,k+1}$	$\cdots$	$x_{1,n}$
r_2	$x_{2,1}$	$\cdots$	$x_{2,k}$	$x_{2,k+1}$	$\cdots$	$x_{2,n}$
$\vdots$	$\vdots$					
r_m	$x_{m,1}$	$\cdots$	$x_{m,k}$	$x_{m,k+1}$	$\cdots$	$x_{m,n}$
	$\boldsymbol{x}_1$	$\cdots$	$\boldsymbol{x}_k$	$\boldsymbol{x}_{k+1}$	$\cdots$	$\boldsymbol{x}_n$

[a] For the sake of following preprocessings and support vector machine (SVM) classification, the intensity observation is recorded in the column vector.

The missing data are ignored in binning. Using 0-1 cost and 10-fold cross-validation, the classification tree achieves a precision of 85–90% on the binned MS data. But since CART is a greedy algorithm that decreases the Gini impurity the most at each step (Duda *et al.*, 2001), in general it does not guarantee an optimal reduction of entropy. The precision of the decision tree is regarded as a reference base line for the ovarian cancer diagnosis.

4 KS-TEST BASED FEATURE SELECTION

For each m/z ratio r_i, we compare the distributions of values in data vectors $X_i = (x_{i,1}, \ldots, x_{i,k})$ and $X_i' = (x_{i,k+1}, \ldots, x_{i,n})$ by a two-sided KS-test (i.e., the null hypothesis H_0 is that X_i and X_i' have

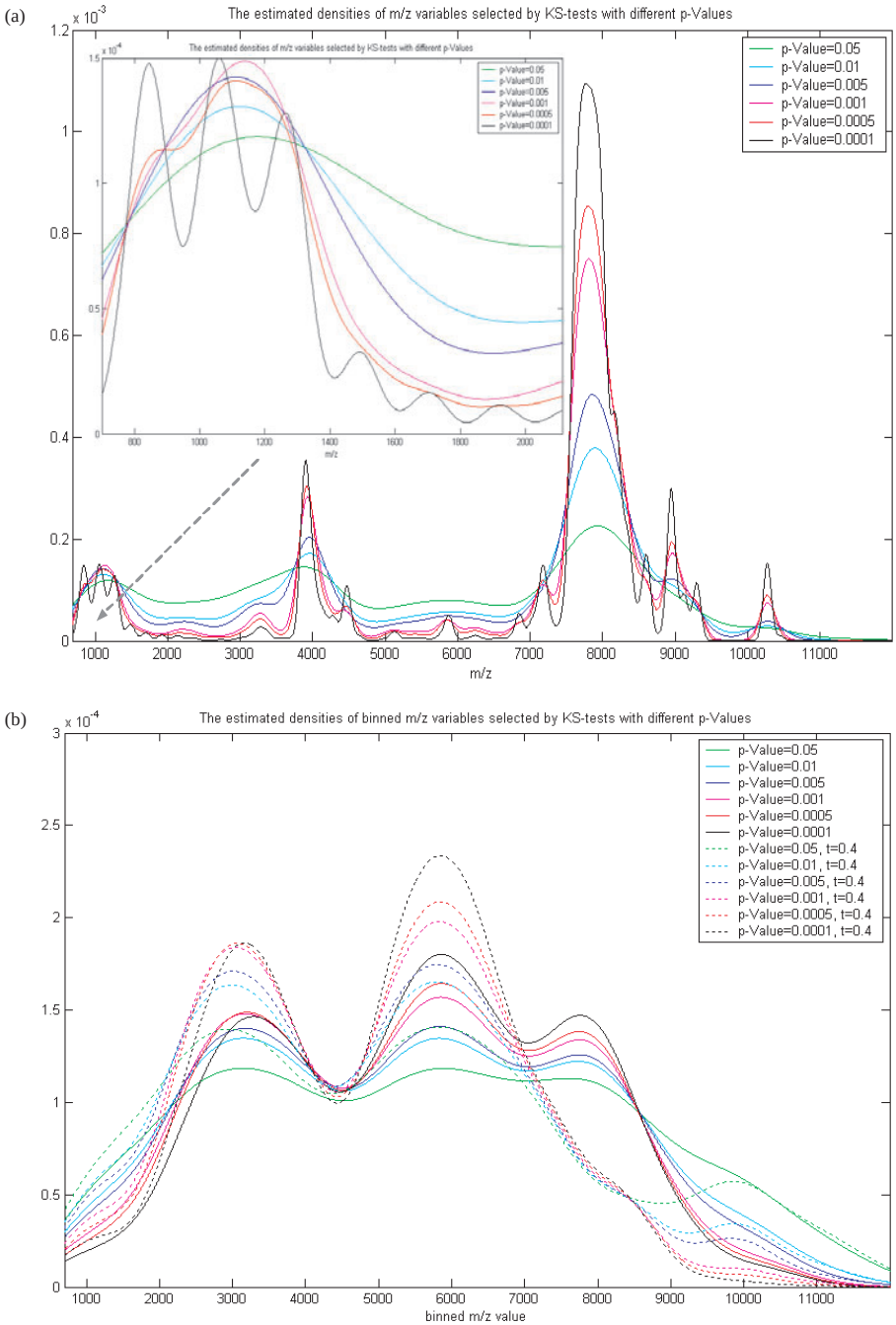

Fig. 2. **(a)** The estimated densities of raw m/z ratios selected by the KS-test with different p-values. The most frequent m/z ratios are around 8000, in accordance with m/z regions found by NCI. **(b)** The same experiments as (a) on the unit-binned m/z values. The dotted lines are the estimated distributions of binned m/z ratios further selected by a coefficient of variation (CV) restriction with threshold $t = 0.4$. This figure can be viewed in colour on *Bioinformatics* online.

the same distribution) with a given significance level α. For instance with $\alpha = 5\%$, H_0 cannot be rejected at $m/z = 703$, but rejected at $m/z = 8000$. The dimension of feature space is reduced to 8094 after choosing only the features that do not pass the KS-test at 5% level.

More flexible feature selection can be based on the reasonably accurate p-values that are guaranteed by $n_i n_i'/(n_i + n_i') \geq 4$, where n_i and n_i' are the sample sizes of X_i and X_i' respectively (Lehmann, 1975). As a comparison, Figure 2a shows the distributions of raw m/z ratios selected by KS-tests with distinct p-values (ignoring the missing intensity observations) which are quite different from those of the binned data plotted in Figure 2(b).

As Liotta *et al.* (2003) pointed out, the traditional single biomarker for a particular cancer makes little sense from a biological perspective, with poor identification of early-stage cancer and the benign. For instance, the widely used biomarker of cancer antigen 125 (CA125)

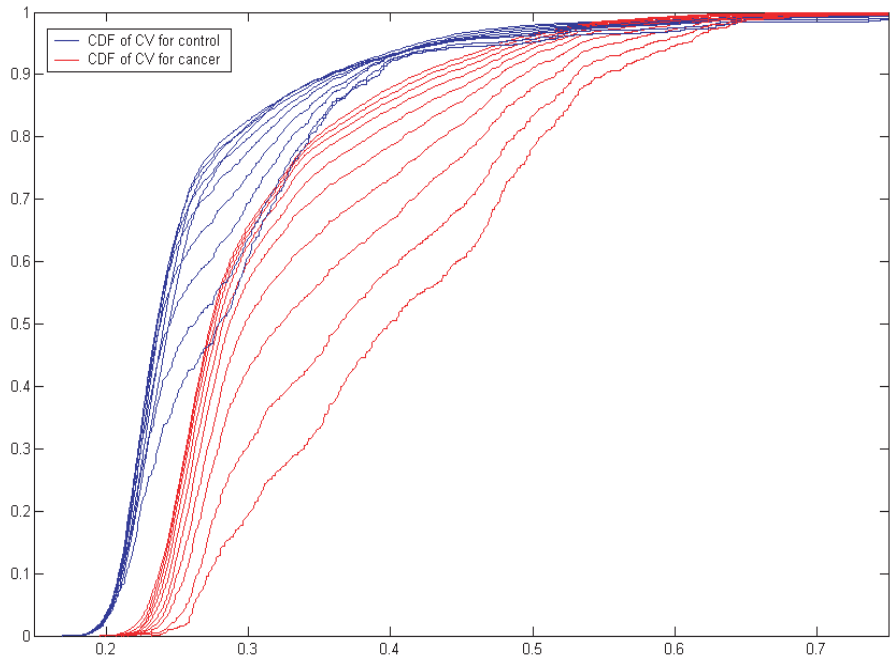

Fig. 3. Empirical cumulative distributions of intensity CV for control and cancer on the sets of binned m/z ratios selected by KS-tests with different p-values. Given a CV threshold $t_H = t_C = 0.4$, for the healthy, the proportion of m/z features with CV $\leq t_H$ is more than that for the cancerous, especially when $-\lg p$ increases. This figure can be viewed in colour on *Bioinformatics* online.

for ovarian cancer can only detect 50–60% of patients with stage I ovarian cancer. In contrast to the traditional way, the LMW biomarkers foreshow a satisfiable clinical application to early diagnosis of ovarian cancer (Petricoin *et al.*, 2002a,b). From the viewpoint of pattern recognition, the biomarkers are those key features that allow a well-done classification. Usually, classification and feature selection are entangled. To avoid the over-fitting problem, several trials of feature selection are suggested, independent of classifiers as much as possible. Also, in our opinion, the biomarkers could be many selected binned m/z ratios, not necessarily particular m/z ratios if they are not able to yield a satisfiable result (Diamandis, 2004).

5 RESTRICTION OF COEFFICIENT OF VARIATION

For a positive random variable X, the coefficient of variation (CV) is defined as $c = \mathrm{sd}(X)/\mathrm{E}(X)$, which can be estimated by $\hat{c} = s/\bar{X}$ where s and $\bar{X}$ are the sample standard deviation and sample mean respectively. The m/z ratio with relatively small CV is considered as a useful feature for the classification. The CV of intensity for the healthy and cancerous will be considered separately.

Given CV thresholds of intensity, for instance $t_H = 0.4, t_C = 0.4$ for the healthy and cancerous, the feature space dimension is reduced from the second to the fourth column in Table 2.

The estimated distributions of binned m/z ratios selected by the KS-test with distinct p-values and a CV restriction of $t = t_H = t_C = 0.4$ are illustrated in Figure 2b. We suggest a threshold such that 85–95% of the binned m/z ratios are included by the respective empirical cumulative distribution function (CDF) of CV for control and cancer. By empirical CDFs of CV, one can choose the probabilities for control and cancer in advance, then get the corresponding CV thresholds (Fig. 3). In the following, we will set p-value $= 0.05$ and $t_H = t_C = 0.4$, which reduces the dimension of the feature space from the original 373 401 to 6757.

Table 2. Feature selection by KS-test and CV restriction[a]

$-\lg p$	Count of selected m/z ratios		
		CV restriction	
	KS-test	$t_H = 0.4$	$t_C = 0.4$
2	6936	6459	5703
3	5757	5366	4678
4	4818	4489	3855
5	3854	3585	3000
6	2950	2749	2208

[a]The italicized numbers are the consequential dimensions of feature space after a CV restriction.

Alternatively, for normally distributed n control (or cancer) intensities at a particular binned m/z ratio, the point estimate of c can be replaced by a $(1-\alpha)$-confidence upper bound $\tilde{c}$ conservatively, determined by the following equation:

$$F_{t(n-1,\sqrt{n}/\tilde{c})}(\sqrt{n}/\hat{c}) = 1 - \alpha \tag{1}$$

where $F_{t(n-1,\sqrt{n}/\tilde{c})}(\cdot)$ is the non-central t distribution function with freedom degree $n-1$ and non-central parameter $\sqrt{n}/\tilde{c}$. Especially when $\hat{c} \leq 0.3$, by McKay's theorem (McKay, 1932),

$$\frac{n(s/\bar{X})^2(1+c^2)}{[1+(s/\bar{X})^2]c^2} \sim \chi^2(n-1) \tag{2}$$

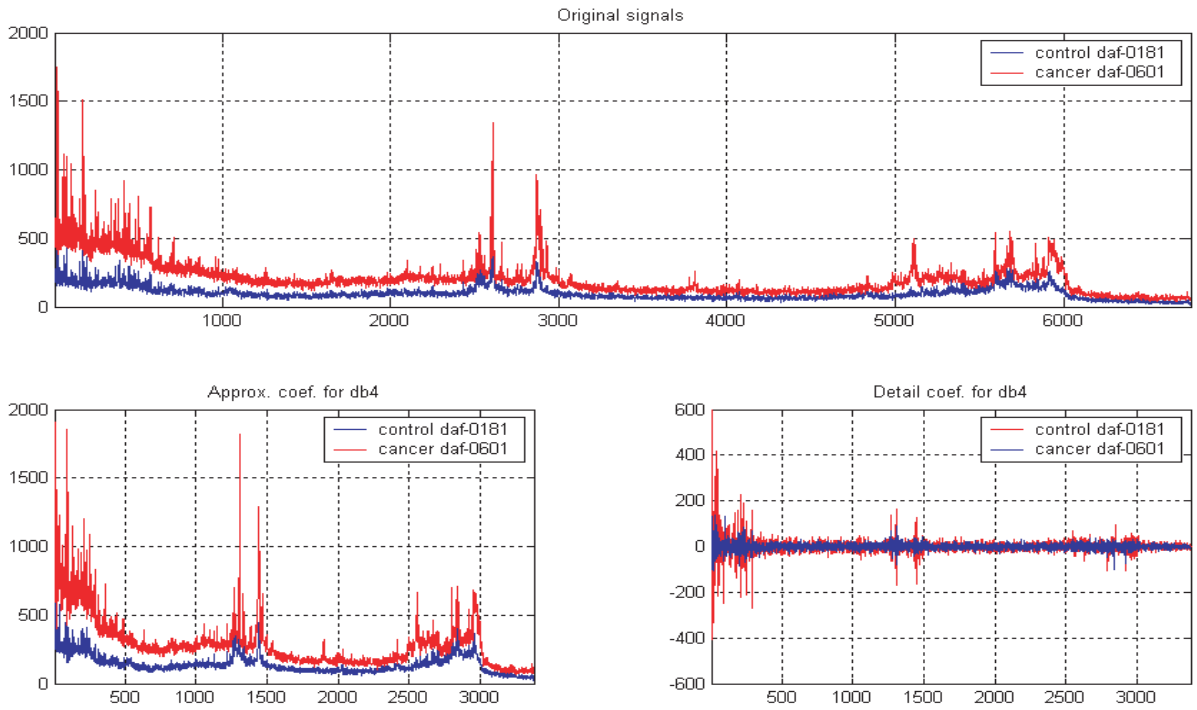

Fig. 4. Single-level DWT of binned m/z data (the X-axis in the first figure is the ordering of binned m/z ratios, not m/z values). This figure can be viewed in colour on *Bioinformatics* online.

we have a $(1-\alpha)$-confidence upper bound

$$\tilde{c} \approx \left\{ \frac{\chi_\alpha^2 (n-1)[1+(s/\bar{X})^2]}{n(s/\bar{X})^2} - 1 \right\}^{-1/2} \qquad (3)$$

6 WAVELET TRANSFORMATION

Wavelet analysis has achieved a broad and successful application to pattern recognition in the last decade, such as in image compression, turbulence or earthquake prediction. It is also an efficient way to compress self-similar data, localizing a signal in both time and frequency (Chui, 1992; Daubechies, 1992). Compared with Fourier transformation, wavelets have advantages in analyzing physical situations where the signal contains discontinuities and sharp spikes. Recently, there is a growing interest in applying wavelet analysis to biomolecular related signals (Liò, 2003).

After applying the pyramidal algorithm (Mallat, 1989) of discrete wavelet transformation (DWT, linear complexity), the binned selected MS spectrum is compressed further to a 3382-dimensional vector of approximation coefficients, which contains most key information for classification (Fig. 4). Considering the dimension reduction, the undecimated DWT without restriction on the length of signal (traditionally it must be a power of 2) is not preferred, although it is superior to the ordinary one in many statistical applications (Nason *et al.*, 2000). The mother wavelet used in the experiment is Daubechies family (db4) and the boundary values are symmetrically padded. The vector of approximation coefficients acts as a fingerprint of the original raw MS data, with a compression rate of more than 100. What is more, the two samples are separated in a manner of keeping the main discriminatory information for classification. Theoretically, a heavier compression rate can be achieved, at the risk of losing some useful information, by choosing a higher level of approximation coefficients.

Table 3. Average performances of SVMs on the low-resolution and wavelet-reduced ovarian data in 1000 independent 2-fold cross validations

Data	Control Mean	SD	Cancer Mean	SD
Approximate coefficients	0.9113	0.0148	0.9975	0.0042
Detail coefficients	0.9116	0.0200	0.9919	0.0102
Original	0.9112	0.0147	0.9975	0.0042

When applying the procedure recursively, a KS-test on the dataset of wavelet approximation coefficients, with a significance level $\alpha = 5\%$, can hardly reduce the dimension of the feature space any more, neither does the reasonable restriction of CV. Therefore, in a sense of data reduction, the vector of approximation coefficients inherits the key discriminatory traits of MS data.

For the high-resolution ovarian data, the vector of detail coefficients contains almost no information for the healthy, since SVMs identify all the data as cancers. In contrast, the detail coefficients calculated from the low-resolution ones lead to acceptable results (Table 3). Without KS-test based feature selection and restriction of CV, the original low-resolution ovarian dataset 8-7-02 (91 controls versus 162 cancers, http://ncifdaproteomics.com/lowresovarian.php) with 15 154 features are well classified by SVMs of Gaussian kernel parameterized by $\gamma = 2, C = 1$ (Section 4). Compressed by db2, even the data of detail coefficients yield a good result, especially on the identification of ovarian cancer.

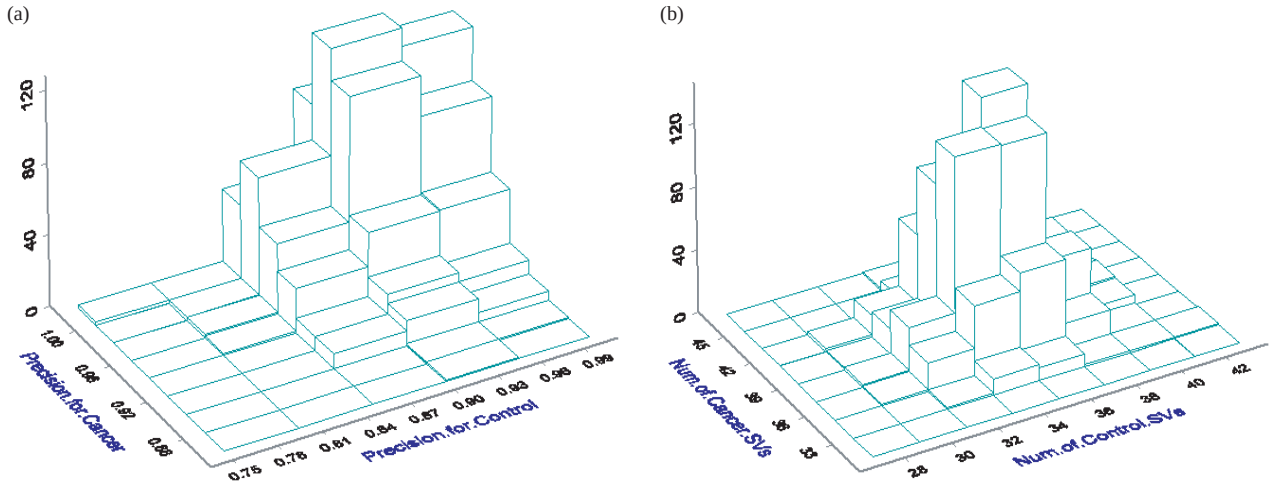

Fig. 5. (a) Distribution of precision in 1000 independent 2-fold proportional validations. There are totally 106 times that the classifier yields more than 98.3% sensitivity and 97.8% specificity simultaneously in the 1000 random experiments, in which eight times of both 100% sensitivity and 100% specificity. **(b)** Distribution of SV number in the last experiment. To achieve both 100% sensitivity and 100% specificity, at least 72 support vectors (35 controls and 37 cancers) are needed. The median counts of control SVs and cancer SVs are 35 and 39 respectively. In general, ♯(Control SVs)/♯(Training Controls) < ♯(Cancer SVs)/♯(Training Cancers), except 24 accidents. This figure can be viewed in colour on *Bioinformatics* online.

7 CLASSIFICATION BY SVM

The SVM method is a widely used classification method of Statistical Learning Theory, originally started by Vapnik and Chervonenkis in the 1960s (Vapnik, 1998). In case that the training set S is linearly separable, the support vector (SV) classifier (Burges, 1998; Cortes and Vapnik, 1995; Cristianini and Shawe-Taylor, 2000; Schölkoft, 1997) is the hyperplane $H: w^{\mathsf{T}}x+b=0$ (where $w \in \mathbb{R}^m, b \in \mathbb{R}$) with the maximal margin separating the two classified subsamples of S.

Generally in the linearly non-separable case, we reach at a soft margin allowing training errors (Shawe-Taylor and Cristianini, 2000), where the classifier H is the solution of the optimization problem that is solved by the method of quadratic programming.

Another approach to the linearly non-separable case is the kernel method (Shawe-Taylor and Cristianini, 2004; Schölkoft and Smola, 2002), by which the training data are mapped into a higher dimensional feature space $\mathbb{H}$ (may be infinite) and become more separable, provided that the map φ makes $K(x,y) = \varphi(x)^{\mathsf{T}}\varphi(y)$ a kernel function guaranteed by the Mercer's condition (Vapnik, 1998).

Besides these techniques, there are still many significant methods of SVM in the published literature that are far more than what it would be appropriate to include here. Quite a few linkages to free SVM softwares or packages, implemented in C (or C++), Fortran, Java, Perl, R and MatLab are available at http://www.support-vector.net/software.html, for instance some popular ones like Joachims' SVM$^{\text{light}}$ (Joachims, 1999) and Lin's libSVM (Lin, 1999).

Applied to the vectors of normalized wavelet approximation coefficients, the Gaussian radial basis function $K(x_i, x_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ was adopted as the kernel of the non-linear SV classifier.

The performance of classification, for a more comprehensive review, was examined by (1) k-fold cross-validation, where $k = 2, 3, \ldots, 10$; (2) k-fold proportional validation: randomly select $100(1-1/k)\%$ controls and $100(1-1/k)\%$ cancers as the training set and test the classifier on the remaining samples; and (3) leave-one- out cross-validation. For each k-fold validation, the random experiment was repeated 1000 times independently (Fig. 5).

In leave-one-out cross-validation, six controls and two cancers were misclassified (NCI's high-resolution ovarian data contain 95 controls and 121 cancers). In addition, by the comparison of standard deviation between the classification precisions, all the k-fold (cross, proportional) validations show that the SV classifier is relatively stable to the cancer samples but mutable to those controls (Table 4), which coincides with our intuition about the nature of noise in control data. For each k-fold validation, the average precisions are estimated by the results of 1000 independent random experiments. Obviously, k-fold proportional validation is stricter than k-fold cross-validation, and better at surveying the robustness of classifier for each category. For instance, the worst specificity and sensitivity in 2-fold cross-validations were 85.96 and 90.30%, respectively, while for 2-fold proportional validations they were 76.60 and 86.67% [Fig. 5(a)]. Moreover, the fact of bigger and bigger deviations of control precision in k-fold proportional validations indicates that the SVM overfitting problem seems more serious for the control samples. The same thing also happens to the SV classification on the more reduced data by principal component analysis (PCA) discussed in the next section.

Another compelling application is the bagging (Bauer and Kohavi, 1999) of one-hidden-layer neural network. In leave-one-out cross-validation, each training set is resampled 100 times to estimate the output (Fig. 6, Table 5). Totally, only one cancer and four controls are misclassified. However, its complexity is inferior to SVMs.

Besides decreasing the computational complexity, the procedure of 'raw MS data → binned MS data → KS-test based feature selection → restriction of CV → wavelet analysis' depurates the original data, and explores their category traits for the coming classification. Some other classifiers are able to benefit from the procedure as shown in the next section.

Table 4. k-fold (cross, proportional) validation of SVM ($\gamma = 20, C = 0.7$) on the preprocessed data

k	k-fold cross-validation				k-fold proportional validation			
	Control		Cancer		Control		Cancer	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	0.9330	0.0174	0.9738	0.0125	0.9335	0.0362	0.9747	0.0217
3	0.9393	0.0188	0.9783	0.0115	0.9411	0.0397	0.9772	0.0234
4	0.9409	0.0200	0.9786	0.0118	0.9431	0.0456	0.9780	0.0279
5	0.9425	0.0203	0.9794	0.0119	0.9451	0.0503	0.9811	0.0266
6	0.9411	0.0223	0.9806	0.0118	0.9429	0.0597	0.9800	0.0320
7	0.9412	0.0210	0.9805	0.0118	0.9432	0.0646	0.9812	0.0326
8	0.9414	0.0222	0.9815	0.0117	0.9401	0.0708	0.9814	0.0336
9	0.9423	0.0231	0.9817	0.0117	0.9433	0.0716	0.9806	0.0397
10	0.9406	0.0226	0.9819	0.0113	0.9447	0.0739	0.9829	0.0385

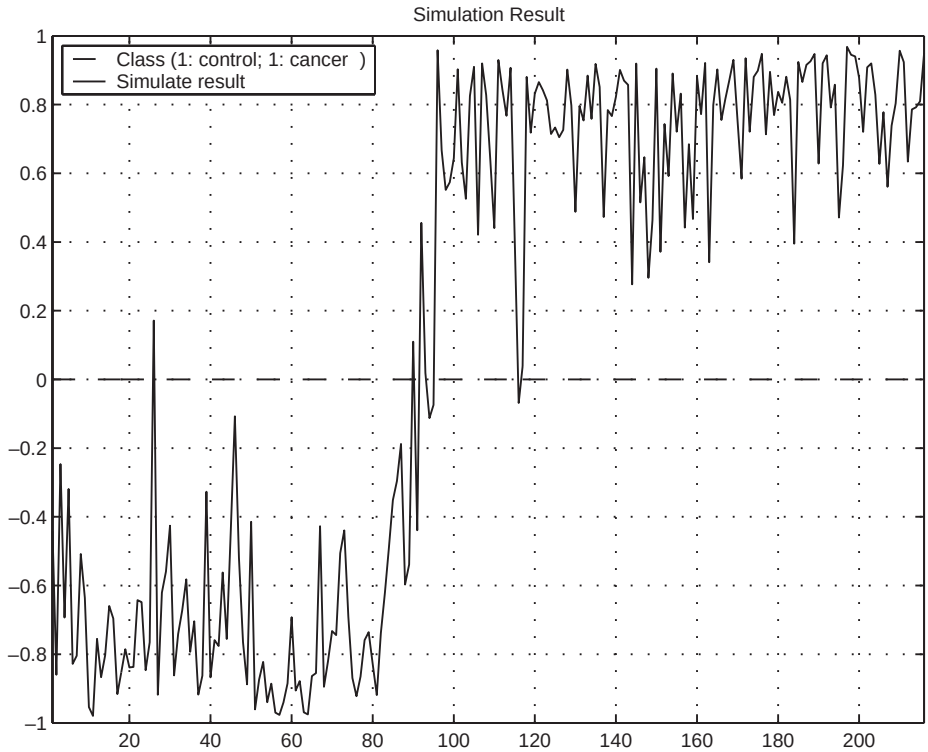

Fig. 6. The average prediction of each testing sample point in leave-one-out cross-validations, based on the resampled training set.

Table 5. 2,10-fold cross-validations for the bagging of one-hidden-layer perceptron by the backpropagation algorithm[a]

Cross-validation	Control		Cancer	
	Mean	SD	Mean	SD
2-fold	0.9312	0.0182	0.9617	0.0155
10-fold	0.9484	0.0201	0.9775	0.0127

[a]The number of hidden units is 300 and the transfer function is tansig.

8 ALTERNATIVE CLASSIFIERS

The method illustrated in the previous section leads to a good result in classification, especially when dealing with cancer samples. In an effort to reach a better precision of identifying controls, we tested several algorithms (voted perceptron, discriminant analysis, decision trees, naïve Bayes, some meta learning schemes like bagging and decorate, random forest) on the preprocessed data as described in Sections 2 and 3, but the precisions were generally inferior to those obtained by SVMs (results are shown in Table 6).

Since the high number of features (3382) could affect the performance of the mentioned classification algorithms, we performed

Table 6. 2,10-fold cross-validations of some methods on the preprocessed data

Method	2-fold cross-validation				10-fold cross-validation			
	Control		Cancer		Control		Cancer	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
VP	0.9254	0.0284	0.9527	0.0231	0.9445	0.0133	0.9700	0.0123
ADAboost	0.8959	0.0304	0.9192	0.0254	0.9205	0.0269	0.9403	0.0164
1-NN	0.8320	0.0301	0.8935	0.0226	0.8598	0.0149	0.9245	0.0103
3-NN	0.8131	0.0352	0.8963	0.0273	0.8362	0.0162	0.9299	0.0117
ADtree	0.8180	0.0432	0.8621	0.0350	0.8377	0.0271	0.8805	0.0299
J48tree	0.7785	0.0441	0.8250	0.0378	0.8022	0.0302	0.8462	0.0305
Random forest	0.7760	0.0306	0.8178	0.0232	0.8226	0.0089	0.8389	0.008
Naive bayes	0.6591	0.0162	0.7186	0.0124	0.6604	0.0055	0.7212	0.0450

Table 7. k-fold (cross, proportional) validation of SVM ($\gamma = 1.7, C = 0.7$) on PCA-reduced data of the first nine components[a]

k	k-fold cross-validation				k-fold proportional validation			
	Control		Cancer		Control		Cancer	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	0.8930	0.0267	0.9492	0.0270	0.8983	0.0465	0.9533	0.0448
3	0.9034	0.0246	0.9650	0.0213	0.9041	0.0506	0.9640	0.0412
4	0.9058	0.0255	0.9722	0.0177	0.9102	0.0575	0.9715	0.0370
5	0.9066	0.0256	0.9740	0.0162	0.9095	0.0604	0.9746	0.0371
6	0.9094	0.0262	0.9760	0.0149	0.9091	0.0695	0.9764	0.0392
7	0.9083	0.0267	0.9773	0.0141	0.9083	0.0788	0.9782	0.0375
8	0.9098	0.0264	0.9784	0.0134	0.9104	0.0838	0.9769	0.0401
9	0.9098	0.0262	0.9801	0.0125	0.9088	0.0871	0.9785	0.0420
10	0.9096	0.0269	0.9801	0.0127	0.9081	0.0920	0.9803	0.0429

[a]Compared with the results recorded in Table 4, the data reduction by PCA does not affect the sensitivity very much, but makes the specificity worse (Fig. 7).

a PCA on the preprocessed data, in order to further decrease their dimensionality. Since finding the optimal number of components for accurate classification is a non-trivial task, we applied several algorithms to the PCA-reduced dataset with a different amount of components, and at last selected the first nine as the coordinates of the updated feature space, which explain 90% variance of the data.

In leave-one-out cross-validation, nine controls and two cancers are misclassified by the SVM with parameters $\gamma = 1.7$ and $C = 0.7$. The other experimental results of this SVM are reported in Table 7. If the training set is large enough, for instance as big as those in 4-fold cross-validation, the PCA reduction affects the distribution of sensitivity little. But it is not the case for specificity (Fig. 7).

The classifier of Voted Perceptron (VP) (Freund and Schapire, 1999), a method of combining Rosenblastt's perceptron algorithm (Duda *et al*., 2001) with Helmbold and Warmuth's leave-one-out method (Helmbold and Warmuth, 1995), is also an approach to small sample analysis, taking advantage of the 'boundary data' of largest margin just as SVM does. Using the more reduced data projected on the first nine components, VP achieves a best performance in k-fold cross-validations, compared with quadratic discriminant analysis (QDA), linear discriminant analysis (LDA), Mahalanobis discriminant analysis (MDA), k-NN, naïve Bayes (NB), bagging (bootstrap aggregating), ADtree and J48tree (Table 8). When the pooled covariance matrix of training data is not positive definite, a common solution is to randomly perturb the training data.

9 IMPLEMENTATION

The software was implemented by OSU SVM toolbox for MATLAB, based on Dr Lin's libSVM-v.2.33. Most of the other classification algorithms were taken or re-adapted from the versions present in WEKA (Witten and Frank, 2000).

10 DISCUSSION

We have developed an efficient method for dimensionality reduction from MS data based on a 4-step strategy: (1) binning; (2) two-sample KS test, (3) restriction of coefficient of variation and (4) wavelet analysis. By efficient preprocessing of high-resolution ovarian MS data, SVMs achieve a satisfiable performance of identifying cancer and the healthy. On the one hand, data preprocessing reduces the dimension of feature space; on the other hand it extrudes the most significant category traits for the coming classification.

Although low-resolution data also lead to a high precision in identifying cancers, a recent study (Baggerly *et al*., 2004) pointed out many of the problems that seem to characterize this dataset, partly connected with experimental procedure and design, partly with technical

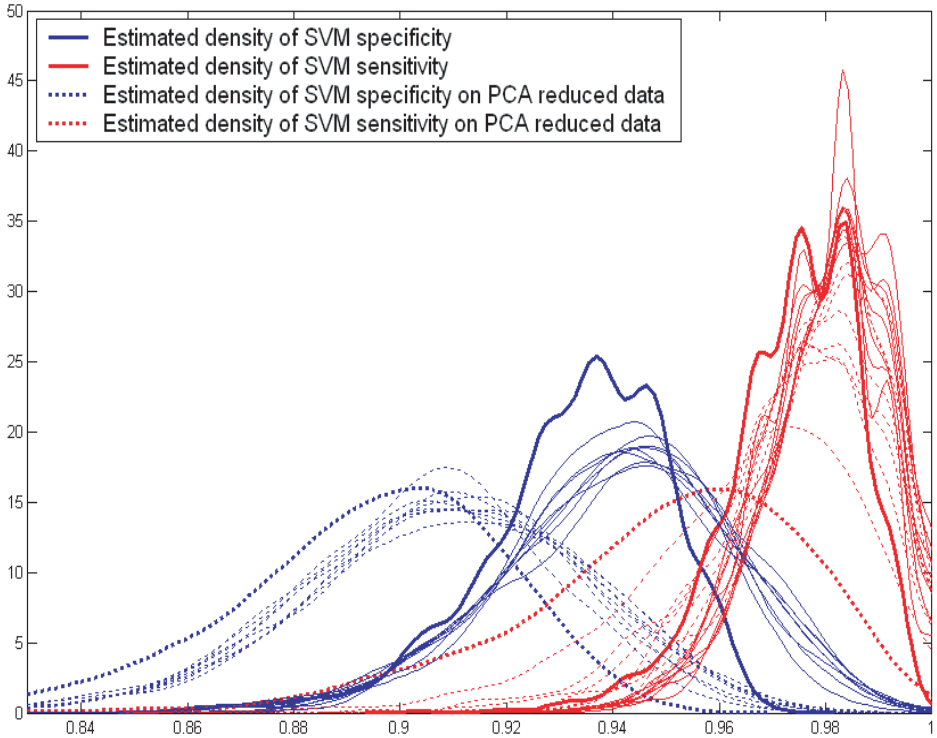

Fig. 7. Estimated densities of precisions in k-fold cross-validations, where $k = 2, 3, \ldots, 10$. This figure can be viewed in colour on *Bioinformatics* online.

Table 8. 2,10-fold and leave-one-out cross-validations of VP, QDA, LDA, MDA, NB, bagging, k-NN, ADtree and J48tree on PCA-reduced data of the first nine components[a]

Method	2-fold cross-validation				10-fold cross-validation				♯(misclassifications) of leave-one-out	
	Control		Cancer		Control		Cancer			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Control	Cancer
VP	0.9393	0.0209	0.9583	0.0192	0.9482	0.0140	0.9691	0.0096	3	3
QDA	0.9202	0.0224	0.9429	0.0226	0.9255	0.0264	0.9647	0.0161	4	2
LDA	0.9179	0.0189	0.9467	0.0156	0.9255	0.0267	0.9522	0.0193	7	1
MDA	0.9392	0.0243	0.9154	0.0291	0.9591	0.0201	0.9267	0.0237	3	10
NB	0.8803	0.0280	0.9190	0.0203	0.8979	0.0130	0.9249	0.0088	9	9
Bagging	0.8835	0.0307	0.9174	0.0224	0.8977	0.0145	0.9232	0.0113	9	10
1-NN	0.8575	0.0288	0.8889	0.0259	0.8902	0.0326	0.9018	0.0269	10	12
2-NN	0.7260	0.0375	0.9641	0.0164	0.8063	0.0409	0.9745	0.0140	17	3
ADtree	0.8238	0.0483	0.8878	0.0343	0.8498	0.0274	0.9025	0.0226	18	8
J48tree	0.7818	0.0220	0.8507	0.0405	0.8245	0.0280	0.8825	0.0201	18	14

[a]Except MDA, all the methods discussed in this paper work better on the identification of ovarian cancer than that of control.

problems (especially calibration issues). One of the most relevant observations in our experiments is the fact that it is possible to get very good classification results employing just detail coefficients of DWT. This can either mean that 'noise' is in fact still enough for classification purposes, or that the whole data is affected by some form of corruption that prevents achieving a perfect classification.

The classifier-independent data preprocessing of proteomic MS data shows a promising approach to the coming classification. More robust classifiers (such as Bayesian SVM and Bayesian neural network) are still urgently needed, as well as their ensemble. In addition, the precisions could be further improved by some resampling method (Gelman *et al*., 2004), which assigns every testing sample point a probability of being cancer.

ACKNOWLEDGEMENTS

This work was supported by the Austrian GEN-AU project BIN (Bioinformatics Integration Network), ÖAD (Austrian Exchange Service) and the EU Marie Curie Training Site grant Genomics of Lipid Metabolism.

REFERENCES

Anderson,D.C. *et al*. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.*, **2**, 137–146.

Baggerly,K.A. *et al*. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, **20**, 777–785.

Bao-Ling,A. *et al*. (2003) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Bauer,E. and Kohavi,R. (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Machine Learning*, **36**, 105–139.

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

Chui,C.K. (1992) *An Introduction to Wavelets*. Academic Press, New York.

Conrads,T.P. *et al*. (2004) High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*, **11**, 163–178.

Cortes,C. and Vapnik,V.N. (1995) Support vector networks. *Machine Learning*, **20**, 273–297.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge.

Daubechies,I. (1992) *Ten Lectures on Wavelets*. SIAM, Philadelphia.

Diamandis,E.P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell. Proteomics*, **3**, 367–378.

Duda,R.O., Hart,P.E. and Stork,D.G. (2001) *Pattern Classification*. John Wiley & Son, Inc., New York.

Freund,Y. and Schapire,R.E. (1999) Large margin classification using the perceptron algorithm. *Machine Learning*, **37**, 277–296.

Gelman,A., Carlin,J.B., Stern,H.S. and Rubin,D.B. (2004) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC Press.

Helmbold,D.P. and Warmuth,M.K. (1995) On weak learning. *J. Comput. Syst. Sci.*, **50**, 551–573.

Joachims,T. (1999) Making large-scale SVM learning practical. In Schölkoft,B., Burges,C.J.C. and Smola,A.J. (eds), *Advances in Kernel Methods—Support Vector Learning*. MIT Press, pp. 169–184.

Lehmann,E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks*. Hoden-Day, San Francisco.

Lilien,R.H. *et al*. (2003) Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Comput. Biol.*, **10**, 925–946.

Lin,C.J. (1999) Formulations of support vector machines: a note from an optimization point of view. *Technical report*, National Taiwan University, Dept. of Computer Science.

Liò,P. (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, **19**, 2–9.

Liotta,L.A. *et al*. (2003) Clinical proteomics: written in blood. *Nature*, **425**, 905.

Mallat,S.G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Pattern Anal. Machine Intell.*, **11**, 674–693.

McKay,A.T. (1932) Distribution of the coefficient of variation and the extended '*t*' distribution. *J. R. Statist. Soc.*, **95**, 695–698.

Nason,G.P. *et al*. (2000) Wavelet processes and adaptive estimation of the evolutionary spectrum. *J. R. Statist. Soc.*, **62**, 271–292.

Petricoin,E.F. and Liotta,L.A. (2003) Mass spectrometry-based diagnostics: the upcoming revolution in disease detection. *Clin. Chem.*, **49**, 533–534.

Petricoin,E.F. and Liotta,L.A. (2004) SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotechnol.*, **15**, 24–30.

Petricoin,E.F. *et al*. (2002a) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Petricoin,E.F. *et al*. (2002b) Proteomic patterns in serum and identification of ovarian cancer. *Lancet*, **360**, 170–171.

Qu,Y. *et al*. (2002) Boosted decision tree analysis of surface-enhanced laser desorption /ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.*, **48**, 1835–1843.

Schölkoft,B. (1997) *Support Vector Learning*. R. Oldenbourg Verlag, Munich.

Schölkoft,B. and Smola,A.J. (2002) *Learning with Kernels*. MIT Press.

Shawe-Taylor,J. and Cristianini,N. (2000) Margin distribution and soft margin. In Smola,A.J., Bartlett,P.L., Schölkoft,B. and Schuurmans,D. (eds), *Advances in Large Margin Classifiers*. MIT Press, pp. 349–358.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Vapnik,V.N. (1998) *Statistical Learning Theory*. John Wiley & Son, Inc., New York.

Vlahou,A. *et al*. (2003) Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *J. Biomed. Biotechnol.*, **5**, 308–314.

Witten,H. and Frank,E. (2000) *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco.

Wu,B. *et al*. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.

Wulfkuhle,J.D. *et al*. (2003) Proteomic applications for the early dectection of cancer. *Nature*, **3**, 267–275.

Yu,J.K. *et al*. (2004) An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. *World J. Gastroenterol.*, **10**, 3127–3131.

Zhu,W. *et al*. (2003) Detection of cancer-specific markers amid massive mass spectral data. *PNAS*, **100**, 14666–14671.